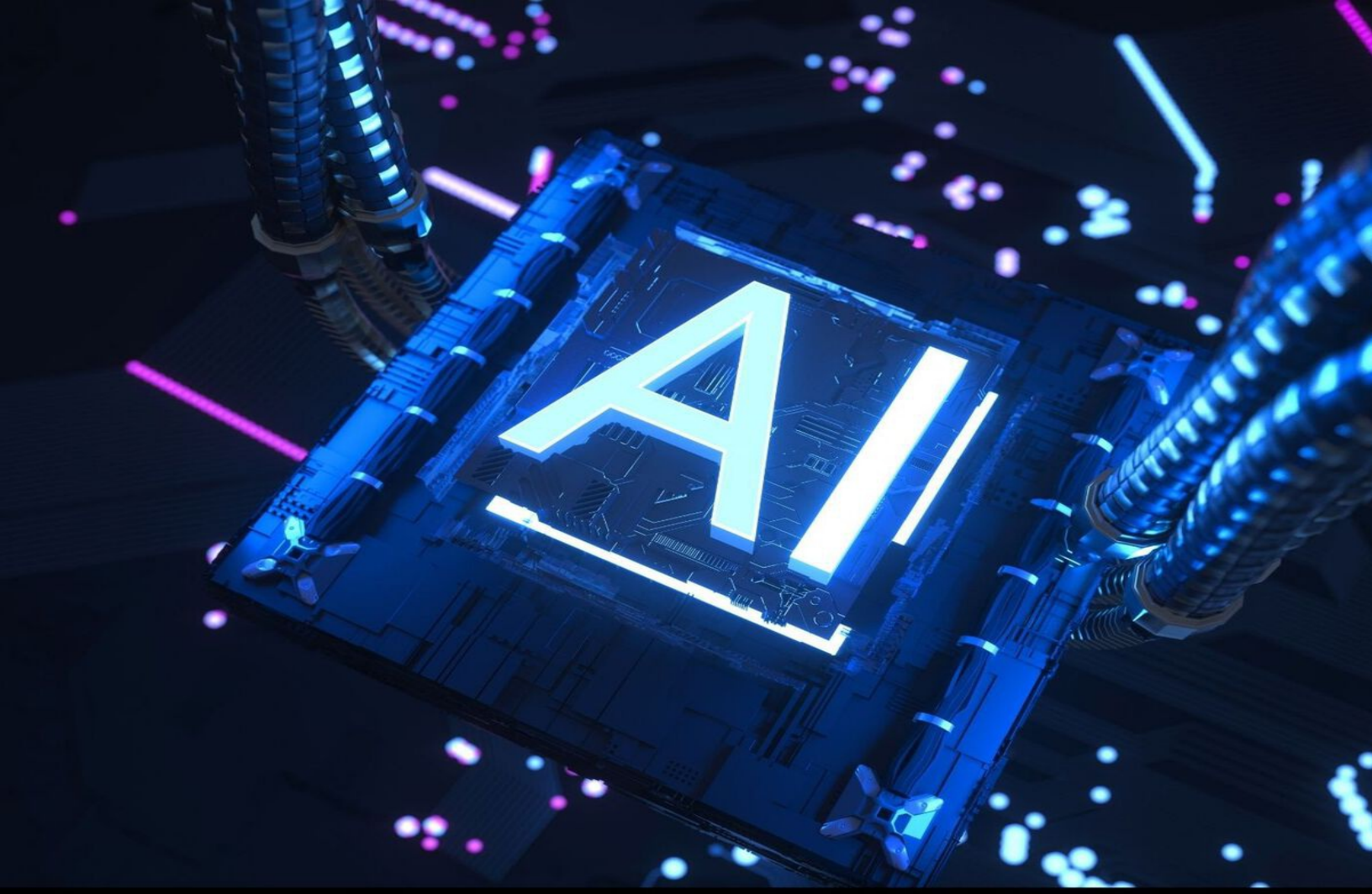


ARTIFICIAL INTELLIGENCE GOVERNANCE AND CYBER-SECURITY

A beginner's guide to governing and securing AI

TAIMUR IJLAL

ARTIFICIAL INTELLIGENCE (AI) GOVERNANCE AND CYBER-SECURITY

A beginner's guide to governing and securing AI

Taimur Ijlal

About this book

This edition was published in **April 2022**

I have tried to keep it as up to date as possible with the latest news and trends regarding Artificial Intelligence risks and cyber-security. The rapid rate at which AI evolves however means that I will be making regular updates to this book whenever any major change happens.

This book is dedicated to my wife and parents. My wife, who regularly pushes me to take on new challenges and risks to better myself. My parents both of whom raised me to be the person I am today and never let me feel that I could not achieve what I set my mind to.

Thanks to all the people who watch my YouTube Channel

"Cloud Security Guy" and appreciate all the comments/feedback I receive

*Last of all, **THANK YOU** for purchasing this book and I hope it helps increase your desire to learn about Artificial Intelligence Risks and Cyber-Security.*

Taimur Ijlal

Contents

About the book

This is a book about Artificial Intelligence (AI) Governance and Cyber-Security. I wrote it to be as simple and to the point as possible about its topic. It will teach you the fundamental concepts of Artificial Intelligence, its impact, and **most importantly** how to secure and govern it.

This book WILL:

- Give you all the core concepts you need to understand AI
- Demystify how AI works so it longer seems so intimidating
- Make you understand the revolutionary impact of AI and what are the changes that are going to happen over its landscape in the coming years
- Teach you some of the scary implications of AI and the real-world impact they can have.
- Guide you how to create a governance framework for AI systems so that they are operated in a responsible manner
- Educate you on how to protect AI systems from cyber-attacks and how to create a cyber-security framework in which AI systems can work securely
- Teach you about the tools and techniques you can use to identify cyber-risks in AI systems and how to mitigate them

This book DOES NOT:

- Teach you a programming language like Python or GO and teach you how to start coding
- Get down into the mathematical details of how AI algorithms work
- Make you an expert in Data Science, Neural networks, Deep learning, etc.

Why should I read this book?

Saying that Artificial Intelligence is having a huge impact on the world is a bit like saying that Godzilla causes a few traffic jams when he attacks a city i.e., a HUGE understatement. AI has the potential to cause seismic changes to everything around us from our jobs to the way we socialize. AI is also causing a skills shortage and is going to be massively in demand in the coming years in fields like risk management and cybersecurity. If you are in either of these fields and do not upgrade your skills, then you will get left behind.

NOW is the best time to invest in your skills and learn AI cybersecurity and risk

However, I know that AI can seem like a very scary and intimidating technology to newcomers

with all the current literature seemingly geared towards people who are already well versed in this field with phrases like neural networks, deep learning, etc. casually thrown around!

A big misconception about AI is that you need to know the deep internal details of how Machine Learning technologies work to secure or govern them, but this is simply not true. Having a fundamental understanding of AI technologies is sufficient for you to start working on securing them. I wrote this book to save people the time and trouble I went through when learning about Artificial Intelligence risks and cyber-security and to share some of the beneficial knowledge I gathered in my journey.

If you want to future-proof your career for the next 10 years and are looking for a good area to invest your time and knowledge in, then you cannot go wrong with Artificial Intelligence

Don't you already have a course on this?

The answer is **YES,** and I do very much have a course on this which you can access below:

https://www.udemy.com/course/artificial-intelligence-ai-governance-and-cyber-security/

However, this book compliments the course and does not replace it. There will be overlap in the topics that are taught but some people prefer reading books while some prefer to listen and follow along so it depends on which style of learning you prefer (or you can do both and get the best of both worlds)

Feedback is always appreciated

I always appreciate feedback whether it is positive or negative as that will help me improve as a writer and make better material. Please leave a review and let me know what you liked and where you think it can be improved.

1

Understanding the impact of AI

This is an introductory chapter that gives you an overview of AI and why it is so important. We are going to set down the foundational knowledge you need about AI and cover two major topics in this section

- Artificial Intelligence and what it means.
- The impact of AI on human civilization

What is Artificial Intelligence?

Let's do a quick exercise as we start. What is the first thing that comes into your mind when asked about Artificial Intelligence? Is it one of the below:

- Computers that can think for themselves.
- Self-driving cars
- Robots working in factories
- **Machines suddenly becoming sentient and taking over the world resulting in the end of humanity**

The last one might seem like a joke, but you will be surprised to know how many people think about doomsday scenarios every time AI is mentioned. It's not surprising given how much we have focused on the concept of self-aware computers in pop culture and how many books, movies, and shows you can find on that topic. Thankfully machines haven't taken over the world (yet) but they have helped improve our lives considerably over the years thanks to Artificial Intelligence.

So how do you define AI?

If we can trace the definition back to a single person, then that would be John McCarthy who was an American computer scientist fondly remembered as the "**Father of Artificial Intelligence**". The term is usually attributed to him in his 1955 proposal at the famous Dartmouth conference where he referred to it as:

"The Science and Engineering of making intelligent machines"

What does that mean you ask? Well, let me tell you a little secret. In most cases, computers are dumb. I mean not dumb in the sense that they have a low IQ but dumb in the sense that they can only do what you explicitly tell them to do and do not possess the ability to make independent

decisions by themselves.

Unlike human beings who learn from their mistakes and experience, computers will do exactly what you tell them to do every time (which is why we use them). The illusion of intelligence that you get from computers is usually just instructions that have been programmed into them by humans. If you time-traveled back a century or two and showed an electronic calculator to people, then they would think that was an intelligent machine!

But we know a calculator just has hard-coded instructions inside it and it expects a specific input based on which it will give a specific output. It is unable to break out the information it is expecting and expected to give and is not going to suddenly recommend newer and more efficient ways of doing something.

The following diagram briefly shows the difference between a human brain and your traditional software program:

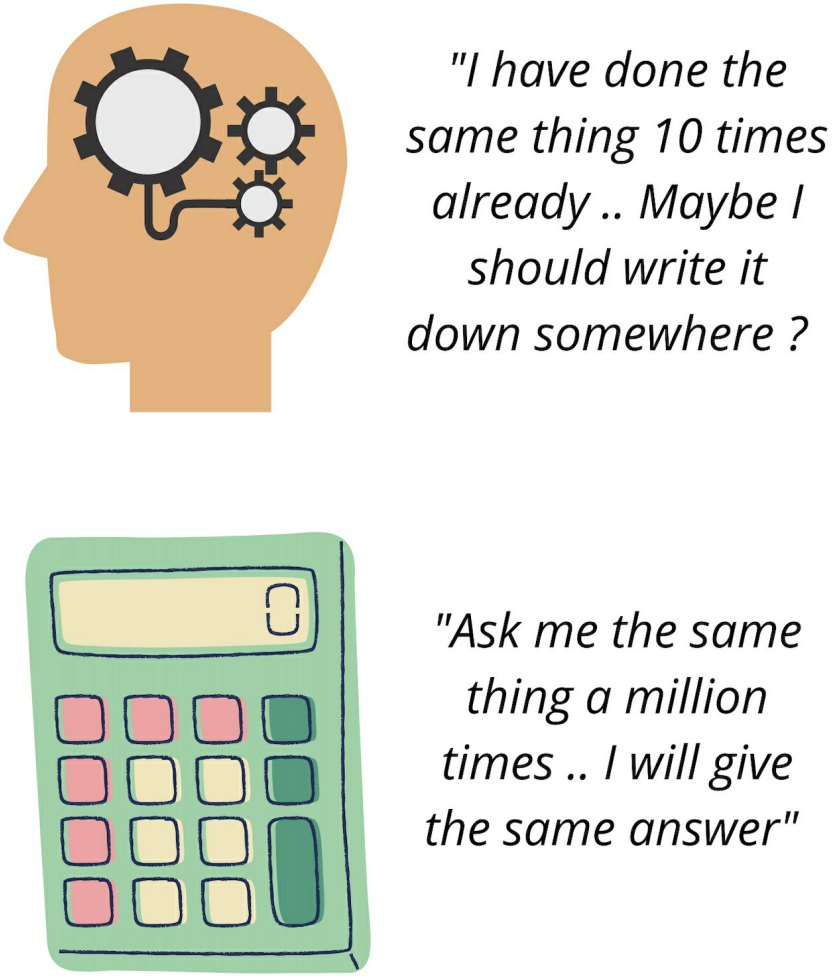


Figure 1.1: Simple difference between our brain and a computer program

Keeping that in mind, I hope now you can start to appreciate why AI is considered to be a big deal as it gives computer systems the ability to **LEARN** and do stuff that human beings can do such as recognizing speech, images, making decisions, etc. **WITHOUT** being explicitly programmed to do so. For example, in an AI-based facial recognition system, you would not need to feed it images of every single person in the world for it to start recognizing people, instead, it would build up its knowledge over time and become more and more intelligent.

This is the main feature that distinguishes an AI system from your regular software program. It is also desperately needed for today's world as the amount of data that is being generated by both humans and machines is far outpacing our ability to absorb and interpret it.

Let's take an example of social media networks which must monitor and stop any potential hate speech from being used on their platforms. With billions of users at any point in time generating billions of texts, you can imagine the cost of doing this manually! By shifting some of these tasks to AI-based systems which automatically flag and remove inappropriate words and content; we can truly take the benefits of technology to the next level.

The Impact of Artificial Intelligence

As an old wise man said one time (or maybe I heard it in a movie)

"To understand where you are going, first look where you have come from "

Artificial Intelligence has been called the Fourth Industrial Revolution due to its potential to transform our lives with researchers estimating that most human jobs could be offloaded to AI systems by 2030. But is AI such a big deal or just hyperbole?

Well, to appreciate AI and its impact you must understand how we have reached this point in history where stuff that seemed to happen only in sci-fi movies (like self-driving cars) is now being taken for granted. AI as a concept has been around since the 50s but only in recent years has it started to catch on and we have started to see its application in all walks of life.

The First Three Industrial Revolutions

To put things in context let's have a short history lesson about the previous industrial revolutions. If we look at the glorious history of humans, we have had three major industrial revolutions which brought about a massive change at a social level and changed the way people worked, lived their lives, and even how cities were structured!

A few centuries ago, people's lives revolved around the farm where work was done at home or in the fields. Goods were made locally and then sold off to local markets which is how people earned their living. All of that changed with the invention of the steam engine which could power factory machinery in industries like textile manufacturing and agriculture. Factories started to spread and job opportunities became plentiful for both men, women, and even children with child labor being an unfortunate reality of this time.

During this first Industrial revolution, society started to fundamentally change as people

migrated from the farms to the cities and into the factories.

Figure 1.1: Work-life before industrialization

However, factory life was far from ideal with people laboring under long hours and unsafe conditions. Along came the Second industrial revolution which brought us things like steel and electricity. These allowed factories to introduce automated machines removing the need for manual labor and increase output to unheard levels. Once again things like the automated assembly line changed the course of human history and how we worked. This also resulted in a lot of people who moved from the farms to the cities losing their jobs as their skills became redundant.

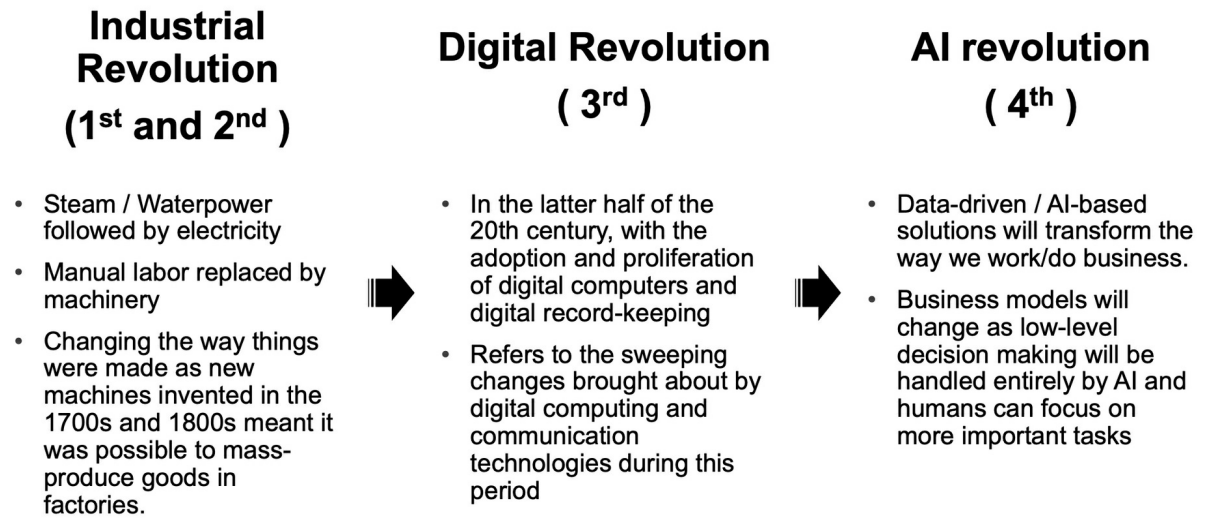
Figure 1.2: From the farms to the factory

The third industrial revolution is one we are still feeling the impacts of with the introduction of digitization and the invention of the internet. If you are reading this book from your smartphone or iPad, then congratulations as you are reaping the benefits of this revolution in which the internet completely changed how people communicated and how business models operated. The fact that most people living nowadays cannot imagine a world without the internet or smartphones goes a long way to show just how dependent we have become upon them in our daily lives.

Welcome to the Fourth Industrial Revolution

Now that we have seen the profound disruptions that happened in the previous era, we can have an appreciation for how major a change AI is having on us as a society. AI is one of the technologies that are being touted as part of the **Fourth Industrial Revolution** along with things like quantum computing, robotics, and the Internet of Things (IoT) that are blurring the line between the physical and digital worlds.

However, one of the major differences between the 4th and the previous revolution is the speed at which things are evolving. Technology is evolving at an amazing (and frankly terrifying) rate which is why it is so important that things like AI are governed and the risks surrounding them are understood.



Why now?

One question you might have been asking is how come the AI hype train is only now catching up to everyone and we see its impact everywhere from the intelligent assistants on our phones to the movie recommendations on Netflix? What has changed now so that every other company is boasting of using "**AI-driven services**" and every other application says, "**Powered by AI**"?

Well, the reasons are threefold:

1. **AI needs computing power and lots of it!** This was simply too cost-prohibitive until recently with the spread of cloud computing. The cloud has given us access to levels of computing power that was simply not possible before and hence the explosion of AI-driven services that can tap into this power.

2. **AI needs storage and lots of it.** Storage costs have dropped significantly (does anyone remember floppy disks?) and again the cloud has given us access to zettabytes of data which AI systems need to build up their machine learning models effectively (more on this in the next chapter)

3. Everyone from tech startups to governments has realized the potential of AI and are investing billions into making sure they do not get left behind in the AI race. AI is now a competitive advantage, and everyone wants their piece of the pie

Chapter summary

That was a summary of the AI revolution to give you an idea of how revolutionary this technology is and how it is not a trend that will fade away over time. AI is going to cause huge changes both at an individual and societal level

Below are the topics we covered:

- What AI is
- The impact of previous industrial revolutions and where AI fits
- Why AI is suddenly becoming so common

2

Machine Learning - The engine that drives AI

To fully understand AI and its associated risks; we need to know about Machine Learning which is the driving force behind most AI services.

We are going to cover two major topics in this section

- Machine Learning and its definition
- How does Machine Learning work?

What is Machine Learning?

A lot of times the terms "**Machine Learning**" and "**Artificial Intelligence**" are interchangeably used together which can be a bit misleading. Machine Learning is a branch of AI and easily the most popular one. You **can** create AI without machine learning but right now it is the primary way in which AI systems are made and pretty much the "**engine**" that drives AI.

So, what exactly is machine learning?

It is defined as the ability for machines to learn from data to solve a task without being explicitly programmed to do so. It is easily the most mature of all the subfields of AI and the one that we use the most in our daily lives. It mimics the human ability to learn from past experiences and apply it to future events as we see in the following diagram:

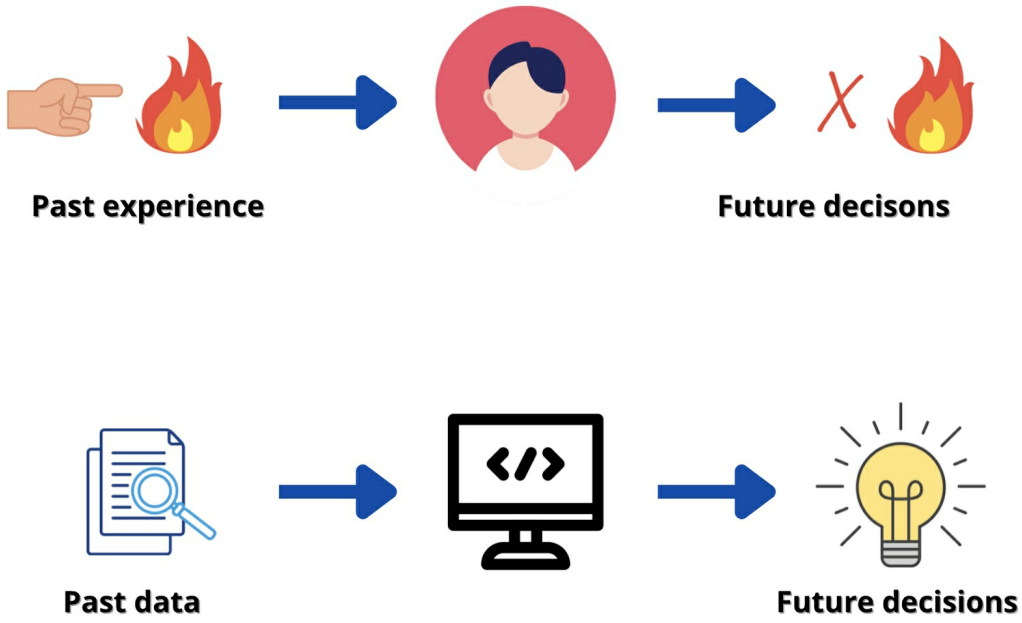


Figure 2.1: How machine learning resembles human learning

To fully understand it, let's look at how traditional computer programs work vs Machine Learning. Normal computer programs take data as input which is processed within software using an algorithm and gives an output. This has traditionally been how computer programs have solved problems for us humans as we can see:

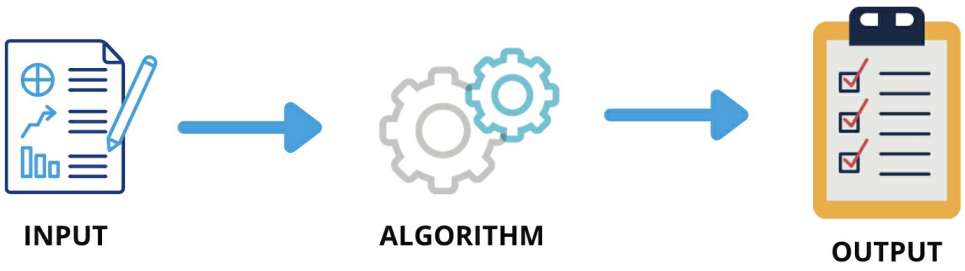


Figure 2.2: A traditional computer program

Machine Learning is slightly different in that you feed the input AND the output into an algorithm to create a program or a "**model**". This model is then used to create predictions and insights about future data as we can see in the diagram:

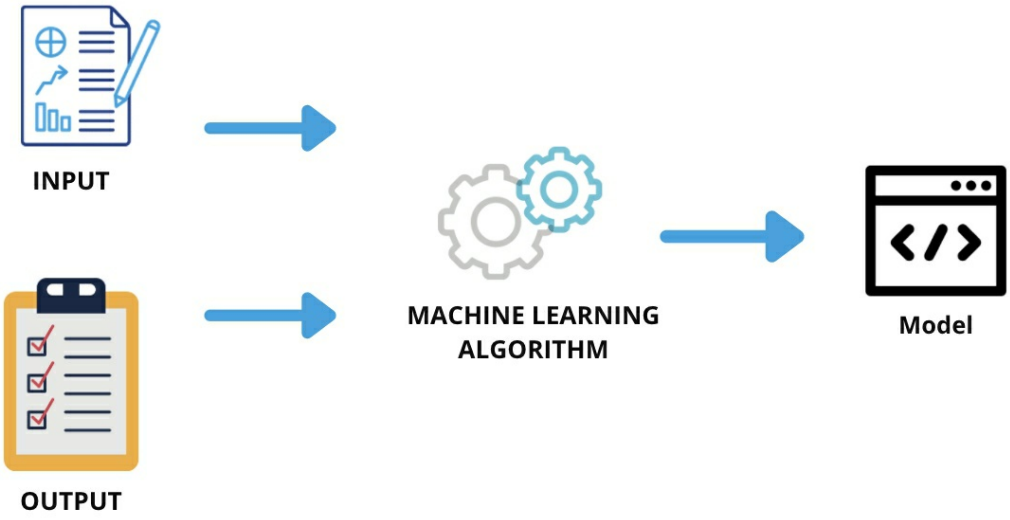


Figure 2.3: Machine Learning

How does Machine learning work?

So, what is the point of putting both the input and the output into this algorithm? Well, the plan is that once the algorithm has gathered sufficient data and seen its output, it will start seeing patterns that will enable it to make future decisions. For example, by feeding a Machine Learning algorithm millions of pictures of felines and canines, it will start to distinguish between cats and dogs by itself without being told to i.e., apply what it "learned" to new and unseen data

If we were to break this down into steps, then machine learning would look like the below:

1. We gather lots and lots of data
2. We create an algorithm to understand that data
3. We train that algorithm using that data in step 1
4. Our software will slowly learn and build a "model" that it will use to predict future results that it has not yet been fed.
5. Now give it new data and see if the model predicted the result correctly.
6. If the results are not correct, then re-train the algorithm multiple numbers of times until the desired output is found.

The machine is basically "learning" on its own and the results will become more and over accurate over time as we see in the following diagram

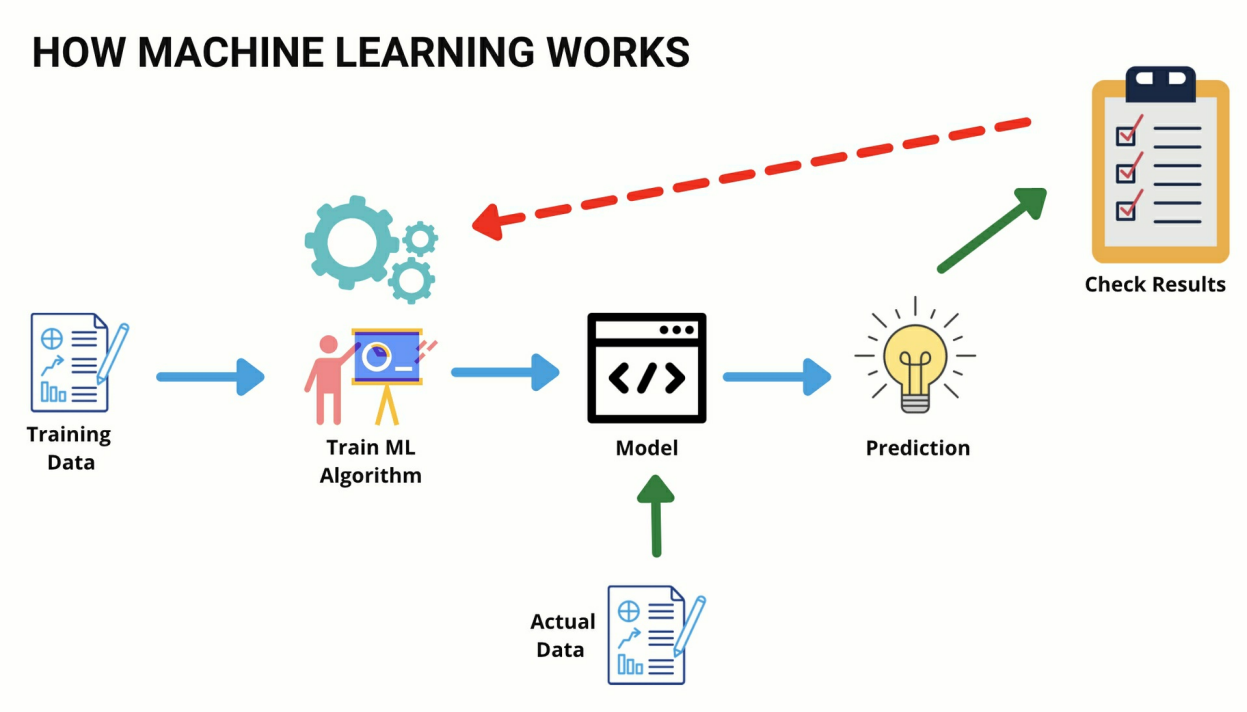


Figure 2.4: Steps of Machine Learning

Types of Machine Learning

Machine learning comes in two flavors which are called supervised and unsupervised with the main difference being the type of data that is provided to the model.

Supervised Machine Learning: The model is "**taught**" via data that is clearly labeled i.e., the data is clean and can be easily understood and doesn't leave much room for ambiguity. So, a machine learning model being taught to distinguish between cats and dogs will be fed data in which the two animals are clearly labeled as such. This is like how children are taught the names of different shapes by teachers, i.e., each shape is given a name and children are taught to distinguish between them

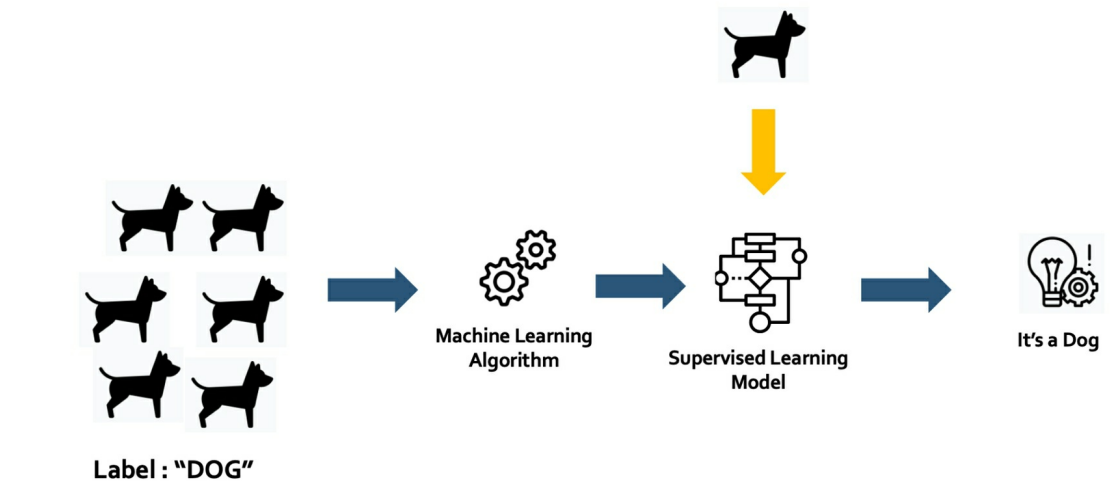


Figure 2.3: Supervised Machine Learning

Unsupervised Machine Learning: The model learns via data that does not have any labels and no guidance is given. Instead, the machine explores the data and learns to identify common patterns to distinguish the data types. So, in this scenario, the model will be fed pictures of both cats and dogs without the data being labeled as to which animal is which. The model will itself identify distinguishing features and use them to predict future results.

This is similar to how human beings learn by experience and trial and error.

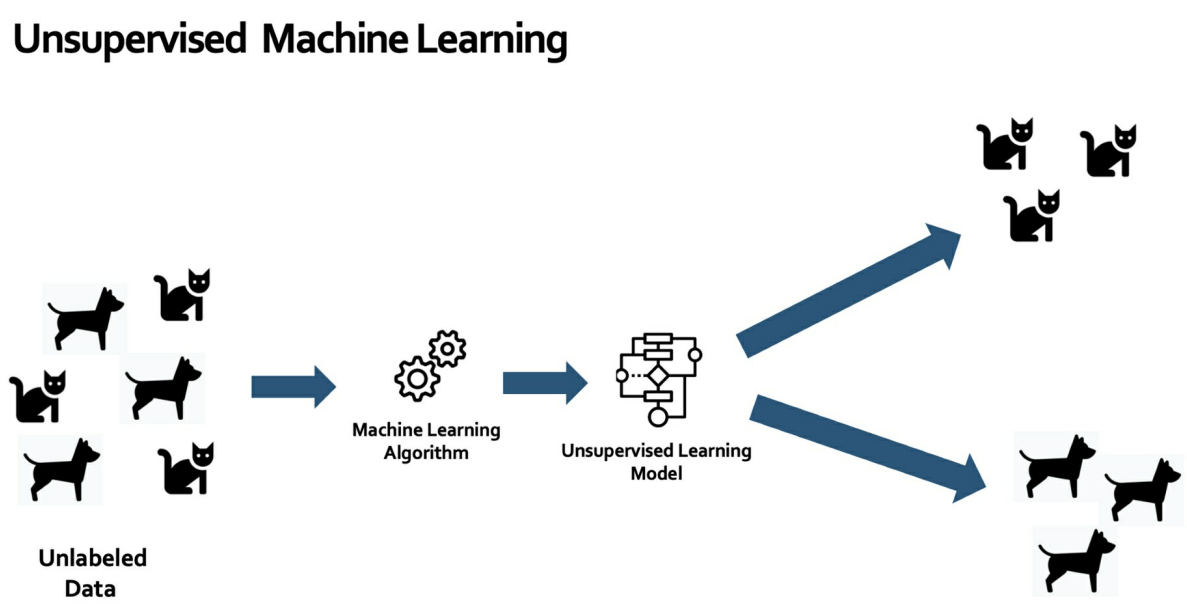


Figure 2.5: Unsupervised Machine Learning

Note: there are also other types of machine learnings present like semi-supervised and reinforcement learning but we do not need to cover them to understand the basics of how machine learning works

Chapter Summary

In this chapter we understood how a machine learning model learns from its data and how it reaches its decisions. This is important as we will see later when discussing cyber-security of AI systems in chapter 7. Some of the most dangerous attacks on Artificial Intelligence attempt to pollute this training data so that the machine learning model is not able to predict data correctly or it reaches a decision which is suited to what the attacker wants. Understanding the process enables us to see the risks behind the same.

In summary, we learned:

- What Machine Learning is
- How it works and replicates human learning
- The different types of machine learning

3

AI governance and risk management

Now that we have set the foundation of what AI is, its importance, and how it works; it's time to get into the need for governing and regulating AI. The first question that might come into your head is WHY? No one likes regulations and standards and the red tape that comes with them so why should we regulate such a revolutionary technology and put hurdles in front of innovation?

Well, the sad reality is that AI, despite all the good that it can do; also introduces new risks which were not there before and there will always be people with mal-intent happy to exploit those risks for their nefarious purposes. Take the example of the Internet which changed the way humans interacted with each other and spread information in a way that was simply impossible before. However, it also opened the doors to cybercrime which is now a multi-billion-dollar industry with incidents like ransomware, denial of service attacks, etc. almost a daily occurrence.

As we will find out, AI can create as many problems as it's going to solve! If you follow technology news or just read about AI on social media you will surely come across headlines about AI potentially causing massive job losses, privacy issues, and of course killer robots!

Let's separate the reality from hyperbole and take a better look at the risks which AI can (and will) cause:

Privacy Risks

One of the biggest blockers to the public acceptance of AI has been the privacy risks that these systems can introduce. As we saw in the previous chapter, data is the lifeline of AI systems and how they become more accurate over time. So, what stops an AI system from collecting and misusing sensitive information such as biometrics, health records, credit card information, etc.?

How do we even know that the AI system is using the data in the way it was envisaged to do?

As more and more countries move towards adopting AI-based technologies such as facial recognition; privacy will become more and more of a concern especially when it becomes commonplace in offices, schools, and other locations. Companies need huge amounts of data for their AI systems to work and unfortunately, they can indulge in unethical practices by collecting this data without consent. Along with the privacy problems, this can also result in huge legal issues for the company when it comes to light that their AI systems are being powered by unethically collected data.

We already have seen cases like the infamous Cambridge Analytica scandal in the 2016 U.S. Presidential election where it was revealed that the personal data of millions of Facebook users was collected without consent by a British firm so they could be targeted for political advertising. Another example was the US company ClearView AI which collected pictures of adults and children in Canada without their consent. Such misuse of AI technologies can lead to

mass surveillance of citizens which is a violation of their privacy rights.

Another deeply disturbing privacy risk is the growing popularity of "deep fake" technology which allows people to create disturbingly real likenesses of real-life personalities such as Tom Cruise and Morgan Freeman. Such videos are virtually indistinguishable from the real persons they are imitating and have serious repercussions for those in the security field.

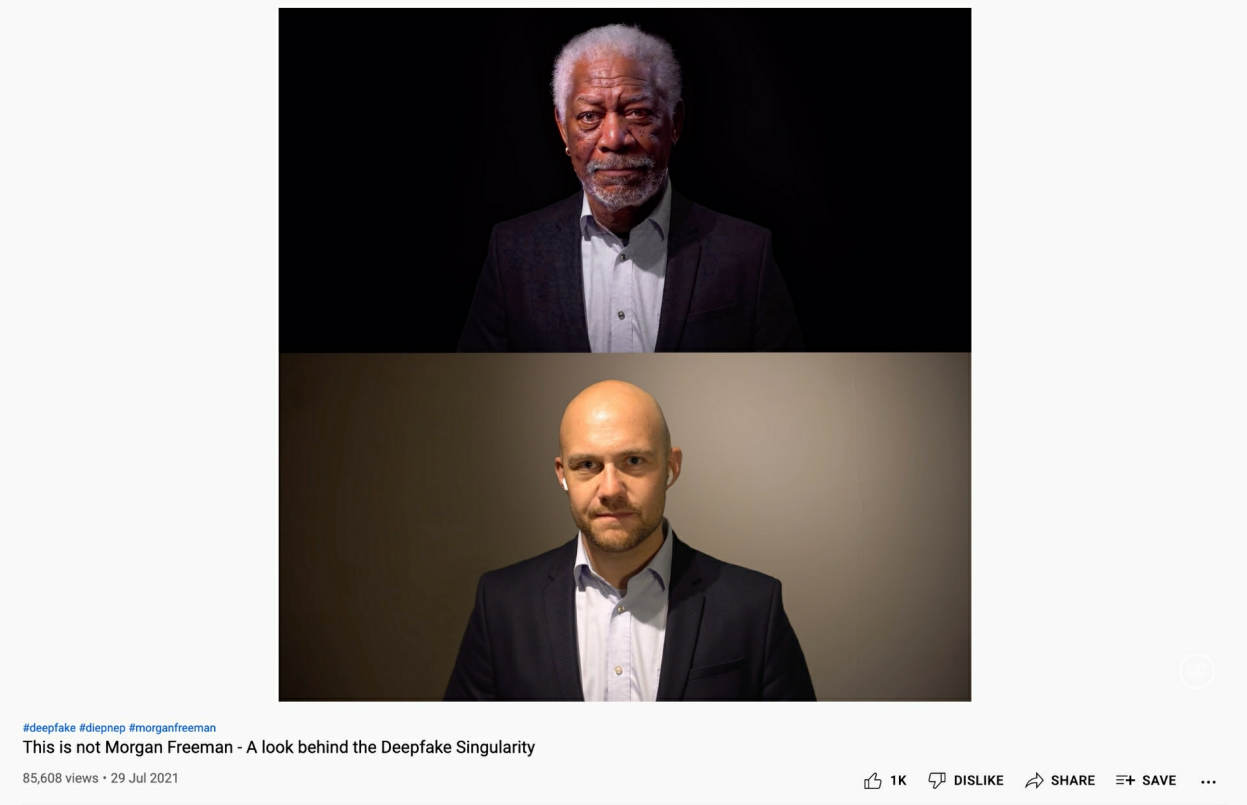


Figure 3.1: This is not Morgan Freeman video

At the rate at which technology is evolving, what is to stop someone with malicious intent from using deep fake technology to manufacture fake audio or video clips of someone they don't like? For example, framing a politician as saying or doing something unethical to ruin their reputation? The implications of such technology during elections can be disastrous given the rate at which false news can now be spread due to social media. These privacy risks are why legislation to govern and control AI is so important as we will see in the coming chapter

Job Disruptions

One of the most valid concerns surrounding AI has been its potential of disrupting the job market and resulting in widespread job losses. This is one risk that seems justified given AI's potential for automation and replacing the need for human beings to do mundane tasks. A study from McKinsey Global Institute stated that robots and automation could replace around 30 percent of

the world's labor by 2030. So, does that mean most of us will be out of jobs in a decade or so?

Well, thankfully the answer is a big fat **NO**.

In every technology disruption and the earlier industrial revolutions, every such leap forward has resulted in more jobs being created than removed. You will always need human beings to make sure things are running along smoothly and the risk of AI taking away all our jobs is simply not rational. There is already a huge shortage of AI and machine learning experts in the market and the demand is expected to dramatically increase in the coming years with AI expected to create an economic impact of around $15.7 trillion by 2030.

However, that does not mean that we do not need to level up and make sure our skills are not made obsolete in the coming revolution. People will need to reskill and be willing to adapt to change otherwise the coming years will become very painful.

AI-assisted Cybercrime

AI like any other technology has the potential of being used for both good and evil purposes. Cyber-criminals are not blind to the potential of offloading mundane tasks onto AI giving them the ability to come up with even more new ways of committing cyber-crime. Machine Learning models can be trained on hacking techniques or socially engineering humans and learn at a much faster rate than any human hacker. Similarly, DDOS or ransomware attacks could reach a new level of danger with AI machines taking over the tasks from humans.

Additionally, AI systems themselves are in danger of being compromised so that attackers can tamper with their decision-making processes. We will take a detailed look at these risks in Chapter 7

Cyber-attacks on AI systems

As AI systems become more and more involved in critical decision making, attackers will start targeting the very algorithm that facilitates this decision-making process. This is easily one of the biggest and most ignored risks present when adopting AI. The reality is that there is not enough awareness of how AI systems work within the cyber-security community (one of the reasons for writing this book!) and hence these risks are almost completely ignored when doing security reviews of AI systems.

Cyber-attacks unique to AI can be a blind spot for many companies and completely bypass traditional defenses like how application-level attacks (SQL injection, cross-site scripting, etc.) started bypassing network defenses in the early 2000s.

Trolling & Misuse of AI

AI has the danger of being misused in a way that was never envisioned by its creators by people for their amusement or more sinister purposes. Take the example of Tay which was an AI

Twitter bot that Microsoft released in 2016. Targeted toward the 18-24 age group, the bot was designed to learn from conversations and become smarter over time. Unfortunately, a group of people realized you could feed the bot racist and offensive information which it would start retweeting resulting in Microsoft having to take the racist bot offline after a few hours! A few of the more colorful tweets which Tay put out are below:

Figure 3.2: Tay being tricked into tweeting offensive information

Figure 3.3: More examples of Tay being offensive

Microsoft admitted to not realizing that people could target their technology with malicious intent and vowed to make sure future bots had these controls in place. That was a slightly harmless example so let's look at something way scarier

While we scoffed at the earlier mention of "**killer robots**" there is one very serious application of AI systems and that is autonomous weapons. Referred to as weapons that can target and engage with the enemies without any human intervention; autonomous weapons are similar to armed drones but much more advanced level.

While some arguments for autonomous weapons have been made such as reducing casualties by removing humans on the battlefield, they do introduce serious ethical and security concerns. If a malicious party could potentially compromise an AI-based missile system, then what is to stop them from being sent back to their origin? Over 30,000 AI and robotic scientists highlighted this risk and signed an open letter which you can read here in which they stated

"If any major military power pushes ahead with AI weapon development, a global arms race is virtually inevitable, and the endpoint of this technological trajectory is obvious: autonomous weapons will become the Kalashnikovs of tomorrow. Unlike nuclear weapons, they require no costly or hard-to-obtain raw materials, so they will become ubiquitous and cheap for all significant military powers to mass-produce"

Bias and prejudices in AI algorithms

We mentioned earlier that one of the biggest benefits of AI will be offloading low-level decision making to machine learning models to save time and money. However, as we replace human decision-making with machine learning algorithms, we may assume that these models do not carry over human prejudices and biases. We know human beings have biases that can cause us to treat other people unfairly but how can AI systems be prejudiced?

Well, the sad fact is that machine learning algorithms are trained on real-world data and that data still has the potential to have biases within it which can unintentionally get carried over causing the model to prefer one group of people over another. These models can then become biased against a particular gender, age, or race resulting in a real-life impact on people's lives, health, and wealth.

What if someone is denied healthcare or a bank loan based on faulty or biased decision-making by an AI system and is not left with any way to challenge this decision? Without human interaction or some way to give context, this can have serious consequences on someone's life.

Let's take an example of a healthcare machine learning model in the U.S. which was found to be racially biased due to the data that was used to train it. The model was being used to predict which patients would benefit from more specialized care which is typically given to people who are chronically ill. The model would predict this based on the patient's previous spending on health care. Unfortunately, black patients despite being considerably sicker than white patients were not given high-risk scores due to several ingrained issues within the healthcare system itself. Due to these biases being carried forward, millions of black people were denied access to the health care they otherwise should have been given.

If the data used to train the model had not used cost as a metric, then this bias could have been avoided and “fairness” present in the algorithm. We will see in Chapter 6 what principles should be present in machine learning algorithms to avoid such situations from happening.

Chapter Summary

In this chapter we saw some of the dark side of AI usage and the negative consequences that AI can unintentionally introduce. This is crucial to understand as AI risks are very much a developing subject and new areas are being discovered regularly.

We covered the below:

1. What are some of the key risks that AI can introduce?
2. How AI can be misused intentionally and unintentionally
3. How wrong AI decisions can harm people's lives

4

Artificial Intelligence laws and regulations

In the last chapter, we saw what risks AI can introduce and why it is so important to have controls in place to stop the accidental or deliberate misuse of AI. The sad fact is that companies usually prioritize profit over controls and will try to reduce costs wherever possible. If it costs too much to secure an AI system, then the company might simply decide not to do it!

This is where regulations come in to enforce compliance to a minimum set of standards that everyone must follow

Regulations are important as they make corporations accountable to the authorities and help to ensure that AI as a technology has minimum safeguards put in place across the board. The consequences of not complying can be regulatory fines or even the removal of an AI system from the market. On the other side, complying with the regulations can help the company market their product as being "**fully compliant**" giving them a competitive advantage over others.

Global AI regulatory landscape

Organizations in the business of making AI systems have historically relied on self-regulation without much oversight. There were no specific regulations in place and AI systems came under the umbrella of other regulations such as **data** or **consumer protection**. Seeing the potential risks involved, governments across the world are rising to the challenge and putting in new regulations to ensure AI risks are identified and mitigated appropriately. Several legislations are being passed in the U.S, China, and other countries to take the lead in the AI race.

This can have good and bad consequences as the ever-growing list of policies and laws for AI systems can cause companies to be hesitant about adopting AI given the risk of not complying with a required regulation. The alternative unfortunately is not to adopt AI and get left behind by their competitors.

The most important regulation by far and the one expected to have the most impact around the world comes from the European Commission which in April 2021 issued a proposal for a new act to regulate AI. Like how it set the stage for global data privacy laws with the General Data protection regulation (GDPR) in 2018, this act is expected to have wide-reaching implications across the world. EU rules usually end up setting the standard for the rest of the world because of all the companies that work in it, so we can expect this act to become a blueprint for other countries to derive their own AI laws.

The EU AI act - What you need to know

As the world's first concrete proposal for regulating artificial intelligence (AI), the EU's draft AI Regulation is going to have a huge impact on the debate on AI and how companies are going to

adopt AI in the future. The act itself takes a risk-based approach to regulating Artificial Intelligence and categorizes AI systems as follows:

1. Unacceptable risk
2. High risk
3. Limited risk
4. Low risk

The basic risk-based principle is that the higher the risk that the AI system poses, the more obligations on the company to prove to regulators how the system has been built and how it will be used. Systems labeled as **Unacceptable AI** are simply banned such as those systems that use facial recognition technologies, systems used for social scoring that rank people based on their trustworthiness, and systems that manipulate people or exploit vulnerabilities of specific groups

The bulk of the regulation focuses on **high-risk AI systems** which must comply with a deep set of technical, monitoring and compliance requirements which we will investigate in detail shortly. Systems classified as limited risk are subject to transparency obligations while the remaining minimal risk systems do not have obligations but **are** recommended to put in codes of conduct to make sure good practices are followed.

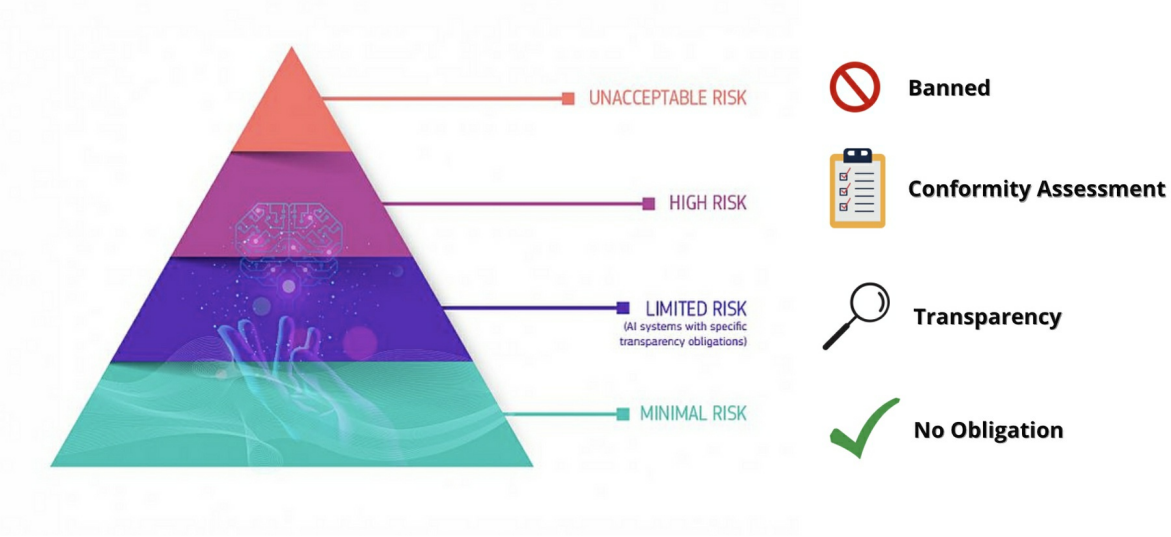


Figure 4.1: How the AI act categorizes requirements based on risk

"High Risk" AI systems under the proposed EU act

The act identifies AI systems as being "**high risk**" when they can potentially endanger the life or health of persons or their fundamental rights. The act has a list of high-risk AI systems some of which are mentioned below:

1. critical infrastructure.
2. education and vocational training.
3. employment.
4. access to and enjoyment of essential private services and public services and benefits.
5. immigration, asylum, and border control management; and
6. the administration of justice and democratic processes.

The key requirement for high-risk AI systems will be to undergo a **conformity assessment**, be registered with the EU in a database, and sign a declaration confirming their conformity. Think of a conformity assessment as an audit in which the AI system will be checked against the requirements of the regulation which are listed below:

- the implementation of a risk-management system.
- technical documentation and record-keeping.
- transparency.
- human oversight.
- **cybersecurity**.
- data quality.
- post-market monitoring; and
- conformity assessments
- and reporting obligations.

These audits can be done as self-assessments by the company making the AI or an assessment by a third party (currently only AI used in biometric systems need to undergo third-party conformity assessments while others can just go the self-assessment route). If the system gets changed after the assessment, then the process must be re-done.

The following diagram illustrates this process:

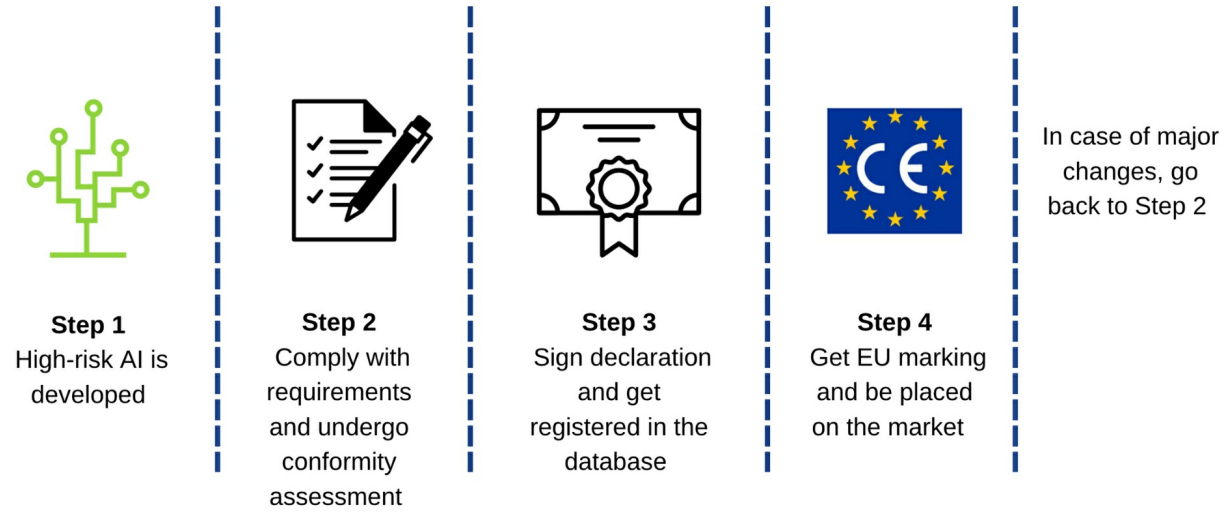


Figure 4.2: Steps for high-risk AI systems to follow under new act

Once the assessment is passed, the result will be a nice CE logo on your product which confirms that it is now ready to enter the market for EU customers.

Who must comply?

Like the GDPR, the scope of the regulation is not just limited to EU also as like the GDPR the law can cross borders and apply to:

- Providers who place AI systems on the market or put them into service in the EU.
- Users of AI systems located in the EU.
- Providers and Users of AI systems located in third countries, where the outputs of the AI system are used in the EU.

How should you prepare?

If you have ever implemented the EU's GDPR then you would understand the EU does not mess around when it comes to non-compliance and can enforce serious fines for breaking its rules. The new AI act also follows this trend and fines for using prohibited AI systems (those presenting unacceptable risks) can go up to **€30 million or 6 percent of annual global revenue** (way above the maximum fine under the GDPR). Companies who provide misleading information to authorities can also get fined up to a maximum penalty of **€10 million or 2 percent of global revenue**.

If your AI system is coming under the scope of the new act, then it is not something to be taken lightly.

While some have criticized the new EU regulation for being too restrictive resulting in Europe possibly falling behind other nations in the AI race; chances are high that this act **will get** enforced so it is best to start preparations now rather than leave it for later. Taking concrete actions now will ensure you are on the right side of this regulation when it gets enforced.

The first and most effective step would be to conduct a gap assessment against this regulation and see where your organization falls and what you must do to be fully compliant. Your company might not potentially have the relevant expertise to conduct these assessments so you would need to reach out to third-party excerpts who can guide you. Another step would be to create an **AI governance framework** in your organization to manage and mitigate AI risks as they appear. We will read more about this in the coming chapter.

Chapter Summary

In this chapter we learned about the regulatory landscape covering AI and the upcoming EU AI regulation which is expected to have the most impact on AI usage across the globe.

We covered the below topics:

- AI regulations and why they are needed
- The new AI regulation and its risk-based approach
- The requirements for high-risk AI systems
- How to prepare for the coming regulation

5

Creating an AI governance framework

In this chapter, we start forming a framework for mitigating the AI risks we talked about in earlier chapters. If your company is planning to use AI systems as a strategic and/or competitive advantage in the long term, then having an overarching AI governance framework is going to be crucial.

AI regulation as we saw is going to be a very powerful tool in enforcing ethical usage of AI, but it takes some time to enact. This means companies need to take the lead and put in frameworks to mitigate the unique governance and security risks that AI systems pose.

An added benefit will be that when the regulations do come in, those companies who proactively implemented governance frameworks for AI systems will be at a distinct advantage and will find it much easier to comply with the new rules.

What makes a good AI governance framework?

AI is becoming more and more viable and easier for companies to adopt and user-friendly AI software which requires little understanding of the underlying models is becoming quite popular. Apart from internal projects, vendor-driven software also may have AI components which may introduce risks if they are not mitigated in time. A company could potentially buy credit scoring software from a vendor without knowing that there is an underlying AI model which was not created properly and potentially discriminates against certain people!

To solve these challenges for risk management professionals, companies need to create an AI governance framework so that a structured system is put in place to **identify, mitigate, and track** these risks to closure.

A governance framework is a structured way of making sure things work properly and in compliance with industry regulations and guidelines. An effective AI governance framework will ensure that the company mitigates the risks of AI systems in a structured and repeatable way. This means that AI systems will not be a blind spot for the company and instead be approved, formalized, and assessed to make sure they are not introducing any unforeseen risks.

An effective framework will be:

- **Technology agnostic**: It does not care about any software technology or provider and instead will apply the same principles regardless of the technology
- **Algorithm agnostic**: It does not care about the underlying technicalities of the AI algorithm but cares about how it has been designed, how its data has been captured and if it is following “AI trust” principles (more on that shortly)

Key Components of an AI Governance framework

While a governance framework can change depending on the nature of the business and its regulatory framework some aspects will be common across industries.

The high-level components of an AI governance framework are as follows:

Figure 5.1: Key components of an AI governance framework

AI and Machine Learning Policy:

Anyone with experience in implementing a governance framework will know that the hardest part is always changing a company's ingrained culture. Companies have a way of doing things that develop over time and introducing new controls is always met with initial resistance. One of the best ways to drive change by management is to formalize a policy that clearly articulates the company's vision about how AI will be ethically used within the company and how AI-associated risks will be mitigated. A **high-level policy** will set down the tone of how AI usage will be controlled across the company and formalize responsibilities for AI usage and the general principles which AI systems must comply with.

In a nutshell, an AI policy informs everyone **who can do what** and where the buck stops if it is found out that AI systems were made in a non-compliant manner. It will also form the basis for the other components which follow.

AI Governance committee:

Another key aspect is a cross-functional governance team that oversees AI systems and makes go/no-go decisions on new AI initiatives. Management will have to identify key people across

the company and empower them concerning AI controls by making them part of this committee.

If a new AI system is being developed in a market that might put your company at risk of not complying with local regulations, then this committee is where the decision to not proceed will be made. By making this committee composed of representatives from multiple departments; it ensures that differing viewpoints are taken, and all the stakeholder input is recorded.

NOTE: Training will have to be a mandatory part before enrolling members as they will need to understand AI risks and how to identify new risks in any upcoming AI models.

At a minimum the committee should have representations from

- **Legal** - To make sure no legal implications are present in any new AI project
- **Cyber-Security** - Usually best placed to flag any security risks in AI systems
- **Technology** - The team that drives adoption of new AI technologies and is responsible for monitoring / managing the underlying infrastructure.
- **Data Science** - The people who are working with the data powering the AI systems.
- **Business** - The driving force of most AI initiatives.
- **Audit and Risk** - Independent members are a necessary part of this committee to ensure effective governance
- Chaired by a member of the Executive level committee

AI risk management framework

An output of the AI policy; a **framework** to identify risks in business-critical AI systems will be set up and be owned by a designated unit. Like a risk management unit in a regular company, this unit will create mitigation strategies to identify and fix AI risks around bias, data collection, cyber-security, resiliency, etc.

The framework consists of several key components which are as follows:

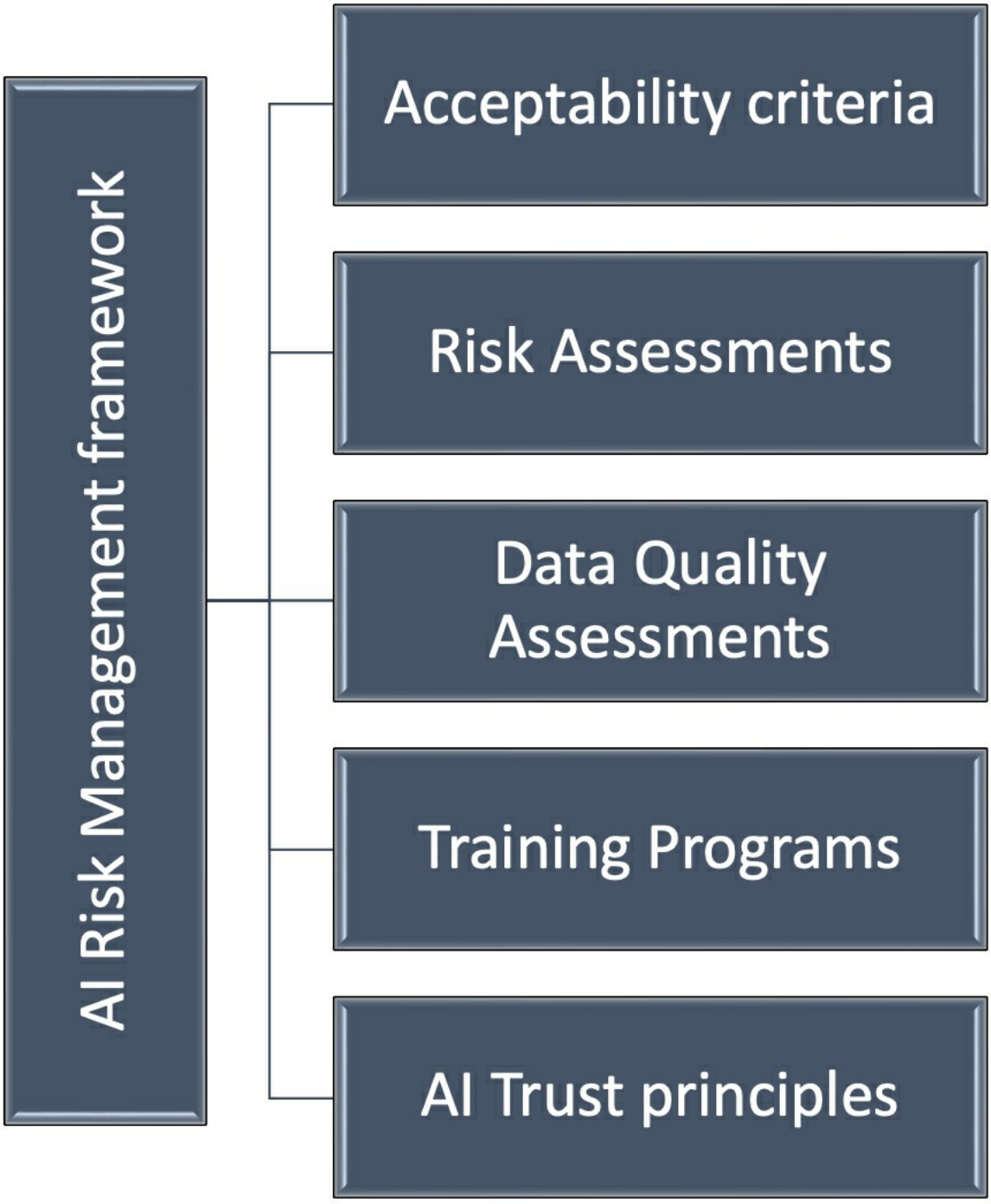


Figure 5.2: AI risk management framework

- **Acceptability criteria for AI systems** surrounding their risk, security, and control. This can be something as simple as a checklist that AI systems must comply with before they can be considered good to go and deployed in the market or a detailed risk procedure. However, what must be ensured is that it covers the entire model

lifecycle from data collection, model selection, deployment, monitoring, etc. Each of these phases has different risks that must be treated separately and must be highlighted early on.

- **Risk assessments for "high risk" AI system**s either by themselves or by getting in touch with a qualified third party who can do an in-depth assessment of their security and governance. These risks will need to be tracked and monitored to closure.
- **Assessments of the data quality** for AI systems to make sure that the data on which they are being trained is suitable and matches the use case. Remember that the machine learning model's underlying logic can change over time which means these assessments have to be done regularly
- **Training programs** to educate data scientists, technology, business, and cyber-security professionals on the risks around AI systems and how to own them. This might be a challenge at first and require outside help but over time the maturity will increase, and teams will become capable of identifying and appropriately mitigating risks due to an ongoing awareness drive.
- **AI trust principles** are consistently enforced across all AI projects. Let us look at what these principles are.

AI Trust principles

For AI systems to be accepted by customers they need to generate "**trust**" i.e., customers need to have confidence that the decisions being made are fair and AI is not discriminating against them in any way.

A part of the framework will be setting up trust principles that every AI system has to comply with, and these must be ingrained within the culture of the company. Data scientists and other teams involved in data collection will be trained to ensure these principles are followed so that biases are minimized.

At a minimum, an AI system must follow the below principles:

- **Integrity** — Make sure that the machine learning algorithm is sound and cannot be tampered with. Any data used to train the algorithm will be used only for what it was gathered for and not for anything additional
- **Explainability** — The AI will not be a "**black box**" and the process by which the algorithm makes decisions will be transparent and documented.
- **Fairness** — decisions will be fair, ethical, free from prejudice, and will not discriminate against any age, gender, group, or ethnicity
- **Resilience** — AI system should be secure and robust enough to defend against attacks on its infrastructure or data by malicious parties.

Why build from scratch?

If you are serious about building an AI governance framework for your company, then the good news is that are numerous ready-made frameworks available that you can use as a template. My personal favorite is the **Model AI Governance Framework** released by the Singaporean regulatory authorities. Introduced in 2019 at the World Economic Forum (WEF), it provides detailed guidance for companies on how to mitigate ethical and governance risks when creating AI systems.

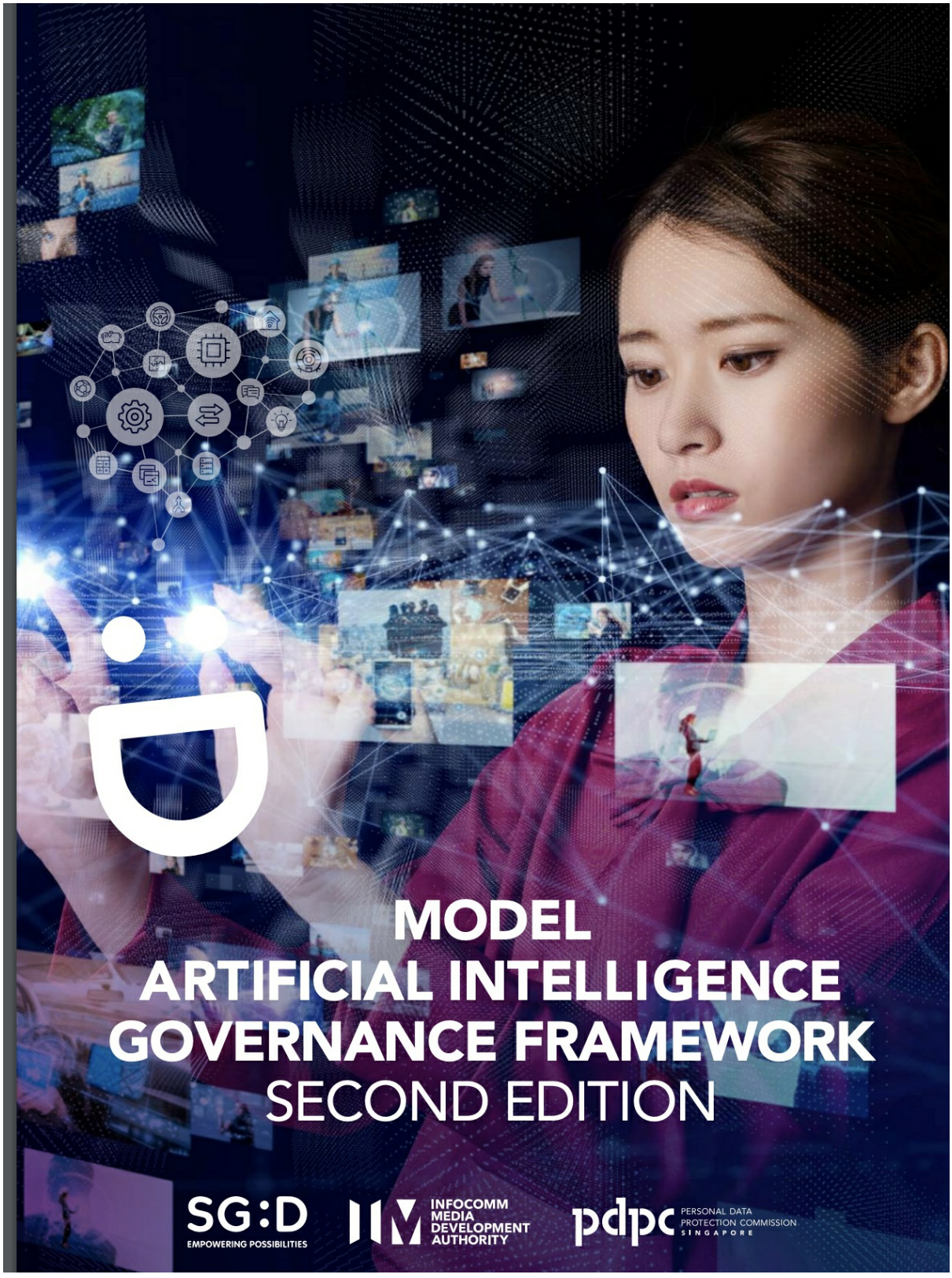


Figure 5.3: Current version of the Model AI governance framework

The framework provides a great blueprint for a model framework and is based on two guiding principles:

- decisions made by AI should be “explainable, transparent and fair”.
- AI systems should be human-centric (i.e., the design and deployment of AI should protect people’s interests including their safety and wellbeing)

You can use and implement parts of the model framework in your organization and tailor it according to your needs. The best thing about the model is that it can be adopted by any company regardless of its size or sector from a large bank to a small tech startup.

Chapter Summary

AI is changing the game for risk management professionals and having a proper governance framework is key to mitigating its risks. Management must realize that de-risking AI is not just red tape that slows down adoption but an actual competitive advantage that can be shown to win customer trust.

In this chapter we learned:

- What an AI governance framework is
- What its key components are
- How to re-use existing frameworks to create your own

6

Cyber-security risks of AI systems

Now that you have a firm understanding of the unique risks which AI systems can create and how to mitigate them; it is time to drill down into possibly the most interesting topic which is AI cyber-security risks. We live unfortunately in a world where data breaches and incidents are almost a daily occurrence with the multi-billion-dollar industry of cybercrime showing no signs of slowing down.

AI has been touted as a game-changer in cyber-security circles with the ability to detect and stop new types of attacks which sounds like a huge relief for overworked cyber-security teams (the term "**powered by AI**" has started showing up in almost all new security products).

However, as we will see AI can also introduce new attack vectors which require new ways of protection, and cyber-security teams need to make sure they are aware of them.

Cyber-Attacks on AI systems can happen in two ways:

1. **AI system gets compromised**: The system itself can get compromised either via its underlying technology infrastructure or through the machine learning model. Most cyber-security professionals will be familiar with the first attack but not so much with the second one.

    - **In the first attack**, it is the AI system that is the target itself and the attacker can compromise it via insecure configurations, missing access control, lack of patching, etc. The attack is like how traditional software systems get compromised.

    - **In the second one**, the attacker manipulates the unique characteristics of how AI systems work to benefit his malicious intentions. Many commercial models have already been manipulated or tricked and this type of attack is only set to increase with **Gartner estimating that 30% of Cyber Attacks will involve AI unique attacks by 2022.** This risk becomes even more dangerous when we realize that most companies adopting AI and Machine learning have cyber-security teams who are unaware of these types of attacks.

2. **AI-enabled cyber-attacks**: AI can also act as an enabler for cybercriminals empowering them to boost their productivity by automation. Nearly every benefit of machine learning systems we discussed earlier can be extended to cyber-crime also and attackers can automate many aspects of their attacks giving them more time to plan out sophisticated attacks with higher damage potential. In this attack, the attacker is using or manipulating the AI to attack someone else i.e., *it is not the AI that is the target.*

To give this more context, the "Malicious Use of AI" was a report written by 26 authors from 14 institutions, across academia, civil society, and industry. It surveyed the landscape of possible security threats that are going to arise from AI technologies and what measures can be put in to better mitigate these threats. An excerpt from the report follows:

A*s AI capabilities become more powerful and widespread, we expect the growing use of AI systems to lead to the following changes in the landscape of threats:*

EXPANSION OF EXISTING THREATS*: The costs of attacks may be lowered by the scalable use of AI systems to complete tasks that would ordinarily require human labor, intelligence, and expertise*

INTRODUCTION OF NEW THREATS*: New attacks may arise using AI systems to complete tasks that would be otherwise impractical for humans.*

CHANGES TO THE TYPICAL CHARACTER OF THREATS*: We believe there is reason to expect attacks enabled by the growing use of AI to be especially effective, finely targeted, difficult to attribute, and likely to exploit vulnerabilities in AI systems.*

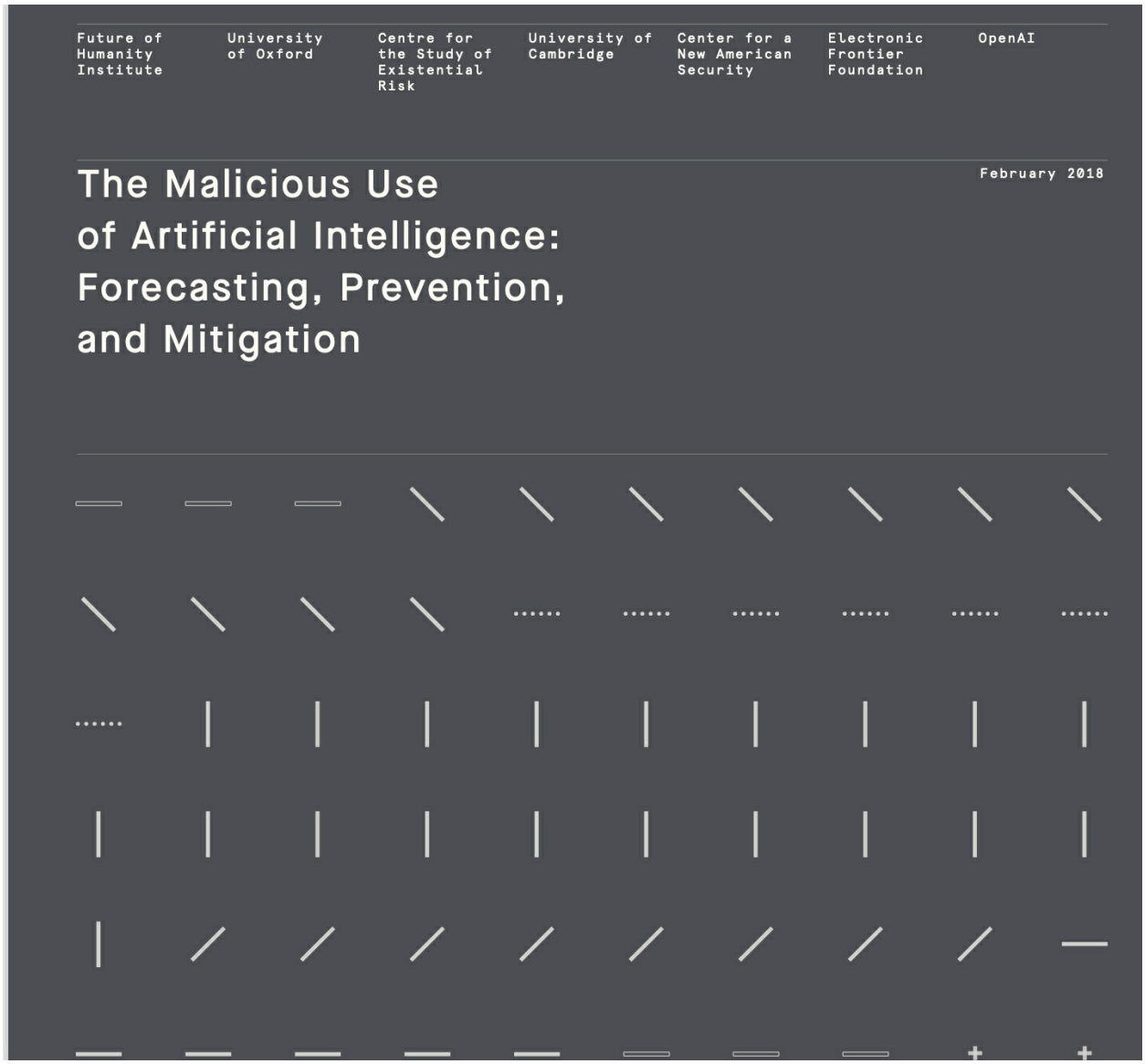


Figure 6.1: The malicious use of AI report

AI cyber-security vs traditional cyber-security

One of the biggest mistakes that cyber-security professionals make is to approach AI security with the approach of securing any other traditional software system. By that, I mean focusing on the security of the application, underlying infrastructure, access control, configuration, patching, logging, alerting, etc. all of which are good and needed.

For example, the following diagram is a typical example of how a traditional application is protected by implementing defense in depth i.e., layered controls at each level of the technology stack. A few decades back **application security** was not considered part of this layered security strategy and became a blind spot for most organizations resulting in a huge number of attacks

targeting the application layer. Now it is considered a mandatory part of any serious cyber-security defense model.

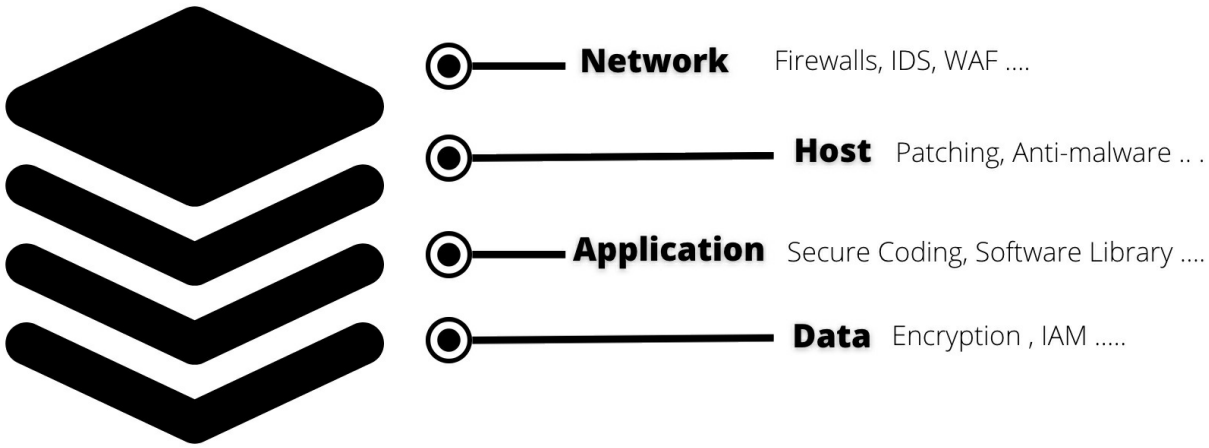


Figure 6.2: The traditional cyber-security defense-in-depth model

The same importance must be applied to AI systems as well. Just as application security became a blind spot for companies in the early 2000s and software supply chain attacks blindsided everybody in recent times; AI has unique cyber-security risks and assessing them the traditional way will miss out on key vulnerabilities and leave the system exposed.

For instance, one of the recommendations from Gartner's Top 5 Priorities for Managing AI Risk is that companies "*Adopt specific AI security measures against adversarial attacks to ensure resistance and resilience,*" and that "*By 2024, organizations that implement dedicated AI risk management controls will successfully avoid negative AI outcomes twice as often as those that do not.*"

What makes AI cyber-security different?

As a generally principle, AI and Machine Learning algorithms rely on their underlying models which analyze huge amounts of data to reach decisions.

What if an attacker was not interested in stealing the data but in tampering with the decision-making process?

Depending on the nature of decisions being made, the potential attack could be far more severe, especially with the rising adoption of AI across a variety of high-risk sectors. These new types of attacks are often referred to as ***Adversarial Machine Learning*** wherein attackers take advantage of these new types of vulnerabilities to bypass production AI systems for their malicious purposes.

AI-based attacks are present across the machine learning lifecycle as we will see and can be categorized as below:

Attack Type	Description
Data Poisoning	Attacker can poison the training data that is being used to train the Machine Learning model. By contaminating this data source, the attacker can create a “backdoor” as he knows the model has been trained on faulty data and knows how to take advantage of it. This can facilitate further attacks such as model evasion mentioned further on.
Model Poisoning	Like the previous attack but this time the attacker targets the model instead of the data. A pre-trained model is compromised and injected with backdoors which the attacker can take advantage of to bypass its decision-making process. Most companies do not build models from scratch but use pre-training models which are commonly available such as ResNet from Microsoft or Clip OpenAI. These models are stored in a Model Zoo which is a common way in which open-source frameworks and companies organize their machine learning and deep learning models. This is like a software supply chain attack in which an attacker can poison the well for many users
Data Extraction	Attacker can query the model and understand what training data was used in its learning. This can result in the compromise of sensitive data as the attacker can infer the data used in the model’s training and is especially dangerous if sensitive data was involved. This type of attack also called “membership inference” does not require access to the model’s functionality and can be done just by observing the model’s outputs
Model Extraction	Attacker can create an offline copy of the model by repeatedly querying it and observing its functionality. The fact that most models expose their APIs publicly and do not properly sanitize their outputs can facilitate these attacks. This technique allows the attacker to deeply analyze the offline copy and understand how to bypass the production model
Model Evasion	Attacker tricks the model by providing a specific input which results in an incorrect decision being made. This is usually accomplished by observing the model in action and understanding how to bypass it. For example, an attacker can attempt to trick AI-based anti-malware

	systems into not detecting their samples or bypass biometric verification systems.
Model Compromise	A functional model in production is compromised through a software vulnerability or via its underlying infrastructure. This is like a traditional attack and the attacker can compromise and take over a live AI system

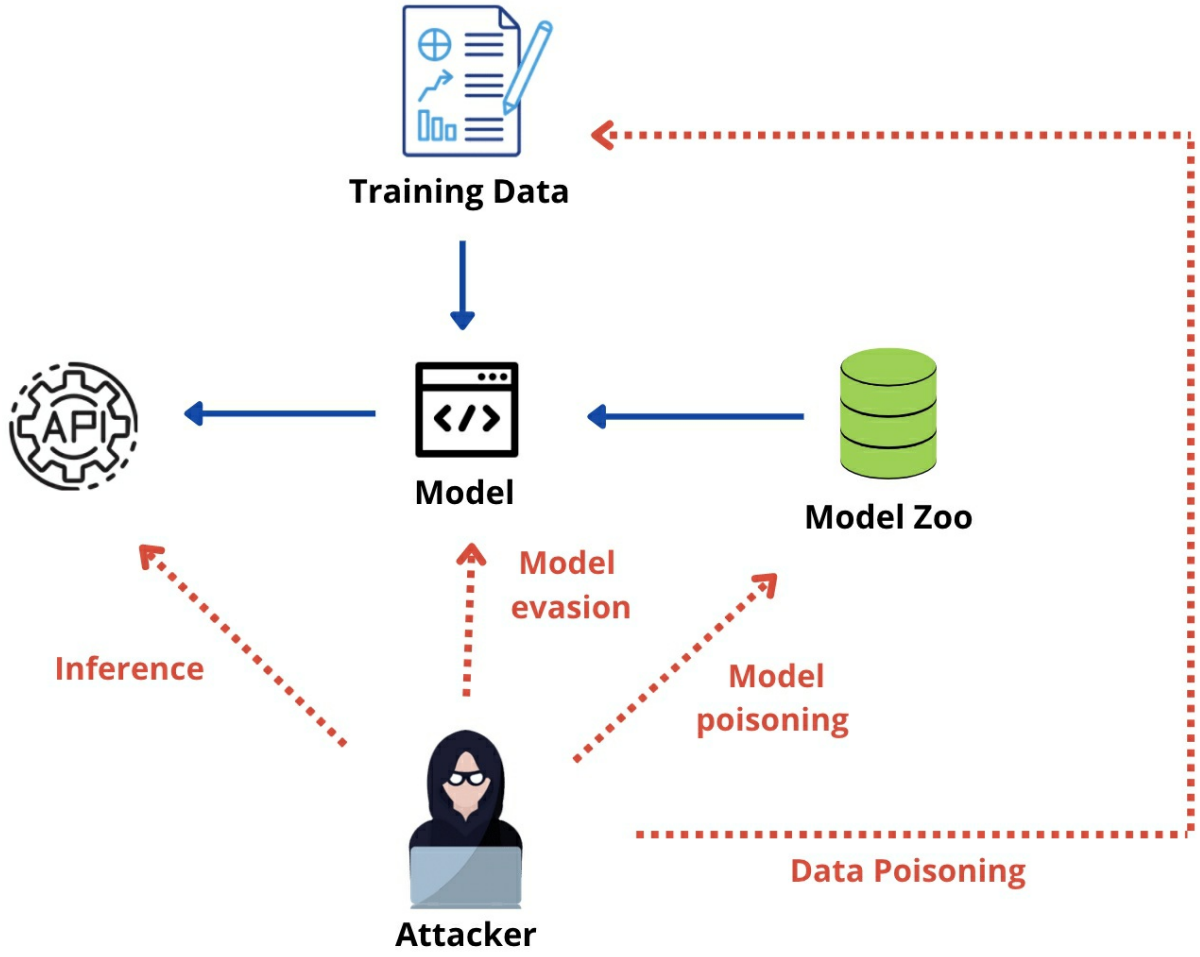


Figure 6.3: Attacks on AI systems

Now to give these risks better context let's look at them from the lifecycle of an AI system. This would enable us to map these risks to stages of an AI and see how these risks arise

Lifecycle of an AI system

Let us take the example of a sample “Company A” that wants to launch a new AI-based system in the market which will give them a competitive advantage and enable them to jump onto the AI hype machine. The model can be anything from something that analyzes customers spending patterns to predicting future trends or a credit scoring model.

For the model to be effective it must be trained on huge amounts of data so it can build a reasonable decision-making capability.

STAGE 1: SELECTING THE MODEL:

The first stage involves choosing the appropriate machine learning model for this use case. In most cases companies do not create these models from scratch as they can be very computationally expensive to train, requiring months of computation on many GPUs. Instead, most companies prefer to either outsource the model training or purchase pre-trained models. This makes sense if the company wants to go to market quickly and does not have the resources (data, computation) on hand.

Risk: Model Poisoning

In this phase a key threat vector would be an attacker poisoning the actual AI model. As mentioned earlier companies usually do not have time to create models from scratch and purchase pre-trained models to make their lives easier. An attacker can potentially compromise this model zoo and inject his malicious instructions or logic.

This is like a backdoor attack in software systems where the application functions exactly as it should apart from a backdoor that the attacker can activate. For example, an attacker can teach an AI-based anti-malware solution to correctly detect all malware **EXCEPT** the one which the attacker will introduce. This is something that the company will not be aware of until the attack happens. By compromising the model zoo, the attacker can also increase his attack surface and poison the well for everyone else.

STAGE 2: TRAINING / TESTING / OPTIMIZING THE MODEL:

In this phase, the model will be trained with sufficient data so that it can start making decisions with accuracy. Training data will consist of sample input and output data which the AI model will correlate to reach decisions. Those decisions will then be checked against actual results to see how accurate or correct they were. This is a crucial phase as the quality and quantity of the data have a direct hand in deciding how effectivity the model will be

Again, in most cases the company is not going to create this training data from scratch and instead either use something freely available or purchase a commercial training dataset

Risk: Data poisoning

An attacker can potentially compromise the training data and pollute it so the training itself is wrongly done. Since most of these training models are outsourced or purchased commercially, the attacker can render the machine learning model useless right from the start leading to months of work being wasted. This can be a nuisance or something much more deadly depending on the risk level of the model.

Let us take the example of a self-driving car that was being trained to recognize objects while driving. An attacker could potentially modify this training data so that the car will not recognize a stop sign leading to injury or death. Models are also regularly refreshed on training data so a smart attacker would not do this immediately but instead use it once the company has gained confidence in the source of the training data. By executing this attack an attacker can manipulate that information to teach AI systems anything they want. If it was a software being trained to detect malware, then they can make them see good software code as malicious code and vice versa.

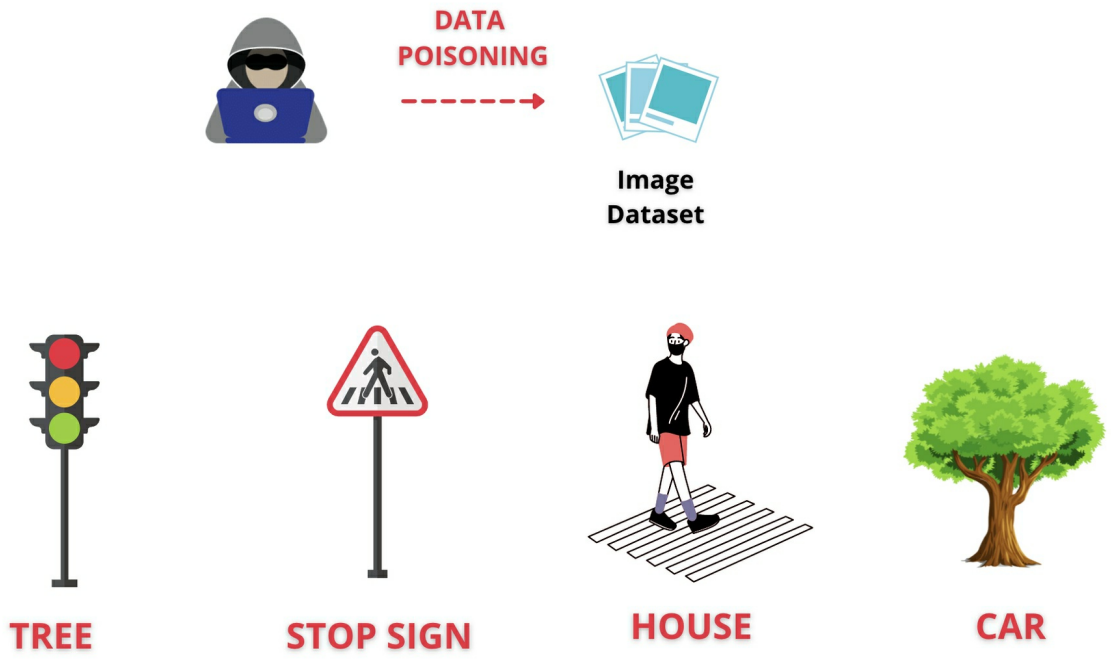


Figure 6.4: Data Poisoning a Self-driving car model

Risk: Data breach

Another risk at this crucial phase would be that of a data breach. Data is the life of AI systems and what drives the quality of decision-making. The more data that is used for training, the more accurate the model will become. However, the problem comes with how this data is handled. Attackers know that AI systems will be trained on real-life data before they are moved to market thus making their data stores a prime area to target. AI data sources are not hardened databases protected from attack but often excel or CSV files stored in folders with a very permissive level

of access. This can be a treasure trove for attackers who can access and exfiltrate this data.

Stage 3: Deploying the Model

This stage involves deploying the machine learning model in production and businesses seeing the value add from their investment. This can be as simple as exposing an API over the internet to be consumed or a much more complex multi-cloud deployment depending on the model.

Risk: Model compromise

In this risk, the attacker attempts to compromise the underlying vulnerabilities of the application or infrastructure on which the machine learning model is hosted. Despite their unique nature, models are still vulnerable to traditional software flaws such as buffer overflows or cross-site scripting and can be attacked in the same way. It is also much easier to try and attack the underlying layer than attack the machine learning model directly.

Stage 4: Maintaining and Monitoring the model

In this phase the model is now fully operational, and the job of the company now shifts to making sure it is running smoothly and monitoring its performance. The model's performance is fine-tuned to become more accurate over time as it keeps learning about new data. This phase continues the earlier risk of model compromise but with some new risks showing up:

Risk: Model Evasion

An evasion attack also referred to as an "adversarial sample" is a technique that attempts to fool the machine learning model. By subtly manipulating the input data going into the model, the attacker can basically "trick" the machine learning model into reaching incorrect decisions. For example, a few pixels added to an image would be invisible to the human eye but might potentially completely throw off a machine learning model and cause it to reach a wrong decision.

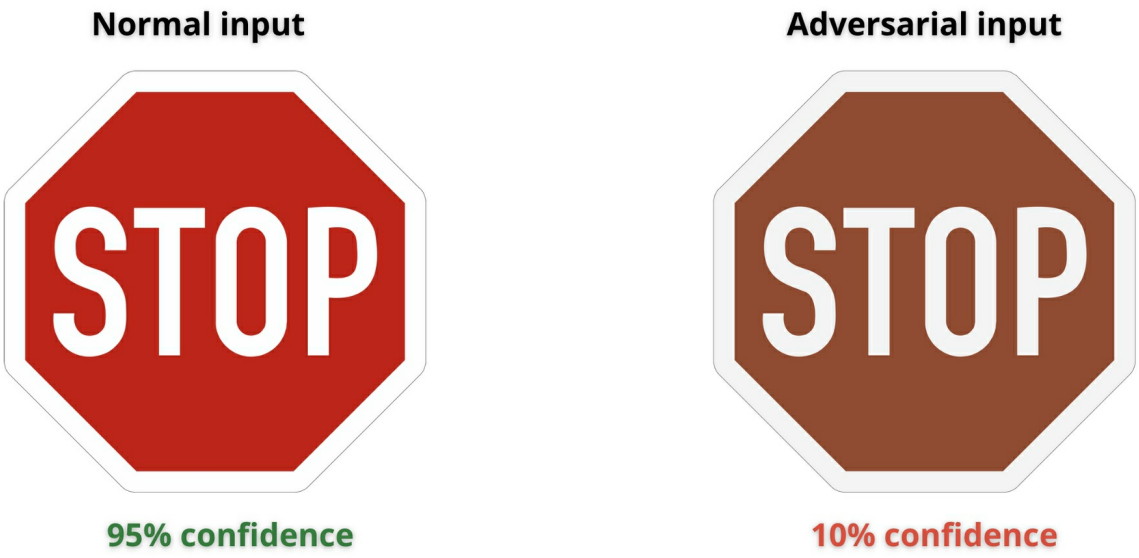


Figure 6.5: Model evasion via adversarial inputs

Next up in this stage, we have privacy-based attacks in which the attacker gleans data from the AI system either about the training data that was used or the model itself. We divide these into Model or Data extraction attacks. Both attacks are quite dangerous as in most cases they do not require the attacker to have an underlying knowledge of the training data, the algorithm used, and the technology. Models are usually exposed via APIs and simply querying the model repeatedly and observing the outputs can be enough to facilitate these attacks:

Risk: Model Extraction

In this attack the attacker has access to the model API and can recreate the model by sending legitimate queries and analyzing the results. The new model has the same functionality as the original model and allows the attacker to uncover how the model was designed or make inferences from the data that was used to train it. The attacker usually does it for two reasons:

- He can use the duplicate model to predict what decisions the original model will make and how to get around it i.e., the evasion attack we mentioned earlier
- He can also steal the functionality of the model which can be a valuable trade secret for your company. Models can take months or years to design for a company and attackers would be more than willing to duplicate the inner workings and sell them

Risk: Data Extraction (Membership inference)

This attack is like the previous but in this case, it is the data that the attacker is attempting to

extract from the model. Machine Learning models can be deployed in several critical industries such as healthcare, banking, government, etc. which can have Personally Identifiable Information (PII) or Cardholder data that can be very valuable to an attacker.

If the attacker knows what he is doing he can reverse engineer a model and make it disclose what information was used to train it. This can have serious consequences if the machine learning model is trained on data deemed to be sensitive. Let's take a few examples of how this attack can happen:

- An attacker can query a model with a name or an identifier to find out if a person is on a patient list in a hospital or a sensitive medical list.
- An attacker could find out if a patient was being provided certain medication or not
- An attacker can provide images to a facial recognition model to find out if a particular face was used in the training or not.

In more advanced attacks a person with malicious intent could even extract credit card numbers and other sensitive information like social security numbers!

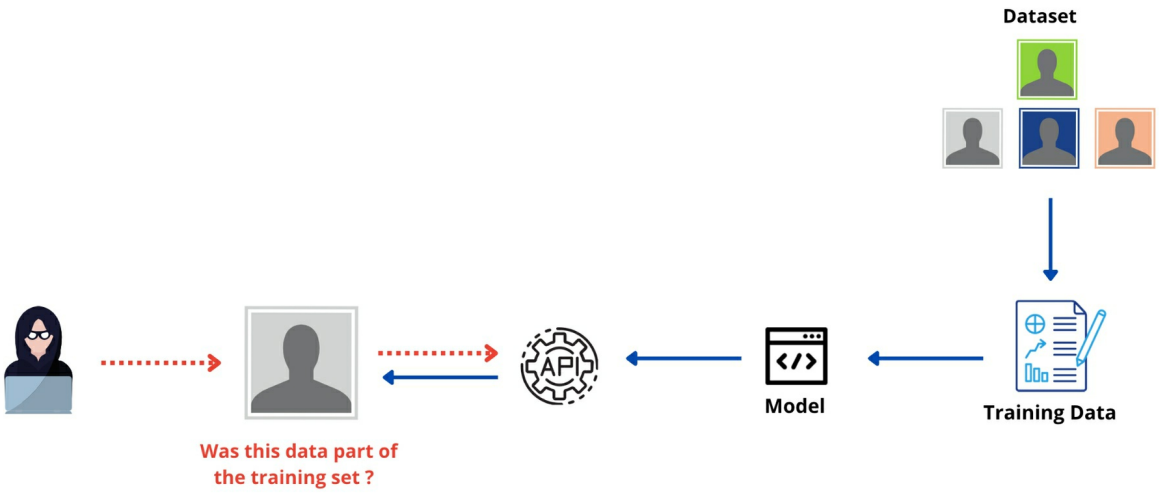


Figure 6.6: Data extraction

Now that we saw the different types of attacks, we can map them to the different phases and see the distribution of cyber-security attacks that can happen over the lifetime of a machine learning model.

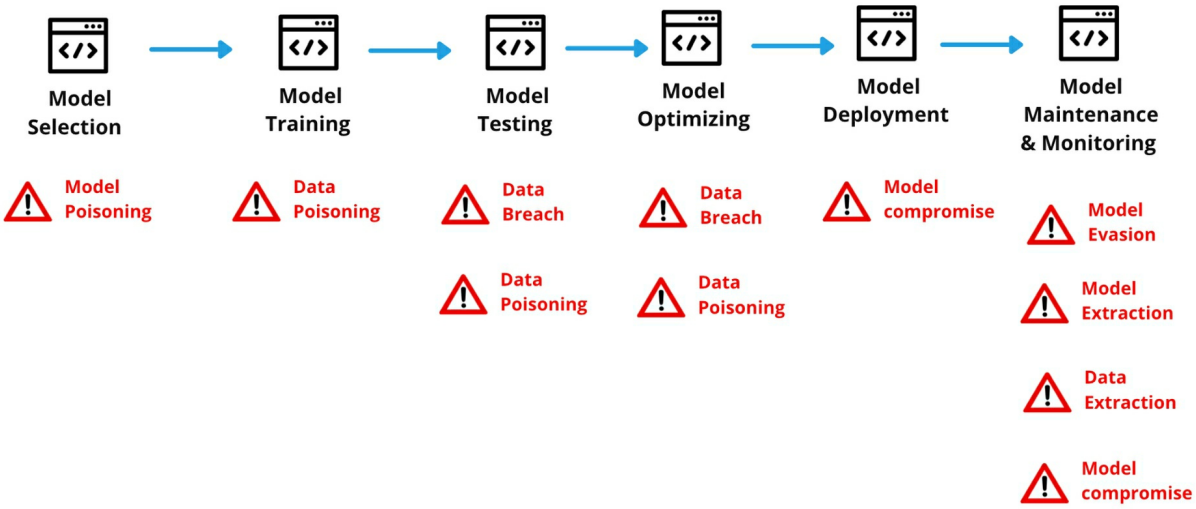


Figure 6.7: AI-specific attacks spread throughout the lifecycle

Chapter Summary

In this chapter we got a bit more technical and saw the different types of attacks that can happen on an AI system. These are just the tip of the iceberg as adversarial Machine Learning is an active area of research with new attacks being added constantly.

We learned:

- How attacks on AI systems can happen
- How cyber-security of AI systems differs from traditional security
- The unique risks of AI systems across their lifecycle

7

Creating an AI cyber-security framework

Now that we have a good understanding of the different types of AI risks that can occur over the lifespan of a model, let us look at how we can mitigate these attacks. **The bad news** is that as of this time there is no AI equivalent of the ISO 27001 framework or the PCI DSS standard i.e., an internationally recognized security standard you can refer to for implementing AI-specific controls. There are some good initiatives in the pipeline but nothing that the industry has universally adopted like ISO.

Companies that are serious about securing their AI systems will have to understand the previously mentioned risks and then select controls designed for these problems. Over time as more awareness is created then we will see standards evolve and form but until then companies must be proactive and mitigate these threats before they are taken advantage of. As always there is a trade-off between productivity and security and cyber-security pros will need to play the balancing act between securing the system and letting it do its job at the same time.

The good news is that cyber-security is already a mature discipline that can quickly adapt and incorporate new types of risks into its existing frameworks and AI is no exception. As security professionals become more and more aware of these risks, we will see AI security controls move into the mainstream. This is like how Application Security was a niche a few decades back but is now considered a given for any cyber-security strategy.

Let us look at how we can go about creating an AI security framework in a company.

How to create an AI cyber-security framework

While we are in uncharted territory here, there are some key steps that a company can take to implement to create an AI-focused cyber-security framework which are listed below:

3. Map the regulatory environment in which AI systems operate
4. Create an AI/machine learning security baseline
5. Maintain an up-to-date inventory of all AI and Machine Learning systems
6. Conduct detailed technical risk assessments of your AI systems
7. Create an awareness program about AI risks
8. Update existing security assurance processes to incorporate AI and Machine Learning technicalities

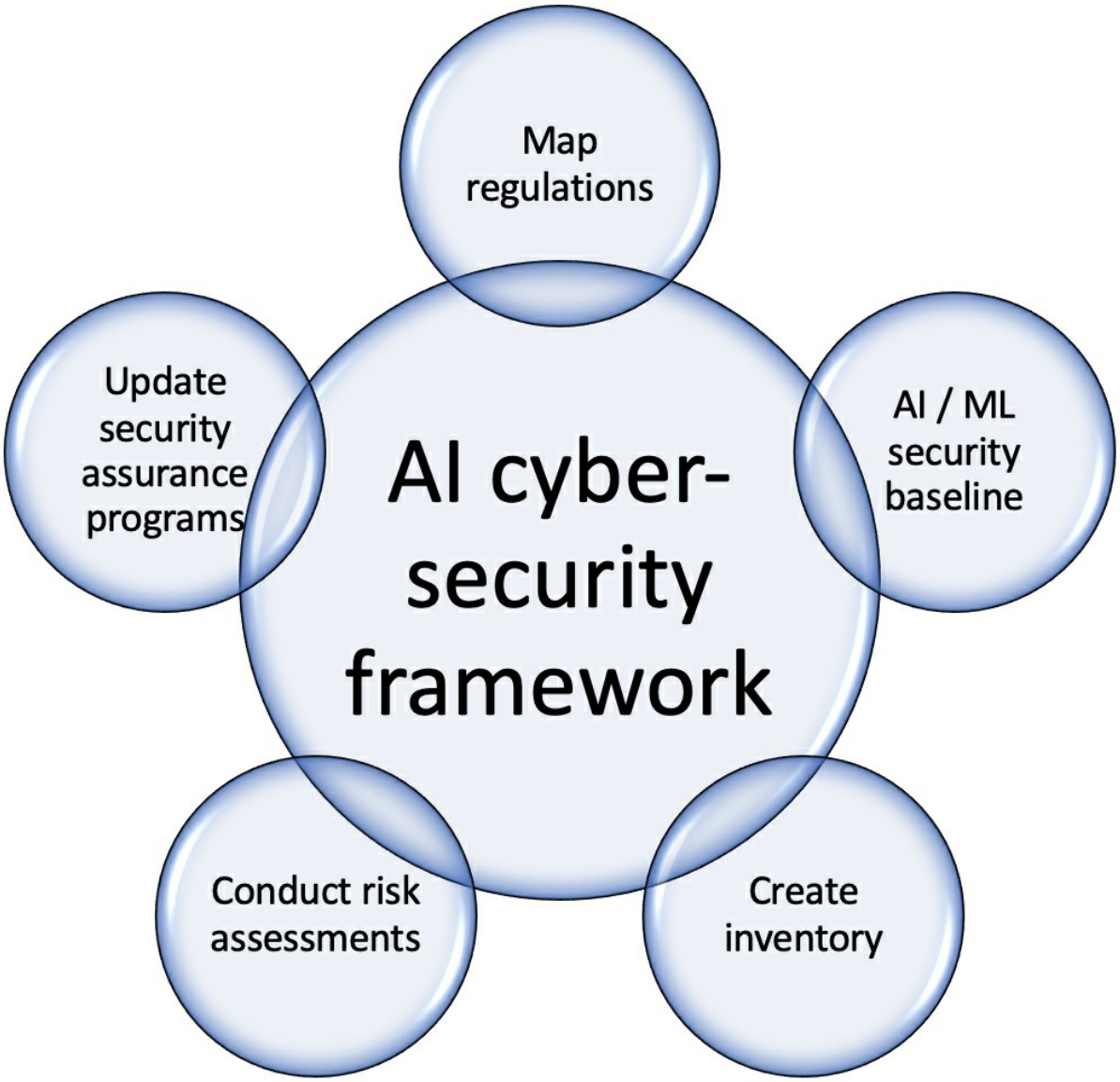


Figure 7.1: Key components of an AI security framework

Let's look at each in detail:

1. **Mapping the regulatory environment in which AI systems operate:**

Regulation will play a large part in determining the risk level of our AI systems and what sort of data they will be allowed to process. Make sure you are aware of the regulatory environment and what requirements are present. If your AI systems are falling under the upcoming EU AI act, then assess the risk level and read the requirements on conformity assessments to make sure your

system is complying with the same.

2. **Create an AI/machine learning security baseline**

To make sure security is consistently applied across AI systems the company will need a minimum baseline of security controls to be applied. This will be at two levels:

- **Security of the underlying infrastructure** on which the AI system / Machine Learning model is hosted. For example, if you are using Azure or AWS to host your machine learning model then you must make sure the services are properly hardened as per best practice guidelines. This is a standard part of any production rollout and does not differ for AI systems. This is where you will turn on encryption, access controls, logging, etc.
- **Security of the AI System** itself to mitigate the unique risks which AI systems introduce. This is where the value of having a security baseline will start to show. We will take a detailed look at this in the next section "Implementing AI controls"

3. **Maintain an up-to-date inventory of all AI and Machine Learning systems**

It is difficult to secure anything without knowing it exists in the first place! Identify all assets in your AI ecosystem as a fundamental step so you know how many AI systems are present and how they can be protected. Make sure the inventory captures the below:

- Business unit using the AI system
- Technology of the AI system
- Customers it services (internal or external)
- Type of data it is trained on
- Risk level of the AI system
- Exposure level of the APIs

4. **Conduct AI technical risk assessments**

Based on the risk level of the AI systems identified in your inventory, detailed technical risk assessments must be carried out and documented with mitigations and timelines. This is a collaborative effort with the business and technical teams. The methodology you use is not important but what is critical is that this is a **repeatable, standardized process** that is consistently followed. This can come under the umbrella of the risk management framework we discussed in earlier chapters. We will see more details on how to conduct threat modeling of AI systems in the next chapter.

5. **Create an awareness program about AI risks**

Easily one of the biggest risks in AI systems is the overall lack of awareness amongst cyber-security professionals. As mentioned earlier the unique risks present in AI systems are either ignored or are treated like any other software rollout with hardening, patching, and

configurations done without regard for the risk of the AI model. If you are serious about AI cyber-security, then it is crucial to upskill your staff and make them aware of these risks. Your company might need to engage with third-party consultants to initially assess these risks and train your teams in parallel until you feel they are at an adequate level

Additionally, it is important to educate the people involved in creating Machine Learning such as data scientists on these risks before machine learning algorithms are used in business environments. These are the people who are interacting with the data and systems on a daily hence they must take ownership of its security

6. **Update existing security assurance processes to incorporate AI and Machine Learning technicalities**

Most companies have security assurance processes present as part of any application rollout in which a full security review of the application is carried out to capture security risks. You must make sure that AI-specific security testing is a part of this process. For example, using adversarial testing samples that simulate model evasion can be done during the model testing phase to assess its level of susceptibility to model evasion attacks as seen in the following diagram. We will see details of AI security testing in chapter 9.

This is not a one-time process but something which has to be carried out annually at a minimum with the AI inventory created earlier acting as an input.

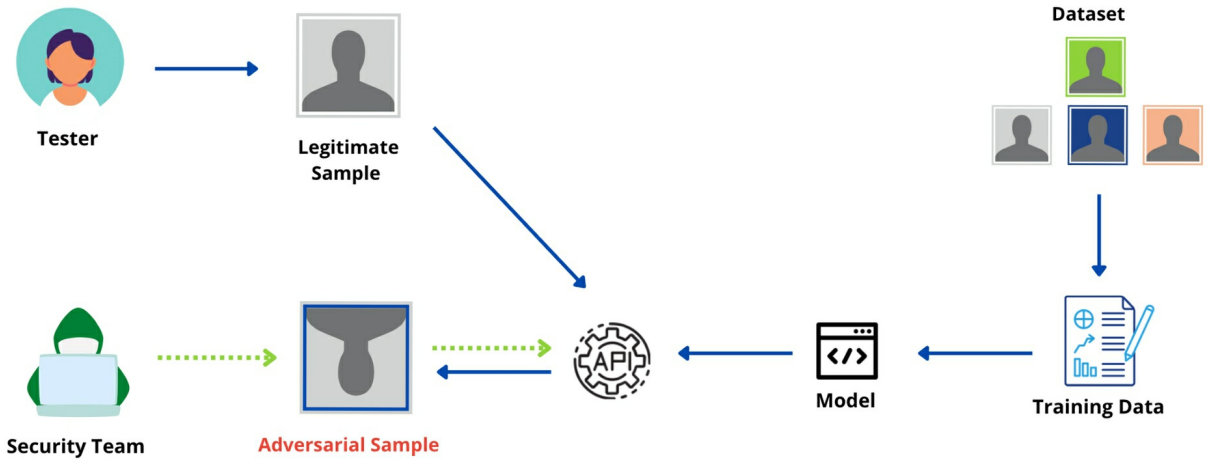


Figure 7.2: Example of AI security testing by internal teams

Implementing AI-specific controls

We talked about creating an AI security baseline to standardize security in AI systems. In the following table let us look at each of the risks we highlighted and the corresponding security control that should be present to mitigate this risk. You can use this as a starting point to create your AI security baseline document:

Risk Identified	Description	Security Control
Data Poisoning	Attacker alters data to modify the ML algorithm's behavior in a chosen direction e.g., poisoning the data fed to an algorithm to make it seem cats are dogs and vice versa or modifying facial recognition data	Data must be checked to ensure it suits the model and limit the ingestion of malicious data. Check the trust level of the data source and protect the integrity of the data pipeline. Have controls in place to reverse the damage done if the data source is contaminated. For example, being able to revert to a clean source of data
Model Poisoning	A type of attack in which the attacker has injected a rogue model in the AI lifecycle as a backdoor. This is especially risky if the company is not creating its model from scratch and relying on publicly available ones (think supply chain attacks like SolarWinds)	Do not reuse models taken directly from the internet without checking them. Use models for which the threats are identified and for which security controls are documented and verified. Ask reports from the vendor in which they detail their internal risk management processes for AI systems
Data Breach	The attacker compromises and can access the live data that is fed to the model in the fine-tuning phase or production phase. This is usually done via traditional cyber-security attacks	Ensure that the data sources are secure from unauthorized access. If third-party data sources are being used, then ensure their integrity is checked. Traditional cyber-security best practices can be used to harden the systems from attack
Model compromise	This threat refers to the compromise of a component or developing tool of the ML application. e.g.: compromise of one of the	Verify that any software libraries used within your AI system do not have major vulnerabilities Define dashboards of key indicators integrating security indicators (peaks of change in model behavior etc.) to allow rapid

	open-source libraries used by the developers to implement the ML algorithm.	identification of anomalies
Model Evasion	Attacker works on the ML algorithm's inputs to find small inputs leading to large modifications of its outputs (e.g., decision errors). This attack tries to trick the ML algorithm to make wrong decisions	Generate adversarial examples for testing the AI model and attempt to perform an evasion attack. Make sure it is part of your testing suite. We will cover this in detail in the Chapter
Data Extraction	Attack attempts to infer data by providing a series of carefully crafted inputs and observing outputs. These attacks lead to more harmful types, evasion or poisoning for example. e.g., AI system can give too much information about its training data in its outputs allowing an attacker to understand what data was provided	One of the difficult attacks to protect against as most Machine Learning models must provide access to their APIs which is what attackers target in this attack. A few of the controls that are recommended are: ● Controlling how much information is provided by the model (i.e., verbosity) in its outputs. This is like how cyber-security experts apply best practices to application error messages to limit how much information is shown. ● Harden the exposed APIs to make sure they are only accessible to specific users (if possible) and do not give away internal confidence scores of results. This makes it much easier for attackers to understand how and what data is being shown and infer the internal workings of the model and attempt to bypass it. ● Implement data sanitization to ensure that no PII or Cardholder data is sent back in the output text and is properly sanitized.

Model Logic Extraction	This threat refers to a leak of the internal working of the machine learning model i.e., the attacker can extract an offline copy of the model and recreate its internal workings. This enables further attacks on the live system.	Same as above, controlling the information (like its verbosity) provided by the model by applying basic cybersecurity hygiene rules is a way of limiting the techniques that an attacker can use to build his knowledge of the model Another key control here is anomaly detection as extracting the logic does not happen overnight. The attacker will have to query the model probably thousands of times before he understands its internal workings. Such queries and their patterns should be monitored and raise red flags within your network like how a port scan of your public IP addresses might be an indicator of an attack. A team monitoring such queries can get alerted and shut off access to an attacker if it indicates a model logic extraction attack

Now that we have a better idea of the different controls that can be implemented to mitigate AI-specific risks, we can map them to specific stages also as shown in the following diagram.

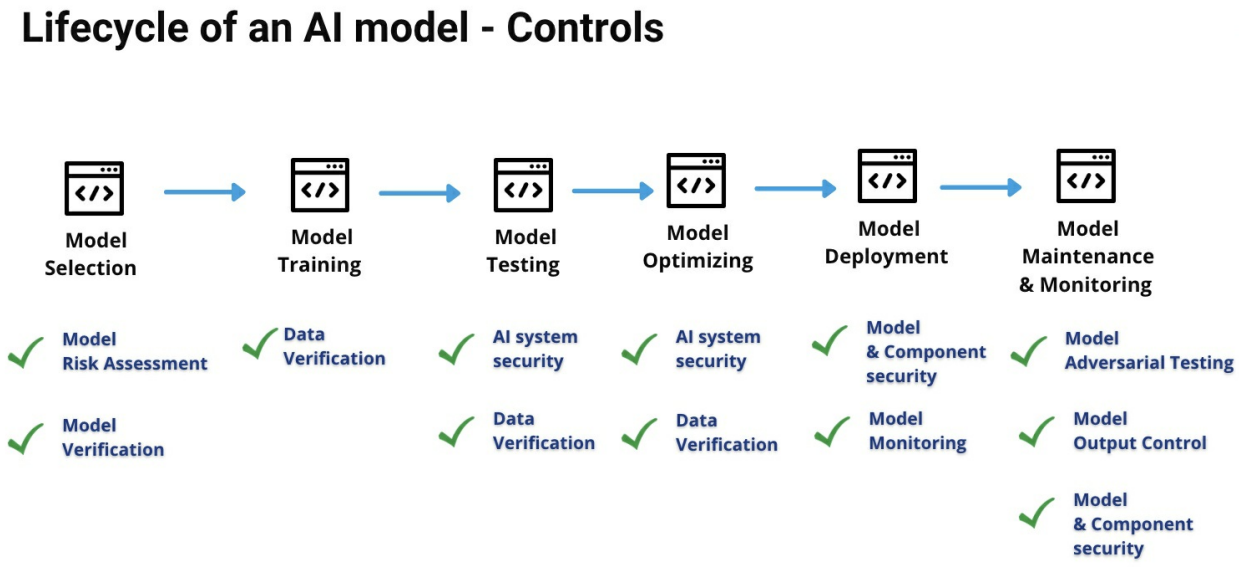


Figure 7.3: AI controls mapped to lifecycle stages

Chapter Summary

In this chapter, we saw how to implement specific security controls to mitigate the unique AI risks that can emerge. This was by no means an exhaustive list as AI security is very much an emerging field. In the next chapter, we will get into more details about how to threat model an AI system to identify risks early in the AI lifecycle.

In this chapter, we learned:

- Key components of an AI security framework
- Detailed controls mapped to AI risks

8

Threat Modeling AI systems

In the last chapter we talked about creating an AI security framework and why it is so important to conduct technical risk assessments of AI systems to identify risks early on. One of the best ways to do that is via **Threat Modeling**.

What is Threat modeling?

Threat Modeling refers to a structured way of identifying security threats to a system and is usually consists of the below:

- A high-level diagram of the system
- Profiles of attackers and their motives
- A list of threats to the system and how they might materialize

Threat Modeling is like risk assessments, but you adopt the perspective of an attacker and see how much damage they can do. There are numerous methodologies and tools available for threat modeling which we do not have to cover here but honestly, you can create a threat model with pen and paper if you understand the core concepts!

Threat Modeling AI applications

In its simplest form, you create a high diagram of your AI application and envision scenarios in which threats will materialize. Let us take the following example of a sample AI system for which we create the following high-level abstraction:

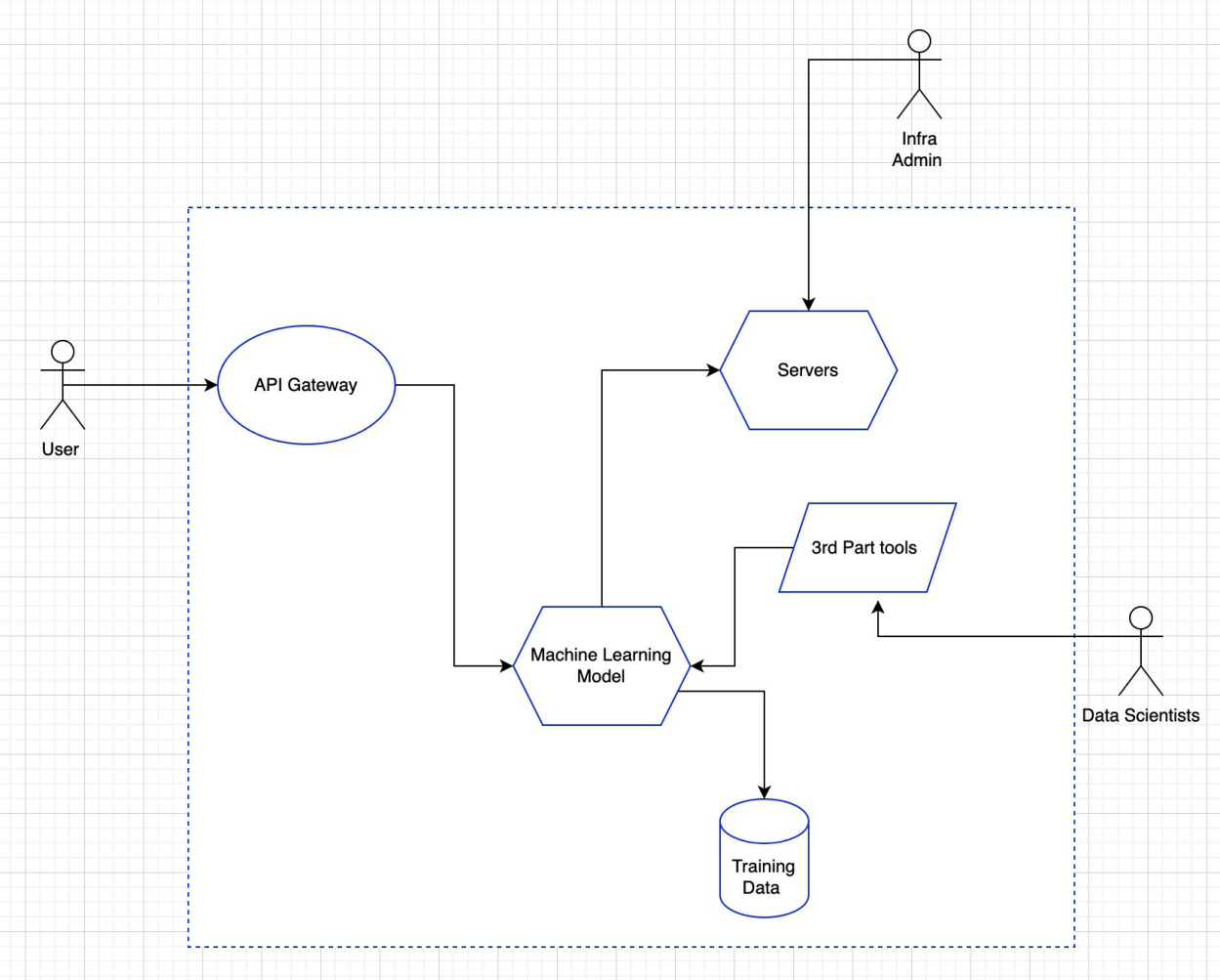


Figure 8.1: High-level diagram of an AI system

Even though this is a very basic conceptualization, you can still use it to decompose the AI system and list down the key assets that might get targeted by an attacker.

Some of the key assets would be:

- The training data
- The public-facing API
- The Machine Learning Model
- The Servers hosting the model
- The infra-admin access keys or credentials

Now assume the viewpoint of an attacker and try to imagine what are the areas they would target. You do not have to do this alone and should involve the data scientist and technology staff that are part of the AI team to help you with the same. Brainstorming threats is a great way to identify weak areas of your technology ecosystem with numerous methodologies present.

STRIDE is a popular threat modeling methodology by Microsoft that I prefer which classifies threats into the following categories:

1. **S**poofing
2. **T**ampering
3. **R**epudiation
4. **I**nformation disclosure
5. **D**enial of service
6. **E**levation of privilege

Try to classify all your threats into these categories and envision at least one for each.

Sample list of identified threats:

Below is a sample list of identified threats and proposed mitigations to discuss further with the teams. Document the threat even if there is a mitigation present so that there is a record for the same.

You can use any tool for creating this list or assign these via a ticketing system.

Threat: Attacker poisons the supply chain of third-party tools used by Data Scientists:

- **Category:** Tampering, Elevation of Privilege
- **Mitigation:** Scan software libraries before usage. Make sure integrity checks are present

Threat: An attacker tampers the training data and attempts to poison the same:

- **Category:** Tampering
- **Mitigation:** Security controls over the training data. Hashing and integrity checks to validate changes

Threat: An attacker causes a Denial-of-Service attack by deleting the training data:

- **Category:** Denial of Service
- **Mitigation:** Regular backups of the images. Restricted access to the training data

Threat: Attacker gains access to the model and attempts to replicate it

- **Category:** Information Disclosure
- **Mitigation:** Throttling limits present on the exposed APIs to restrict the number of requests that can be made. Alerts for an abnormal number of calls. Limited information in the output requests.

Threat: Attacker gains access to the training data and attempts to exfiltrate it

- **Category:** Information disclosure
- **Mitigation:** Restricted access is given on a least-privilege basis. Training data is encrypted at rest and requests specific keys to decrypt.

Threat: Attacker gains access to admin credentials or keys

- **Category:** Elevation of privilege, Information disclosure
- **Mitigation:** Admins use multi-factor authentication to access servers via hardened endpoints.

Threat: Attacker exploits software vulnerability on Machine Learning system

- **Category:** Elevation of Privilege, Denial of Service
- **Mitigation:** Machine Learning only exposes an API publicly. Application is hardened to attacks and checked via periodic scanning and tests

Threat Modeling sample questions:

As you can see threat modeling is more of an art than an exact science and you will get better at it the more you do it. The specific methodology you use is not important but what is important is that you do it consistently without fail and track these risks to closure. This should be done in collaboration with the technology and business teams who will be involved in the training/creation of the AI system on a periodic basis.

Below is a list of sample questions you can refer to when doing Threat Modeling in your organization. These questions are just samples to help you which can be converted into a checklist and used by anyone from Technology Auditors to Cybersecurity professionals when carrying out reviews

- Is the training data downloaded from public data stores? if yes what sort of authentication mechanism is present? How do we ascertain the quality of this data?
- If the online data store is compromised or has a data breach, then will we be notified?
- If our training data is poisoned, then how would the company find out? Do we have any metrics/indicators which can indicate if the training data qualify has been

tampered with?

- If the training data was poisoned, then would it be possible to re-train the models and roll back to a safe version?
- Does the training data contain sensitive information which can be classified as Personally Identifiable Information (PII) or cardholder data?
- Is there any input validation or sanitization done on the data before being consumed by the model?
- Can the model APIs contain data that can be classified as PII? Was consent taken by the owner for this data?
- Is there any process to sanitize data on the output being sent back to the caller?
- Does the model output contain raw confidence scores that can help the attacker gain information about its internal workings?
- If the model decision-making accuracy experiences a sudden drop, then are there any alerts or notifications configured to alert the responsible persons of possible tampering?
- Has the model been trained with adversarial inputs to be sufficiently hardened against model evasion attacks? The goal here is to see if sufficient testing was done with adversarial inputs to trick the model
- Can the model be tricked into denying service to a particular set of users? For example, making the model blacklist a particular word that denies access to users who legitimately use that word in their operations. *This can be a common attack against conversational bots*
- What is the exposure level if the model is stolen by an attacker?
- Can the outputs provided by the model be used to deduce information about a particular person (membership inference attack)? Does the model limit response to only the desired information? For example, an attacker could find out if a particular person went through a specific medical procedure based on other attributes e.g. age, gender, location, etc.
- Can the model be queried repeatedly to disclose what data was used to train it? Are there alerts to detect such queries?
- How are users authenticated who access the model and its APIs? Can this access be revoked in case of suspicious API calls or compromise?
- Does your Machine Learning model rely on third-party dependencies? If yes then are those third-party libraries scanned for security vulnerabilities before use?
- If you are relying on an outsourced machine learning model, then has its logic been

reviewed for backdoors? How is the risk of malicious Machine Learning providers being mitigated?

- If you are relying on an outsourced machine learning model, then has its logic been reviewed for backdoors? How is the risk of malicious Machine Learning providers being mitigated?
- Is there a policy that prevents sensitive machine learning models from being outsourced? This is to mitigate the risk of malicious machine learning models being used which have pre-trained data that contains hidden malicious instructions. Otherwise, is there a process to assess the security of the third party creating the model
- Have the technical systems hosting the Machine learning model/data been reviewed/signed off by the cyber-security teams for adequate security controls?

Chapter Summary

In this chapter we saw how threat modeling can help us envisage AI risks and threats in a clear and easy-to-understand manner. Essentially, threat modeling is asking the question “What can go wrong” and answering it while thinking like an attacker. Carry out this process throughout the AI lifecycle and whenever there is a major change, and you will see a tangible improvement in your AI security posture

Topic we covered:

- What is Threat Modeling
- Using Threat Modeling to envisage AI risks
- What sort of questions to ask during AI threat modeling?

9

Security testing of AI systems

In the last chapter we looked at **threat modeling** which is a high-level overview of your AI security risk posture and its exposure to security threats. In this chapter, we drill down to the technical level and see how to carry out security testing of AI systems to find out their weaknesses to specific attacks.

While penetration testing/vulnerability scanning is a standard process carried out by most cyber-security teams, AI-based security testing is still a relatively new field that is still maturing.

Where to start?

The good news is that you don't have to start from scratch. If you have ever been involved in penetration testing or red teaming as part of a cyber-security team, then you might be familiar with the **MITRE ATT&CK** framework which is a publicly accessible framework of adversary attacks and techniques based on real-world examples.

It is used globally by various public and private sector organizations in their threat models and risk assessments. Any person can access it and understand the tactics and techniques that attackers will use to target a particular system which is very useful for people involved in penetration testing or red teaming.

This popular framework was used as a model to create MITRE ATLAS (Adversarial Threat Landscape for Artificial-Intelligence Systems), which is described as

"a knowledge base of adversary tactics, techniques, and case studies for machine learning (ML) systems based on real-world observations, demonstrations from ML red teams and security groups, and the state of the possible from academic research"

ATLAS follows the same framework as MITRE, so it is very easy for cyber-security practitioners to study and adopt its techniques when they want to test their internal AI systems for vulnerabilities and security risks. It also helps in creating awareness of these risks amidst the cyber-security community as they are presented in a format, they are already familiar with.

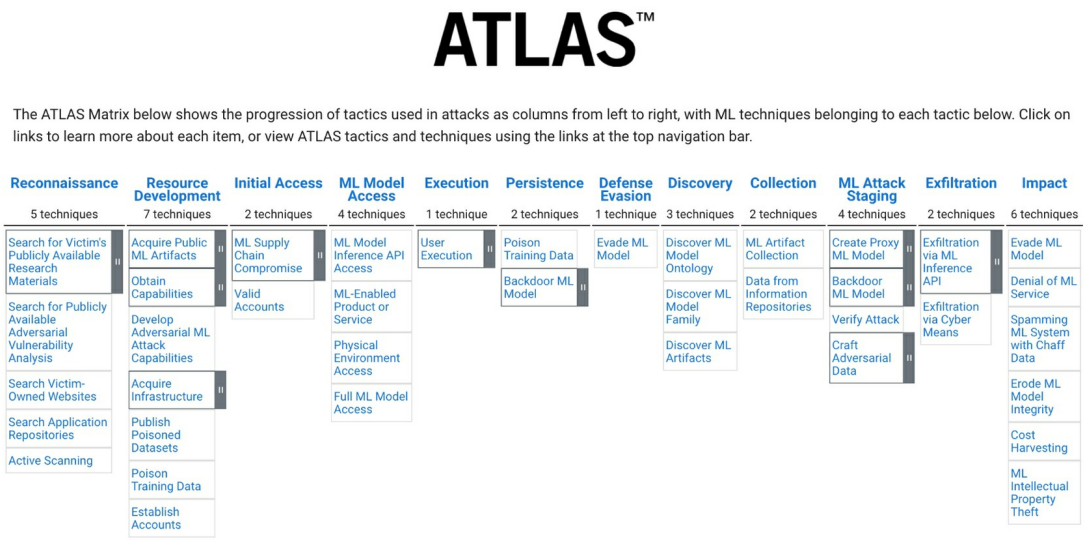


Figure 9.1: ATLAS framework for AI security testing

Types of tests

Cyber-security teams can refer to the ATLAS framework and use it as a launching board for deciding their security testing strategies. When security testing AI systems there are numerous approaches you can take some of which I have listed below:

1. **Standard security testing/penetration testing of the underlying platform**. This will highlight vulnerabilities of the platform hosting the AI system/model and is usually a standard process for most companies. The system does not go live until cyber-security teams validate the security of the platform.

2. **Security testing of AI systems**: This will cover specific AI-specific attacks using specialized tools which we will see. Due to the complex nature of AI systems, the best approach is to follow the red team / blue team model so that the security of the AI system is validated from multiple levels.

    - **Red Teams**: Like penetration testing but usually more targeted and can be done by internal or external trained personnel. The purpose is to test the effectiveness of the AI security controls. The red team will carry out a series of targeted attacks against the AI system usually as part of a campaign.

    - **Blue Teams** would be the internal security team that checks the effectiveness o the AI controls against both attackers and the red team. Blue teams have to adopt the perspective of the attacker and have experience in penetration testing/red teaming also. *It is generally not recommended to mix your red team / blue teams with your standard cyber-security teams as it requires a different level of expertise to be effective*

If done effectively then your security testing program would look like the following:

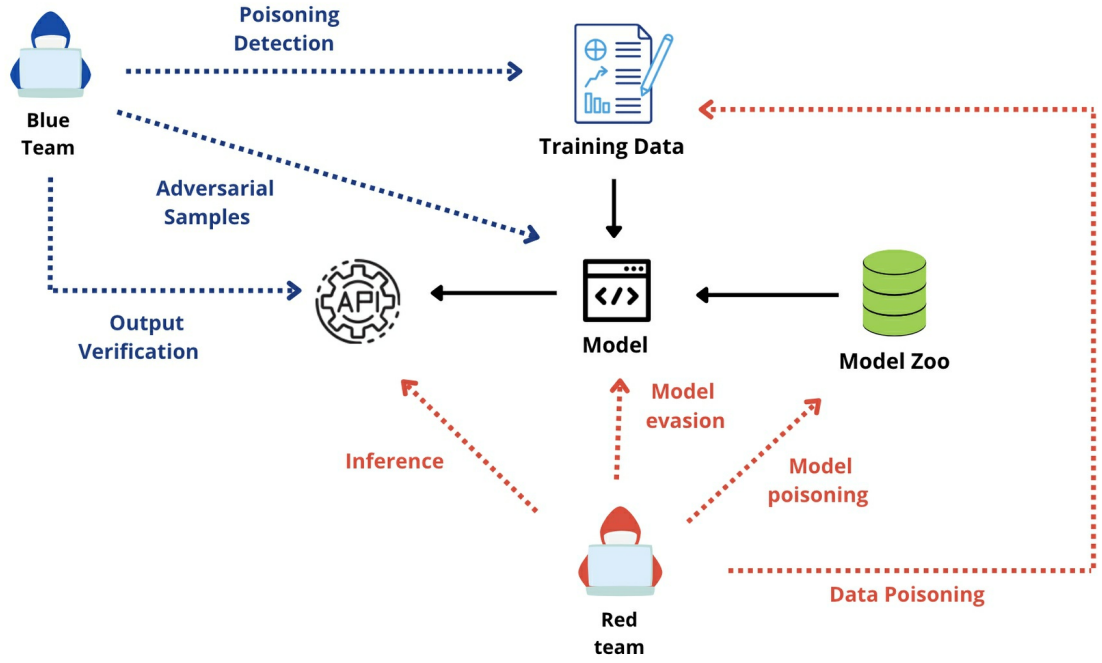


Figure 9.2: Red / Blue teaming of AI systems

Tools of the trade

Standard security tools usually do not have AI-based techniques built into them which can assess a model's vulnerability to risks like model inference or evasion. Thankfully there are free tools available that cyber-security teams can use to supplement their existing penetrating testing toolkits. These tools are open source, but you can look for commercial alternatives also.

Whichever type you prefer make sure the tools have the following features.

1. **Model agnostic:** It can test all types of models and is not restricted to any specific one
2. **Technology agnostic:** It should be able to test AI models hosted on any platform whether it is on-cloud or on-prem.
3. **Integrates with your existing toolkits**: Should have command line capabilities so that scripting and automation are easy to do for your security teams.

Some of the free tools you can find are

- Counterfit by Microsoft: Described by Microsoft as "an automation tool for security testing AI systems as an open-source project. Counterfit helps

organizations conduct AI security risk assessments to ensure that the algorithms used in their businesses are robust, reliable, and trustworthy". Counterfit provides a great way to automate and test attacks against AI systems and can be used in red teams and penetration tests. It contains preloaded AI attack patterns which security professionals can run from the command line via scripts and can integrate with existing toolkits

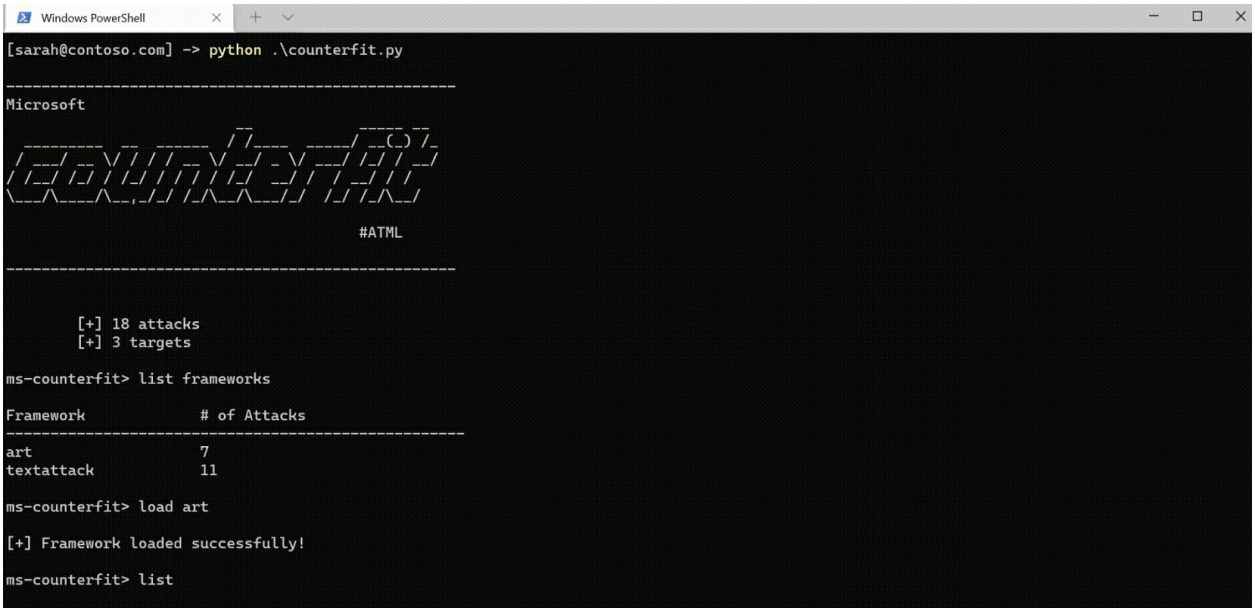


Figure 9.3: Counterfit by Microsoft

- Adversarial Robustness Toolbox (ART) is described as "*a Python library for Machine Learning Security. ART provides tools that enable developers and researchers to defend and evaluate Machine Learning models and applications against the adversarial threats of Evasion, Poisoning, Extraction, and Inference. ART supports all popular machine learning frameworks*"

For these tools to be effective make sure you map it to the ATLAS framework so that you can align it with a common standard. You can use these tools both for red teaming / penetration testing and for conducting vulnerability assessments of AI systems. Use them to regularly run scans of your AI assets and build a risk tracker of AI-specific risks. By tracking these risks over time, you can see an improvement in your security posture and monitor progress over time.

Another valuable resource to get better context and awareness of attacks would be the Atlas case studies page listed here. This page is a listing of known attacks on production AI systems and can be used by security teams to better understand the impact on their systems.

Chapter Summary

In this chapter we saw how to conduct security testing of AI systems and what is needed to create an effective AI security testing program. We also looked at the MITRE ATLAS framework which maps the popular MITRE ATT&CK framework to AI-specific attacks

Topics we covered:

- Enhancing security testing/penetration testing programs to cover AI-specific risks
- MITRA ATLAS framework for AI attacks
- AI-specific security testing tools

10

Where to go from here?

Congratulations on reaching the end of this book and hopefully now you have a firm understanding of how to govern and secure your AI systems in a real-world setting. This is just the start of your journey as AI is a vast field with new risks and threats emerging regularly and there are lots of ways to take your knowledge to the next level:

1. Get involved in AI projects in your company and try to implement some of the knowledge your learned here. Nothing substitutes experience with AI, and this is a great field to start in today
2. Download some of the frameworks/tools we discussed and customize them for your own environment
3. Subscribe to my blog and YouTube channel where I regularly discuss AI risks and threats.
4. Subscribe to my LinkedIn newsletter "**Cloud Security and AI thoughts**" here

Training and Courses

A great way to supplement this book is to enroll in my course below which I regularly update:

https://www.udemy.com/course/artificial-intelligence-ai-governance-and-cyber-security/

Get in touch

Feel free to get in touch with me on LinkedIn if you liked this book and want to discuss something. Always happy to hear from my readers!

Feedback time

Thank you for reading this book and I hope you liked it.

I would **really** appreciate you leaving me a quick review on Amazon and feedback will help me to further improve this book and grow as a writer. It only takes a few minutes, and I would be extremely grateful for the same

Feel free to reach out to me on LinkedIn if you want to connect.

I wish you all the best in your AI security journey!

About the Author

Taimur Ijlal is a multi-award-winning, information security leader with over two decades of international experience in cyber-security and IT risk management in the fin-tech industry. For his contributions to the industry, he won a few awards here and there such as CISO of the year, CISO top 30, CISO top 50, and Most Outstanding Security team.

He served as the Head of Information Security for several major companies, but his real passion is teaching and writing about cyber-security topics. He currently lives in the UK where he moved with his family in 2021.

Taimur has a Blog and a YouTube channel "Cloud Security Guy" on which he regularly posts about Cloud Security, Artificial Intelligence, and general cyber-security career advice. He has also launched several courses on Cyber-Security and Artificial Intelligence and can be contacted on his LinkedIn profile or via his YouTube Channel.