The Positive Way

WAVESTONE

ARTIFICIAL INTELLIGENCE AND CYBERSECURITY

PROTECTING TOMORROW'S WORLD TODAY

AUTHORS

CAROLE MEZIAT
carole.meziat@wavestone.com

LAURENT GUILLE
laurent.guille@wavestone.com

This publication was produced with the help of Erwan Nicolas, Cybersecurity and Digital Trust Consultant.

The statistics are alarming: 44% of FTSE100 companies talked about launching AI projects in their last annual report, and only seven linked these to cybersecurity[1]. How can future disasters be prevented—pragmatically and concretely?

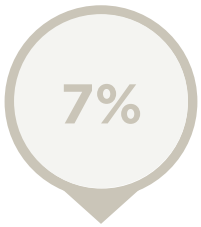
As the facts show, attacks on artificial intelligence systems are already happening. When the iPhoneX was released in 2017, Apple boasted of having created an extremely robust facial recognition system. A week later, and at a cost of less than $150, the Vietnamese cybersecurity company, Bkav, managed to create a mask that was capable of duping the application.

Artificial intelligence (AI) is revolutionising our daily lives: autonomous cars, behavioral biometry, predictive medicine, intelligent chatbots and content suggestion. New uses are appearing every day, but the cybersecurity risk management that using this new technology entails is rarely discussed.

What are the specific risks associated with AI? What questions do you need to ask before tackling them? What security solutions can be applied in this innovative area? And how can you select one and put it in place?

[1] Wavestone - Financial communication cybermaturity index - 2019 Edition, FTSE 100 findings.

of companies mention
AI or machine learning



Only mention **AI or ML in relation to cybersecurity**

mentions **AI or ML presenting a cyber risk** – the others use it to help mitigate cyber risk (mostly through predictive analytics for financial crime prevention).

ARTIFICIAL INTELLIGENCE: LOOKING BEYOND THE BUZZWORDS — WHAT'S BEHIND THE CONCEPT?

Artificial intelligence can reproduce **human intelligence**. Its scope ranges from medical analysis programmes that detect tumors, to intelligence capable of driving your car completely autonomously.

In AI applications, we can distinguish between those whose rules are fixed in advance by experts and those with the **ability to adapt their behavior** to the situation. For this second case, we use the term **machine learning**. This second type of system relies on large amounts of data, manipulated using increasingly powerful computers and analysed to automatically recognise patterns, which then serve as the basis for decisions. There are various learning methods, including supervised learning, unsupervised learning, and reinforcement learning.

These systems, with their own learning mechanisms and the ability to modify decision thresholds in an adaptive way, **offer a fundamentally different approach compared with historical systems**. In addition, AI and machine learning are often conflated, including when considering cybersecurity. This Insight is no exception, and behind the broad title of Artificial Intelligence, **it addresses, in particular, the management of cybersecurity risks related to the use of machine learning.**

NEW CHALLENGES FOR CYBERSECURITY TEAMS

Why attack AI?

AI is booming, and applications that use machine learning increasingly form part of our uses, manipulate our data, or even act physically in our daily lives. Studying the vulnerabilities of these systems can be a **cost-effective investment for attackers—who can resell the data they harvest or monetise certain decisions made**. In addition, compromising the facial recognition module of an iPhone, or deflecting the trajectory of an autonomous car, are likely to attract the kinds of attackers **who relish a technological challenge or want to gain personal recognition.**

Attackers mainly seek to:

- **Disrupt the AI application's operation,** by deliberately causing an erroneous decision using a chosen data set. One example is bypassing a facial recognition system to obtain non-legitimate logical or physical access and carry out a theft.
- **Sabotage the operation of the AI itself,** preventing or disrupting the application's operation. Here attackers aim to damage the corporate reputation, or disrupt the activities, of a target company. The 2016 Microsoft Tay chatbot attack, covered next in this Insight, was a totemic example of sabotage.
- **Understand and «reverse engineer» the model** by studying it behavior. Data processing and modeling work is often time-consuming and costly for companies, and the results represent high added value. Stealing, and then reselling, a model can be very lucrative, and can attract buyers who want to cut corners in the endless race for digital innovation.
- **Steal the data used by the application,** by querying it directly, or by cross-checking the results it provides and then attempting to draw conclusions, or even by stealing the databases the AI has access to.

How is AI attacked?

Attacks that specifically affect machine-learning-based applications can be grouped into three categories.

Poisoning... or shifting the AI's centre of gravity

This technique is the one least related to traditional methods because it **specifically targets the automatic learning phase**. With

poisoning, an attacker seeks to modify the AI's behavior in a chosen direction by influencing the data used for learning.

The attacker mainly seeks to disrupt the application's operation to their advantage, or to sabotage it.

This, for example, is what happened in 2016 to Microsoft Tay, a chatbot built by Microsoft to study the interactions of young Americans on social networks, in particular Twitter. Overnight, Microsoft Tay was flooded with abusive tweets by a group of malicious users from the 4chan forum; in less than ten hours, they shifted the chatbot's behavior from that of a "normal" adolescent to that of a rabid extremist.

All applications that use automatic learning can be affected by this type of attack. And such techniques are **particularly powerful when the data used for learning is poorly controlled:** public or external data with a high-learning frequency.

Inference... or making the AI talk

With inference, an attacker experiments, successively testing different queries on the application, and studying the evolution of its behavior.

The attacker looks either to collect the data used by the AI (in learning or in production) or steal the model (or some of its parameters).

This is one of **the most widely used techniques to disrupt the operation of security solutions**. Attackers subscribe to the solution, send multiple requests, gather the associated outputs, and use the data to train a model identical to the security solution being studied. They can then initiate attacks more easily, having gained ownership of the algorithm, and therefore access to all the information that leads to a decision taken by the solution. They will then use the adverse examples generated to propagate attacks that will not be detected.

An attacker will use all the information provided as outputs from the application to aid their attack. As a result, these techniques are all the more effective if **the application is involved in a large number of uses and the results of its decision-making are highly detailed.** For example, an image recognition application that returns a result, as well as its reliability level: «this image is a dog with a confidence level of 90% and an ostrich with a confidence level of 10%» will be much more vulnerable to this type of attack than an application that returns only the most likely answer: "This picture is a dog.».

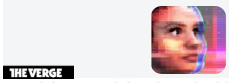
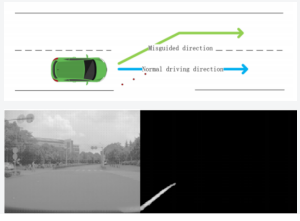
Evasion... fooling the AI

With evasion, an attacker plays with the application's input data to obtain a decision different than the one that the application would normally make. They seek to create **the equivalent of an optical illusion for the algorithm, known as an adversarial example,** by introducing a «sound» that is carefully calculated to remain discreet and undetectable.

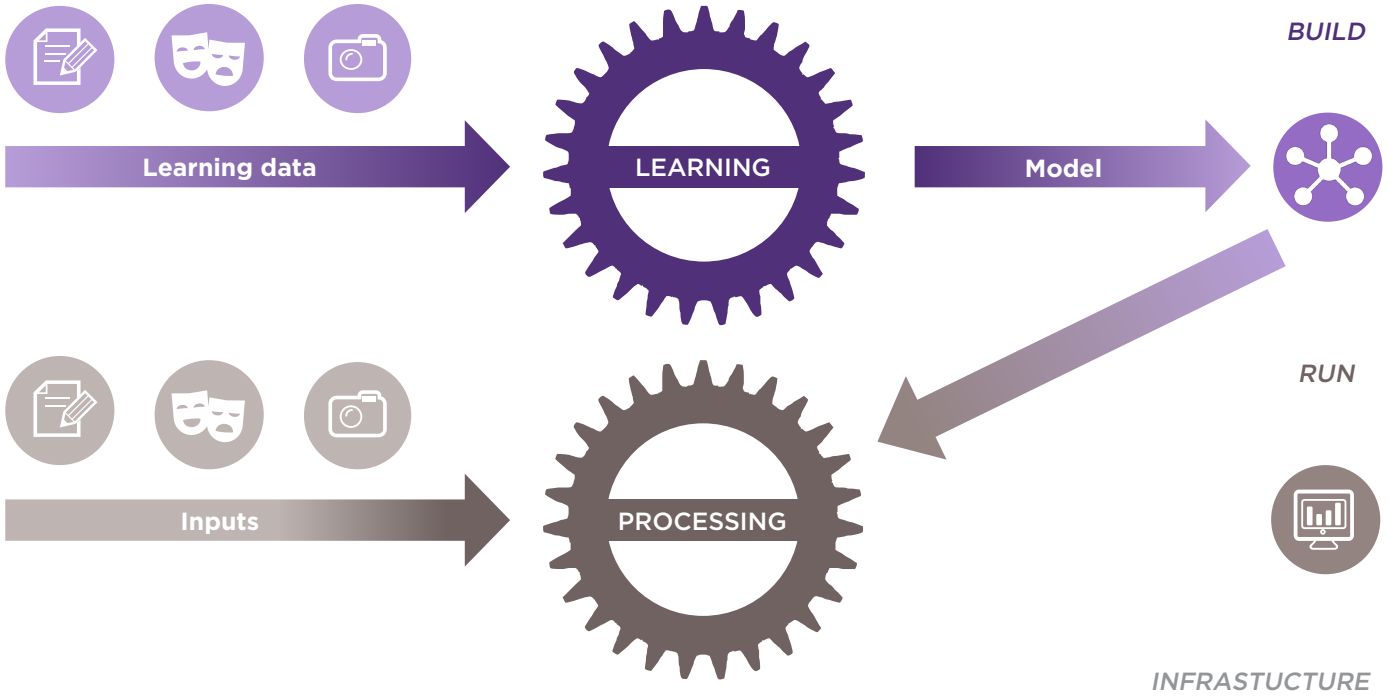
By doing this, the attacker seeks to shift the application's behavior to their advantage and target it in production, once the learning is complete.

The compromise of the autopilot software on Tesla's autonomous car by a group of Chinese researchers working for Keen Security Labs is an illustration of this type of attack. The researchers managed to deflect a Tesla S from its path using simple stickers stuck to the guidelines on the tracks used by the car. The recognition of road markings and environmental factors, gained through

SUMMARY OF THE MAIN METHODS OF ATTACK USED ON MACHINE LEARNING

Attack category	POISONING	INFERENCE	EVASION
Example of an attack	*Shifting the AI's center of gravity*  THE VERGE AI picks up racial and gender biases when learning from what humans write	*Extraction of information from the IA*  iPhone X's Face ID supposedly got hacked. We have questions A Vietnamese security software company says it tricked Face ID, which Apple touts as more secure than Touch ID. Specially processed area 2D images Silicone nose 3D printed frame	*«Optical Illusions» for AI*  Misguided direction Normal driving direction
Main motives	**Disrupting the application's functioning** **Sabotaging the application**	**Understanding and reverse engineering the model** **Data theft**	**Disruption of functioning**
Phase targeted	**Learning** *(build)*	**Learning** *(build)*, **Processing** *(run)*	**Processing** *(run)*
Aggravating factors	**Learning frequency** (continuous, etc.) **Uncontrolled data** (public, unauthenticated, etc.)	**Level of detail in the output** (provision of the level of reliability) **Exposure of the application and learning data**	**Input complexity** (images, sound, etc.) **Exposure of the application** (public places, internet without authentication, etc.)

THE FUNCTIONAL BRICKS OF AN APPLICATION BASED ON MACHINE LEARNING



machine learning, are the basis of the steering decisions autonomous cars make. The attackers managed to distort the road markings to create a situation that the Tesla S had not learned about.

The exposure here to such evasion techniques is all the more important in this case **since the application is broadly exposed (used by the general public, limited access control, etc.) and uses complex input data (images, sounds, etc.).**

This third category completes the suite of major types of attack that are particular to the use of AI. The ultimate scenario is the complete takeover of an AI system by an attacker—in order to use it as a means of attacking another target.

SIX KEY POINTS TO SECURELY PURSUE AN ARTIFICIAL INTELLIGENCE PROJECT

As we've discussed, AI solutions, which are booming and highly attractive, are far from immune to cyberattacks. The methods used are sometimes new, meaning that the security mechanisms applicable to existing systems need to be assessed, and in some cases rethought. Here are six essential areas to consider in mastering the risks related to AI business projects.

1\ Protect data at every step of the project

Data is the foundation of machine learning projects in companies. These projects require the manipulation of a large volume and variety of data. Before even talking about protecting against data leaks, it's essential to ensure that the desired use **complies with current regulations,** especially those related to data protection (the GDPR, medical data regulations, PCI DSS, etc.). This means defining, as clearly as possible, the project's purpose and the associated data processing.

This must be thought through, not only for the target use of the application, but also **for the entire development and testing phase of the machine-learning solution,** which is particularly data heavy. One of the key opportunities machine learning offers is an ability to establish correlations from large amounts of data, to an extent that a human would be incapable of producing. Models are tested with as much data as possible to determine which are the most efficient. Once the model is selected and validated, **the amount and type of data used can then be reduced to the strict minimum needed to move to production.** Therefore, in any machine learning project, there's a need to:

- **Desensitise the learning data** as soon as possible, starting with personal data. Solutions that generate fictitious (or synthetic) data sets are emerging on the market; these can produce desensitised data while conserving its statistical value. These solutions, such as Mostly AI, HAZY, or KIProtect, offer high potential to address data privacy issues.
- **Recognise the possibility that sensitive residual data might be used**, by protecting it appropriately (in terms of access rights, encryption, etc.), especially in the development phase.
- **Raise awareness among, and empower, the teams involved in model development** in terms of handling of sensitive data—and not just among data engineers and data scientists, but also in the business functions involved.

Because decisions made by machine

learning solutions are not based on rules explicitly defined by humans, but on ones learned automatically, they can sometimes be difficult, or even impossible, to explain. At this stage, then, it's also important to assess **the level of interpretability** required for the algorithms. This is especially true of algorithms that make decisions using personal data within the scope of the GDPR, a regulation which requires that any important decision, or one that has a legal nature, can be explained.

2\ Protect the big data platform

In machine learning projects, this step takes on some special dimensions. There are large amounts of highly concentrated data, and therefore particularly exposed to the risk of the theft or modification of information. The models themselves and their learning sets constitute a valuable industrial secret, something which, in fiercely competitive settings, could be highly sought after: through theft of the model or its decision thresholds, etc.

Since machine learning systems exhibit natural evolutionary behavior, subtle changes in behavior, and therefore possible malicious modifications, are more difficult to detect. It's essential, then, to **apply good practices to protect big data,** regardless of the environment: perimeter security and compartmentalisation, authorisation, and access management (on every brick), the management of privileged accounts, hardening, maintenance in a secure condition, encryption of media, traceability, and the security of suppliers and contracts.

In parallel, the development of machine-learning projects often requires the use of specific technologies not used in the company's standard development frameworks; for these, there's a **need to evaluate and validate security levels** before going into production or rolling out widely.

These building blocks, which support the initial stages of a machine-learning initiative, often serve as a basis for scaling up such services—and implementing good practices, right from the start, will **benefit all future projects.**

3\ Securing the learning process

The machine learning stage is both the key step in which the solution's effectiveness and relevance is based, and the genuinely new part of the initiative in relation to existing systems. So, it's not difficult to understand why it can be a prime target for attackers. Careful and focused thinking is therefore needed to protect this stage. Protection needs to be at two levels: at the data training level and at the learning-method level.

The algorithms are trained on a data sample known as the learning set. This data is used as the base set, from which the algorithms developed must be able to be generalised. Designing the learning set is an essential step in any machine learning project. The data set must be large enough to be generalised, sufficiently representative not to introduce bias, and readable enough for the data extraction to maximise the value added to the model. **Influencing this learning set can affect the behaviour of the machine learning application.**

Several measures can be put in place to **ensure the reliability of the learning set used** and guard against the poisoning-type attacks discussed above.

/ Initially, extending the learning time to widen the training data set as much as possible, using **advanced learning,** makes it possible to reduce the impact of each piece of input data on the model's operation; this makes the solution less sensitive to partial alterations to the learning set.

/ Next, **various control steps can be implemented throughout the learning phase:** systematic validation at several points during the learning phase by an expert professional in the field; carrying out data set integrity checks that can detect alterations; the definition of data-usage thresholds originating from the same source (localisation, individual, IP, etc.), to reduce the risks of oversampling; the definition of blacklists (key words, patterns, etc.) that need to be systematically removed from the learning set (for example, offensive vocabulary in the case of a chatbot); the detection of changes in model behaviour in the course of retraining (the detection of abrupt changes from one training session to another, comparisons of developments in the frequency or number of regular retrainings, etc.).

> Projects of a new kind, which require creative thinking about risk scenarios and the protective measures to deploy

It's clear that the less the learning set is well managed (where public data or data from an external supplier is used) and stable (one-off vs. continuous learning), the greater the risk of compromise, and, therefore, the more important protective measures will be for the learning set. The key is to **properly balance the business requirements for quantity, variety, and up-to-date learning data with the need to secure and control the data.**

In parallel, there's also a need to define safe learning practices, which, until recently, have been virtually non-existent. **Securing learning methods is a whole new field in cybersecurity research.** The discipline is about defining which algorithms to prioritize from a security point of view, how to use them, and which options to select when developing such algorithms. Examples of relevant good practices that can help build more robust solutions are: RONI (Reject on Negative Impact), which enables data that has a negative impact on the precision of the model to be deleted from the learning set, and Bootstrap Aggregating, which allows models to be stabilized

4\ Securing the application

Many attempted attacks can be contained by **applying the secure development best practices that are already widely deployed in companies** (for example, Open Web Application Security Project [OWASP] rules). This is all the more important because data scientist profiles have developed from statistical rather than computing-based backgrounds; the scientists often have little awareness of security issues compared with more traditional developers, and projects are typically carried out directly by business functions independently of the IT teams, who are more familiar with managing projects' security aspects.

However, these measures aren't enough to protect against all types of fraud related to the use of machine learning. Most machine-learning-specific security measures focus on three areas: managing inputs, making processing reliable, and controlling outputs.

Managing inputs

- **Protecting the data acquisition chain:** here it's essential that no one can modify the data— from the moment it's acquired to the point when it's supplied to the algorithm to make the predictions. To ensure this, the data processing chain must be designed to guarantee end-to-end protection—by limiting the access channels available to users, applying strict access-management controls, and encrypting the associated data flows.
- **Filtering the input data:** this task is similar to filtering input data in traditional web applications. It requires verifying the format (checking the type of data, the completeness of the information entered or extracted, etc.) or the consistency of input data (differences compared with the anticipated data, historical data ,etc.) by detecting noise (weak signals), and then rejecting or cleaning the data before the application processes it. For example, noise detection (Noise Prevention), in particular, is used to protect image recognition applications against attacks.
- **Detection and the blocking of suspicious user behaviours:** this consists of setting up mechanisms that can detect attempted inference attacks; for example, by detecting a large number of similar requests, or requests from the same source, by detecting successive anomalies in the format of inputs, etc. User and Entity Behaviour Analytics (UEBA) is also a domain with major applications for using machine learning to improve cybersecurity.

Making processing reliable

- ***Adversarial Training:*** this method is applied during the learning phase, before production; it consists of teaching the model examples of possible attacks and associating decision-making with them. This technique is widely used to recognise images, using Generative Adversarial Networks (GANs) to automatically generate conflicting images that include noise, and make models more robust to evasion attacks.
- ***Randomisation*:** this method serves the same purpose as adversarial training and consists of adding a random noise to each unit of data. Adding such random noise before the data is processed makes it more difficult for an attacker to predict how to disrupt each entry in order to achieve their goal. The algorithm is trained on data to which this type of random noise has been added; noise intensity is optimised to strike the best balance between algorithm accuracy and resilience against evasion attacks. In contrast to other algorithm reliability techniques which offer only empirical guarantees, randomisation's effectiveness is mathematically proven. The results it achieves against evasion attacks are comparable to adversarial training, but the randomisation is much less costly in terms of computing time.
- ***Defensive Distillation*:** this technique involves using two successive models to minimise decision errors. The first model is trained to maximise the algorithm's precision (for example, being able to determine with 100% probability that an image is a cat rather than a dog). The second model is then driven using outputs from the first algorithm that have low levels of uncertainty (for example, outputs that represent a 95% probability that an image is a cat rather than a dog). The second, «distilled" model, makes the task of guessing the algorithm's operating mode more complicated for attackers—in turn making it more robust. This technique puts one more obstacle in an attacker's path; although by investing more time and resources, they would likely be able to analyse, understand, and eventual-

LEARNING BIAS, ETHICS AND REPUTATIONAL RISKS

The influence of data sets and the associated risks of learning bias, go well beyond cybersecurity issues: they also raise ethical issues. If considerable care isn't taken, natural learning biases can be introduced, which can have extremely damaging reputational consequences. An example is Google Image's image-recognition algorithms, which, in 2015, classified photos of black people as gorillas.

The European Commission is taking a close interest in these topics and has recently published recommendations on the ethics of AI (Ethics Guidelines for Trustworthy Artificial Intelligence—which can be found on the Commission's website).

ly compromise the solution's operation.

- ***Learning Ensemble:*** to make predictions reliable, several models can be used simultaneously, with each based on different algorithms and approaches, which can combine and/or compare their respective results before making a decision. This technique is costly in time and resource terms and works on the assumption that not all algorithms will be sensitive to the same variations in input data; therefore, an attack on one algorithm may well be ineffective on another. This assumption is, however, being questioned today: research shows that models which use different approaches, yet are trained on the same data, are sensitive to the same types of evasion attack. For the most sensitive use cases though, employing this technique can make an attacker's life more difficult—by making the model's decisions more opaque.

Safeguarding your outputs

- ***Gradient Masking:*** this aims to reduce the risk of reverse engineering of models by limiting the "verbosity" of the machine learning application, especially in terms of decision scores used by the algorithm. It reduces the risk of the algorithm being reverse engineered via an iterative game played on its input parameters
- ***Protection of the output chain:*** just as with the data acquisition chain, the set of predictions an algorithm makes needs to be protected against access attempts. Only the result of a decision should be accessible. Consequently, the same techniques are employed: encryption, access control, etc.
- ***Detection of suspicious outputs:*** as systems based on machine learning are progressive, decisions may evolve despite an apparently similar context (depending on the time, individual, etc.). In order to limit disruptions, checks can be put in place; for example, by comparing results against a benchmark and raising an alert when doubts arise. For example, before making a bank transfer to a specific recipient, a check can be made to verify that the amount isn't over ten times the average of the amounts paid to the same recipient in the previous year. This enables both errors, and potentially malicious involvement, to be detected. The way in which these anomalies are handled must then be decided on a case-by-case basis: halting the processing, a request for re-authentication, teams being alerted, etc.
- **Moderation and blacklisting** dof output data: as a check on model outputs, a list of prohibited answers can also be put in place after the results have been generated, in addition to manual moderation. Applying automatic constraints to the outputs is also an option; for example, forcing the values to remain within authorised ranges.

5\ Defining your risk management and resilience strategy

As discussed in the introduction, the term, artificial intelligence, encompasses a very wide range of applications—**and not all of them have the same degree of sensitivity as an autonomous car.** Take smart chatbots, for example: the level of risk involved for a passive chatbot that provides personalised advice in response to questions such as «What will the weather be like this afternoon?" will be lower than that of a transactional chatbot that is capable of creating an online payment card. It's therefore also essential to conduct a risk analysis, in order to know where to prioritise your efforts. Such analysis must take into account the specific risks related to using machine learning (which were set out in the first part of this Insight).

In order to adopt approach at large scale, it's important to develop **security guidelines according to the type of AI project.** For example, the guidelines can be structured by input data type (image, speech, text, structured data, etc.), by learning frequency (continuous, ad hoc, or regular) or by the solution's level of exposure (public facing, internal, etc.). You may need to invest time and seek expertise to address the relatively new issues being discussed here.

As no system is immune from threats, it's important to anticipate attacks and define **a resilience strategy that is designed for machine learning's particularities:**

Consider approaches that will ensure high availability and safeguarding of the application. In particular, anticipate that the application could be sabotaged during learning, compromising the data used and rendering the model useless. It's vital to **keep a copy of the model's previous states and be prepared to restore it to these, if necessary.**

Think about the audit trail needed **if an investigation has to take place.** Machine learning sometimes produces a black box effect, limiting the ability to understand how a model came to a particular conclusion; this will act as an obstacle to investigating an incident. Traditional good traceability practices doesn't take account machine learning's complexity into account, so it's essential to define specific, machine-learning traceability requirements, and to **retain an audit trail of the parameters that influence the decisions the algorithms make.** Adopting practices like these can, for example, facilitate cooperation with the courts, if it comes to that.

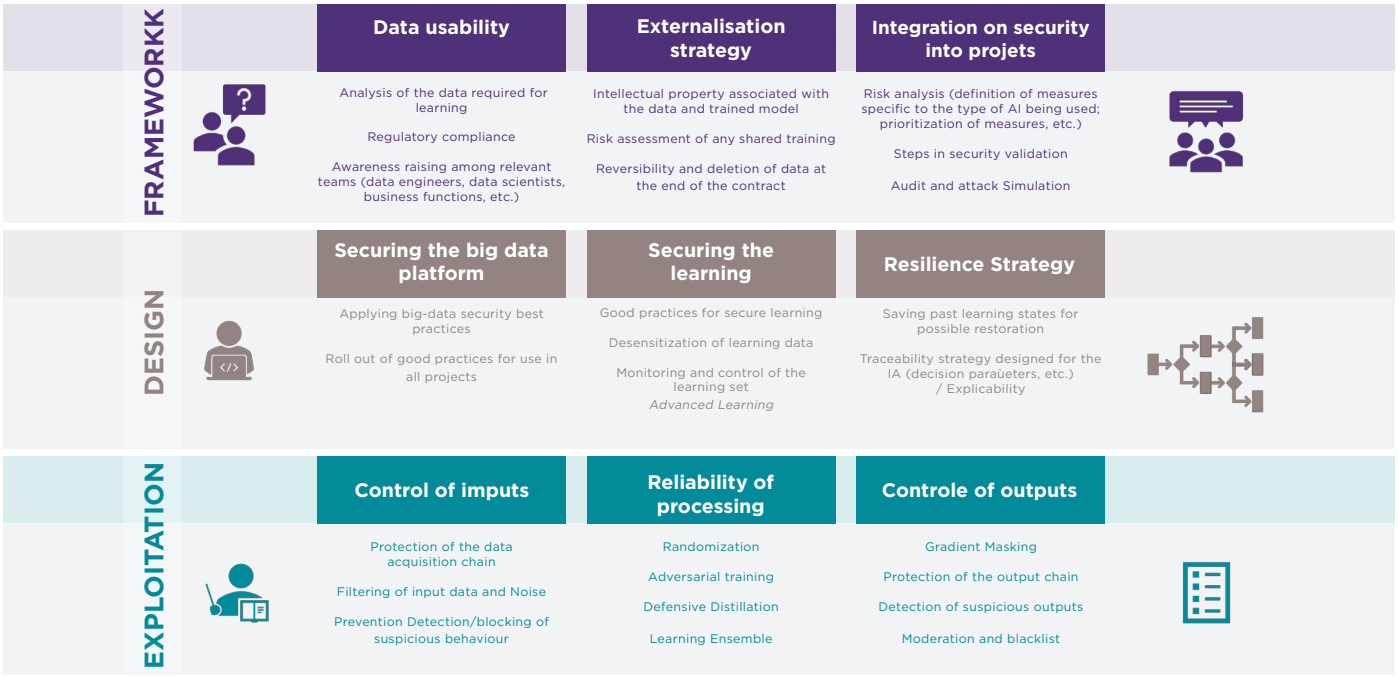
6\ Think carefully before outsourcing

In some settings or use cases, AI solutions are delivered by external providers. Opting for an external solution always carries risks, and, over and above the need to **ensure publishers consider the points listed above,** there are other additional areas, again specific to machine learning, that need to be covered during the contracting phase.

The first is **intellectual property:** a question that concerns the ownership of both the data and the trained model. Who owns the trained model? Who's responsible for it? At the end of the contract, who will take it over? And in what form?

Asking these types of questions is critical

THE SECURITY MEASURES TO APPLY AT EACH STAGE OF AN AI PROJECT



before making a decision to outsource. The AI market is awash with new, innovative, and highly specialised solutions, and market consolidation is highly likely in the years ahead. What might happen if a direct competitor decides to invest in one of these innovative technologies—and buys out a supplier who's been working as your partner for the last few months? Such risks must be mitigated at the contracting stage if you are to avoid being faced with a fait accompli later.

A second area to watch closely concerns exposure to a risk mentioned above. One of the advantages of using an external solution is the possibility of a model trained on a larger amount and wider variety of data—because the supplier's solution benefits from being able to train it using data from several of its customers. This means the model may be more powerful than one trained on data from a single customer only. But this same advantage raises the question of whether **Chinese Walls are needed between the different customers who use the application.** If this becomes an issue in any context linked to sharing during machine learning, the task becomes more than a question of checking whether the supplier is applying good practice in compartmentalising its infrastructures and applications: **it's also a matter of determining whether the sharing aspects of model training could lead to the disclosure of confidential data or customers' personal information.** Questions like this might arise, for example, when bringing a new customer within the scope of learning activities: here, the new data used could alter the solution's decisions in a way that means certain information about the customer could be deduced. Such sharing can also have other consequences, like a divergence in the model's behaviour when it's trained on a new data set following the introduction of a new customer. It's a matter, then, of paying careful attention to these types of problems when you select the solution, and of consciously striking the best **balance between the business benefits that shared training might bring and the associated security and privacy risks.** There's also a case for including specific, risk-limiting, clauses in contracts, for example maximum-weighting clauses for the involvement of different customers in terms of using their data in learning, or putting in place performance monitoring indicators for the machine learning solution, which can be used to verify that the model improves over time.

Finally, it's also essential that, before entering into any agreement, build in **the need for reversibility** at the end of a contract with a supplier. Beyond the intellectual property issues already discussed, the need to recover or remove training data or learned rules must be clearly specified in the contract. And the questions may be more subtle than those for more conventional solutions; for example, in the case of shared training: "Do I have to specify that the provider's model is retrained, without my data, once the contract is over?" or "Do I want to retrieve the rules from the vendor's analytics engine at the end of the contract?" If yes, "What must I specify that the supplier puts in place so that I can operate these rules with my infrastructure?" "Will we have the in-house technical skills to exploit them?" or "Should I make provision for a transfer of machine learning skills in the contract?"

A PREREQUISITE: ENGAGE YOUR BUSINESS FUNCTIONS TO HELP PROTECT YOUR NEW ARTIFICIAL INTELLIGENCE SYSTEMS

Many of the AI initiatives being pursued by companies today are still in their research or experimental stages. Their initial objective is to demonstrate AI's value to their various business lines. These proof of values (POVs) are often done on specific use cases,

based on the company's existing data. **Timescales are generally short to be able to demonstrate an evidence-based return on investment in the space of a few weeks.** Security rarely features strongly in these experimental phases. However, sensitive data is often already being used, and the production deployment schedules that will follow a successful POV is also likely to be short, which makes integrating security measures challenging. In addition, rolling out the solution to other areas, or extending the use cases, often happens quickly too.

Against this backdrop, planning for the need to integrate upstream security measures is key, if you want to keep pace with scale-up requirements and ensure that cybersecurity teams take them into account.

It may help to take the lead and **produce a position paper or overarching security memo on the topic.** This will raise awareness among decision makers and the business functions about security issues relevant to AI projects even before an initiative starts; it also helps forge the right mindset throughout the corporate ecosystem.

At the same time, **identifying and categorising current** and future corporate **AI experiments and projects** enables the process of defining the security measures to begin, which needs to be integrated by prioritising the company's actual needs: "What types of data do we handle: (sensitive, personal, health-related, etc.)?" "Are the solutions to be developed in-house or outsourced to specialist suppliers?" "What types of data will be used (text, images, sound, etc.)?" Depending on the answers to these questions, the security measures needed may be very different.

Then, if this audit reveals that a large number of sensitive projects are underway, or a significant increase is expected in the future, the set of projects will need to be more closely managed. Being able to call on employees with **cybersecurity profiles, including data science and the ability to define concrete security measures** (logs, backups, assessment of approaches being used, etc.), is rapidly becoming essential.

THE NEW MARKET SECURITY STANDARDS BEING DEVELOPED

Because uses of AI are in their early stages, maturing the security approaches is hampered by the technology's newness. Many attack and defence mechanisms are still at the stage of theoretical research.

Yet, numerous initiatives are emerging to define the standards that will govern the security of future AI applications. It makes sense to pay close attention to the activities of the think tanks and working groups active in the area in order to supplement your AI security guidelines with concrete measures as standards emerge; these include the AI Security Alliance (in the US), the Information Commissioner's Office AI Auditing Framework (in the UK), and the Centre for European Policy Studies (in the EU).

THE NEW THREAT TO PREPARE FOR: DEEPFAKE

In March 2019, duped by an artificial voice that imitated its CEO, a company was frauded out of €220,000. This incident, which came to light in the summer of 2019, created a considerable buzz. The head of an energy-sector company received a call from the CEO of its German parent group asking him to make a transfer of €220,000 to the bank account of a Hungarian supplier—a transfer he then made. At the other end of the phone was, apparently, **a synthetic voice**—one created by **an AI-based, voice-generation software** that could imitate the CEO's voice.

While this claim is still to be confirmed, with doubts over whether the voice really was AI-generated, such an attack is all the more likely to happen today given the advance of systems that can replicate voices or even video footage—the **well-known "Deepfakes."** And these attacks are going to be difficult to control. Why would you suspect anything when you hear your boss's familiar voice?

WHAT IS DEEPFAKE?

Deepfake is the modification of images, audio, or video, using AI (and especially «deep learning») to present a fake version of reality.

A Deepfake file is created using two competing AI systems: one known as the generator, and the other the discriminator. The generator creates a false file (image, audio, video, etc.) and then asks the discriminator to determine if the file is real or false. Together, the generator and the discriminator form what is called a **Generative Adversarial Network (GAN).**

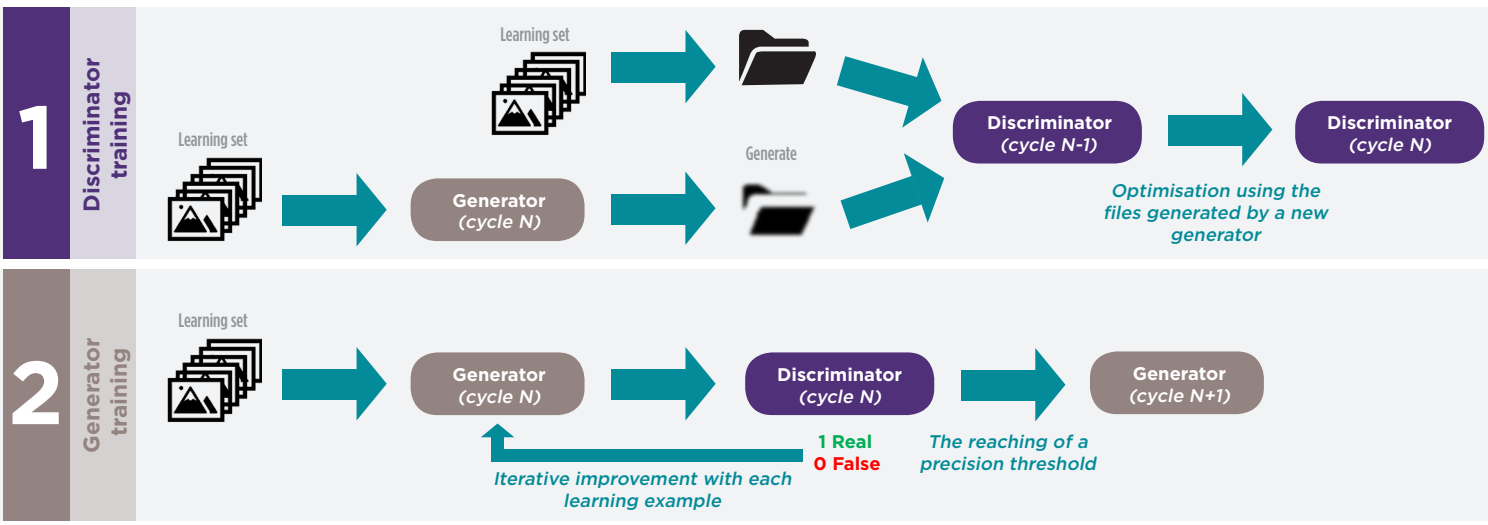
A training data set is provided to the generator to initiate the process. The method then works in two-step cycles:

1. **Discriminator training:** The discriminator is trained to differentiate between real files and the fake files created by the generator.

2. **Generator training:** The generator creates files, which the discriminator then evaluates. This enables the generator to improve by producing more and more credible files that the discriminator eventually considers real.

Once the generator has progressed sufficiently in creating credible files, the discriminator is trained again in order to differentiate between real files and further fake files created by the generator (a new "Step 1"). The retrained discriminator is then used to further develop the generator as part of a new "Step 2." This cycle is repeated as many times as necessary to achieve the desired level of accuracy.

THE FUNCTIONING OF A GENERATIVE ADVERSARIAL NETWORK



There are different forms of Deepfake, including:

- **Deepfake audios,** which mimic the voice of a targeted person using samples of their voice. These make it possible to make the person "speak" a given text input
- **Face-swapping,** which replaces the face of a person in a video by that of a person being targeted, using a photo as raw material.
- **Deepfake lip syncing,** which in video, adapts the targeted person's facial movements using an audio file of another person. It enables them to "speak" the words in the audio file, even though they have never uttered them in reality.
- **Deepfake puppetry,** which generates footage of a targeted person using a video of an actor as an input. This makes it possible to create a video where the targeted person gives a speech that is actually spoken by the actor. A famous example is a video of Barack Obama speaking the words of Jordan Peele, in which the former president's gestures are reproduced in a highly realistic way.

TECHNIQUES THAT ARE AVAILABLE TO ALL

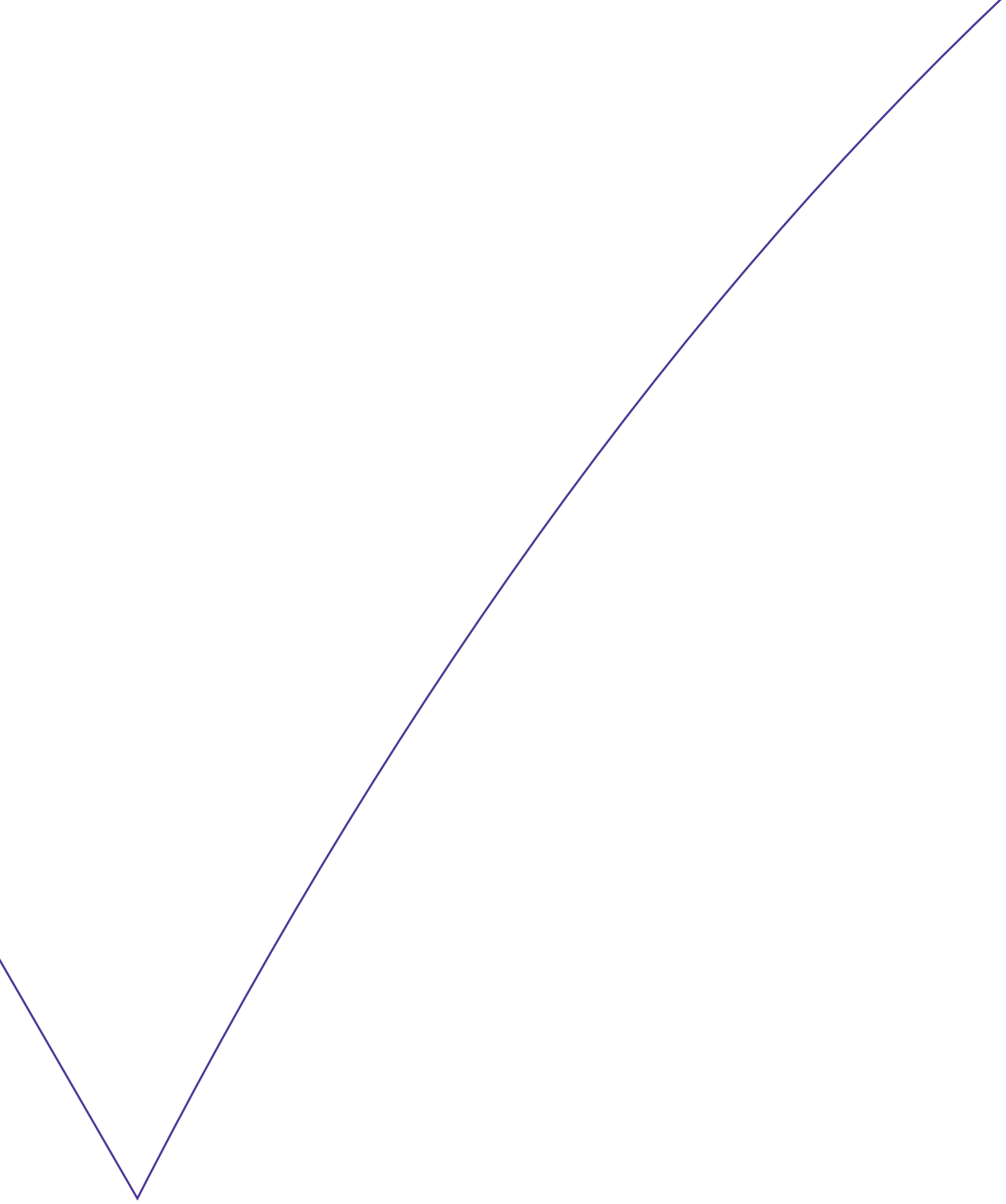
The boom in machine learning is resulting in increasingly powerful algorithms and the democratisation of access to them; examples are mass market applications like Lyrebird (a free application for Deepfake audio) or Zao (a Chinese face-swapping application that caused a stir on its release, in the summer of 2019). These applications, which were created for entertainment purposes, represent a new and powerful tool: something accessible to all—and easy to use for those with malicious intent. There's no doubt that the number of attacks using such applications will increase in the coming years. Faking the words of a president, the violation of images, the creation of false legal certificates, bypassing biometric authentication, etc.: the range of possible use cases is frightening.

DEEPFAKE: DEEPLY FATAL?

Given the outlook for the risks, solutions are being actively sought to safeguard against the use of Deepfake. It's a complex area, and one being taken very seriously. In the run-up to the 2020 US election campaign, **Facebook** launched its **"Deepfake Detection Challenge,"** a public competition to develop anti-Deepfake tools—with no less than $10m on offer for the winner.

The manipulation needed to create Deepfakes leave clear signs. **Non-natural elements,** such as the number of eye blinks, the relative orientation of facial elements, or distortions, **can all be detected.** Such methods target the shortcomings of Deepfake techniques and will remain effective until attackers can master and adapt their methods to remain below detection thresholds.

Preventive solutions are also being studied, such as the creation of **anti-Deepfake filters** which are applied to media prior to the publication of content. These introduce "noise" into an image which is undetectable to the naked eye but disturbing to the learning process of Deepfake algorithms (something that works along similar lines to the adversarial examples discussed earlier). Other solutions use innovative ways of guaranteeing file authenticity, like Amber Authenticate, a Blockchain-based solution.



The Positive Way

WAVESTONE

www.wavestone.com

In a world where knowing how to drive transformation is the key to success, Wavestone's mission is to inform and guide large companies and organizations in their most critical transformations, with the ambition of a positive outcome for allstakeholders. That's what we call «The Positive Way ».

Wavestone draws on some 3000 employees across 8 countries. It is a leading independent player in European consulting, and the number one in France.

Wavestone is listed on Euronext Paris and recognized as a Great Place to Work®.