INTERPOL

unicri
United Nations
Interregional Crime and Justice
Research Institute

RESPONSIBLE
AI INNOVATION IN
LAW ENFORCEMENT
AI Toolkit

Technical Reference Book

Funded by
the European Union

REVISED FEBRUARY 2024

INTERPOL

unicri
United Nations
Interregional Crime and Justice
Research Institute

Funded by
the European Union

DISCLAIMER

The contents of this document are for information purposes only. INTERPOL and UNICRI assume no liability or responsibility for any inaccurate or incomplete information, nor for any actions taken in reliance thereon. The published material is distributed without warranty of any kind, either express or implied, and the responsibility for the interpretation and use of the material lies with the reader. In no event shall INTERPOL or UNICRI be liable for damages arising from its use.

INTERPOL and UNICRI take no responsibility for the content of any external website referenced in this publication or for any defamatory, offensive or misleading information which might be contained on these third-party websites. Any links to external websites do not constitute an endorsement by INTERPOL or UNICRI, and are only provided as a convenience. It is the responsibility of the reader to evaluate the content and usefulness of information obtained from other sites.

The views, thoughts and opinions expressed in the content of this publication belong solely to the authors and do not necessarily reflect the views or policies of, nor do they imply any endorsement by, INTERPOL or the United Nations, their member countries or member states, their governing bodies, or contributory organizations, nor does it imply any endorsement. Therefore, INTERPOL and UNICRI carry no responsibility for the opinions expressed in this publication.

INTERPOL and UNICRI do not endorse or recommend any product, process, or service. Therefore, mention of any products, processes, or services in this document cannot be construed as an endorsement or recommendation by INTERPOL or UNICRI.

The designation employed and presentation of the material in this document do not imply the expression of any opinion whatsoever on the part of the Secretariat of the United Nations, UNICRI or INTERPOL, concerning the legal status of any country, territory, city or area of its authorities, or concerning the delimitation of its frontiers or boundaries.

The contents of this document may be quoted or reproduced, provided that the source of information is acknowledged. INTERPOL and UNICRI would like to receive a copy of the document

OVERVIEW

WHAT

The Technical Reference Book contains technical explanations of key concepts and terms in AI, as well as in associated fields. The aim is to establish a shared understanding for the purposes of the Toolkit for Responsible AI Innovation in Law Enforcement and to shape realistic expectations around the potential risks and benefits of the use of AI in law enforcement.

WHEN

The Technical Reference Book is designed to help readers/users to better understand the technical aspects of the other resources in the AI Toolkit. It should therefore be consulted as and when necessary as readers/users work through the AI Toolkit, in order to clarify unfamiliar concepts and terms or for further exploration of the specific technical elements presented within the AI Toolkit.

WHO

Anyone in a law enforcement agency using the AI Toolkit or its resources may benefit from the additional technical guidance contained in this document. The Technical Reference Book has been developed for individuals who have no knowledge in AI, but it also contains more detailed and technical information to allow more advanced readers to build upon an established foundation.

Table of Contents

Basic definitions

Artificial Intelligence (AI) and many of the related terms in this growing field, such as machine learning and deep learning, are becoming increasingly widely used – both in personal and professional settings. In the context of law enforcement, AI can be extremely advantageous but, without a technical background, this subject can often be intimidating and may appear impenetrable to law enforcement officers.

The starting point to build the necessary background knowledge is understanding the fundamental definitions of terms used in the AI field. This is especially important because there are many popular misconceptions about AI. The most prevalent misconceptions are that AI, machine learning and deep learning are all the same thing, that all algorithms are AI, or that AI is synonymous with so-called **artificial general intelligence** (AGI) or even **artificial super intelligence** (ASI). In this section, we will clarify the meaning of all these AI-related terms, their differences and how they relate to one another, in order to gain a better understanding of the AI field.

WANT TO LEARN MORE?

Long history of AI

Although the AI field has exploded recently, the concept itself is not new and has its roots in the 1950s, during a conference in Dartmouth where the term "Artificial Intelligence" was given to a new field of research.[1] Due to the limited storage and processing capacities of computers at the time, investment in research and development in the field was limited for several decades, resulting in what became later known as the "AI Winter".[2] Recent technological developments, particularly over the course of the last decade, have provided cheaper hardware solutions for mass data storage and faster processing. These developments, along with increased connectivity which allowed access to large volumes of data, as well as a growing community of human experts, have provided all the necessary ingredients for a new era of AI.

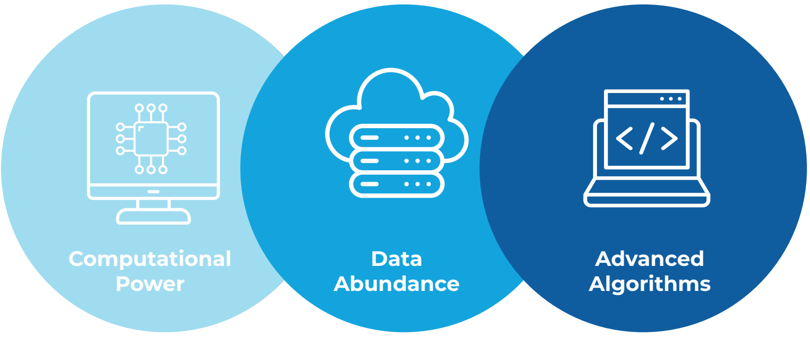


Figure 1. The three key ingredients that made AI development possible.

ARTIFICIAL INTELLIGENCE

Although the concept of AI is frequently cited, there is in fact no universally accepted definition of the term. It is, however, generally understood to describe a discipline concerned with developing technological systems that simulate human skills such as visual perception, speech recognition, decision-making, and problem-solving. It is considered a subfield of computer science with connections to other disciplines such as robotics, statistics, linguistics, cognitive science, psychology, and even philosophy.

While AI is a field of study, an “AI system” or “AI technology” is a product of that field, i.e., a practical application of AI designed to perform a specific task.

WANT TO LEARN MORE?

A more technical definition of an AI system

AI systems are a combination of software and hardware that operate with a certain degree of autonomy, executing actions to achieve a given goal. The system-generated outputs are the predictions, recommendations or content which can help humans to make better decisions.

The software includes AI algorithms. The algorithm decision-making processes may be based on data (representing the environment which the decision will affect) or on instructions and symbolic rules provided by a human.

The AI algorithm runs on computers. Given the high processing demands of some AI tasks, specialized hardware is required to improve performance and efficiency.

The data may be generated or collected from the environment using specific hardware such as sensors. AI systems can adapt their behaviour by analysing the way the environment was affected by their previous actions, which requires collecting data in real time. AI systems can also act in the physical environment through specialized hardware including robotic arms.

AI systems, therefore, comprise both software and hardware, where the software defines the algorithm, and the hardware provides the processing power and storage necessary to run the algorithm and any additional devices which may be used to capture and act on the environment.

AI became popular due to advances in machine learning which made the analysis of vast amounts of data significantly easier and faster. Machine learning models can analyse large amounts of historical data and extract patterns to create rules that can then be used to make “predictions” about the future. However, these outputs or predictions are not some kind of magical crystal ball that can see into the future, and they may not necessarily happen - they are purely extrapolations based on the data, context and situation. **AI systems’ outputs are just probabilistic outcomes, not facts.** In any law enforcement use of AI systems, law enforcement personnel must be aware of this important difference and the system should only be used as a tool to support decision-making by officers. For this reason, the **human in the loop** concept is particularly relevant in the context of law enforcement.

▶ *Learn more about the importance of the human in the loop concept in the* ***Principles for Responsible AI Innovation.***

PRACTICAL EXAMPLE

Police patrol robots – a complete AI system

A police patrol robot is an example of an AI system combining **software** and **hardware**. Police patrol robots can be used for security patrols and environment monitoring, allowing access to situations that may be inaccessible or dangerous for law enforcement personnel. Police patrol robots can manipulate objects as well as gather data through the integration of sensors. These sensors can be cameras to be able to move avoiding obstacles but can also include other sensors, such as temperature detectors and microphones, which can provide a better sense of the surroundings and collect more information to transmit to the operators in the command centre.

In this example, the AI system includes hardware such as the robot body, processors and sensors that capture, for example, images from the environment. It also includes software such as an AI algorithm that performs image processing and identifies objects that obstruct the robot's path, and another algorithm that helps decide which direction should be taken to avoid obstacles.

When people talk about **AI**, they usually refer to **machine learning** or even **deep learning** since these are the most popular types of models and the ones that generated more discussion in recent years. However, from a technical point of view, it is incorrect to use these terms interchangeably. AI describes a general domain, whereas machine learning and deep learning are both subfields within this domain, describing specific types of AI algorithms. Besides machine learning algorithms, AI systems can also use rule-based, symbolic and agent-based methods, evolutionary/genetic algorithms, or others to learn and mimic human intelligence.

There are also other relevant fields that play a role in the development of AI such as **data science**. While machine learning uses data to learn and improve performance on a specific task, data science is a broader field that studies data and how to extract meaning from it, including techniques such as data management, descriptive analytics and data visualization.

All these terms are connected as described below:

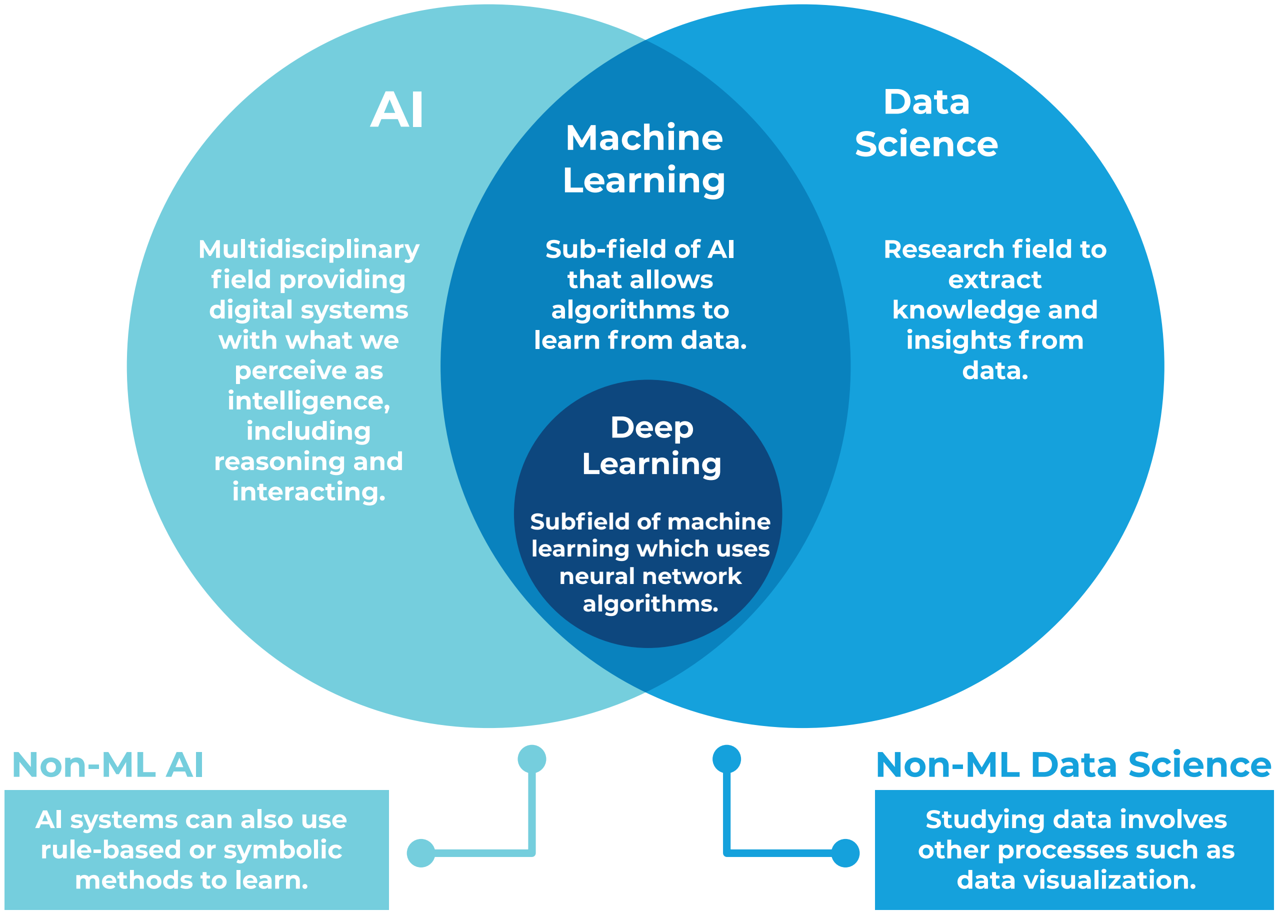

AI
Multidisciplinary field providing digital systems with what we perceive as intelligence, including reasoning and interacting.
Machine Learning
Sub-field of AI that allows algorithms to learn from data.
Data Science
Research field to extract knowledge and insights from data.
Deep Learning
Subfield of machine learning which uses neural network algorithms.
Non-ML AI
AI systems can also use rule-based or symbolic methods to learn.
Non-ML Data Science
Studying data involves other processes such as data visualization.

Figure 2. Relationship between AI and notable subfields.

The AI systems that exist today are all what are known as **narrow AI**. This means that they are programmed for one specific objective, such as finding the best way between point A and point B, and cannot be used for purposes other than those they have been specifically created for – or even for the purpose they have been created for when a small change occurs in the environment. For example, a licence plate recognition system that has been trained to read alpha-numerical licence plates will be unable to function properly if it encounters licence plates with symbols. In short, these systems can be exceptionally good at identifying and picking up patterns, but they are not well suited to adapting to a changing world.

It is also important at this stage to distinguish two further concepts that have already been briefly mentioned in the introduction, namely **artificial general intelligence** (AGI) and **artificial super intelligence** (ASI). Currently both of these concepts are confined exclusively to works of science fiction, but significant work is underway to develop more advanced models that are moving towards AGI. To avoid any misunderstandings about the type and nature of AI systems discussed in the *AI Toolkit*, both of these concepts will be briefly discussed here.

Firstly, AGI. Unlike narrow AI, an AGI system would not be trained for a specific purpose, and would have a form of intelligence more comparable to that of a human. In this regard, it would be able to understand, plan, reason, communicate in natural language, and integrate all of these skills and apply them to a broad range of tasks. An AGI system would also have the ability to learn and adapt to new tasks.[3] AGI is something that has been pursued and actively researched by experts in the field of AI for a long time, with various opinions on *whether* – and, if so, *when* – it can be achieved.[4] However, the current rapid development of large generative AI models appears to increasingly indicate that such generally intelligent systems could conceivably be possible at some stage.

Secondly, ASI. If AGI has yet to be created in any meaningful form, the concept of ASI is even further away. ASI refers to an AI system that would not only be comparable to human intelligence but would, in fact, *surpass* human intelligence in all aspects.[5] From creativity to problem-solving, super-intelligent machines would overcome human intelligence in terms of individuals and society. ASI has generated extensive philosophical debate, with some arguing that it may even present an existential threat to humanity.[6]

MACHINE LEARNING

Moving deeper in this exploration of the AI field, we turn to machine learning (ML), which, as previously noted, is a subfield of AI rather than a synonym for the term. Machine learning is a technique that teaches computers to do what comes naturally to humans – learning by example. More specifically, machine learning is a computational method that finds patterns in data when given a specific goal (such as the classification of objects as yellow or red). Once the pattern has been learned, the template for that pattern can be applied to new data to make predictions.

This differs from **traditional** or non-ML-based computer software, which operates through explicit programming or symbolic rules that consist of a set of instructions. Such instructions could be, for instance: *if variable X has a value of less than 5, classify object as A. If not, then classify object as B.* A machine learning algorithm, on the other hand, extracts patterns and implicit rules from a number of examples included in a database. After analysing several data points, the algorithm may conclude that: *if variable X has a value of less than 4.987, the object should probably be classified as A, and if it has a value between 4.988 and 99.884, the object is more likely to be classified as B.*

WANT TO LEARN MORE?

A more technical explanation of machine learning

Machine learning uses a variety of algorithms that learn from data in an iterative fashion to predict outcomes. The algorithm ingests a large number of data points or training data, and shapes a numerical model based on that data. A machine learning model is, therefore, the computation that generates a specific output for a given input. When training the machine learning algorithm with data, more precise models are created, and its accuracy tends to increase. After training, when a model is provided with an input it has never seen before or a testing data point, it estimates an output or a prediction based on what it has learned from the training data.

In the law enforcement context, there are also cases where the machine learning algorithm is pre-trained on a trial database and then re-trained on a real database of law enforcement cases. This is the case for facial recognition technology systems. First, the algorithm is trained to recognize faces and match two similar faces. Then, it is re-trained on a real criminal database where it learns to recognize the faces of suspects.

Machine learning can be extremely useful in helping law enforcement officers recognize patterns. Large amounts of information are collected during any investigation, including witness statements, forensic reports, crime scene photos etc. Analysts usually work with national or local data and information, making use of rudimentary search and analysis tools to generate intelligence. However, while an officer usually uses these tools to work on a single investigation, machine learning applications can work on multiple investigations, thus making searching for crime patterns more efficient and effective and as such providing better support and information for more complex police work. However, it is important to emphasize once more the limitations of such systems, which should always be seen as tools to support human decision-making and not tools for autonomous decision-making.

PRACTICAL EXAMPLE **Recommender systems**

A recommender system could be used in a police supplier setting to buy equipment. On an e-commerce website that sells police equipment, officers may encounter a section with headings such as "*similar products*" or "*others also looked for...*". These recommendations are not selected and hard coded by developers but are compiled by an unsupervised model. This machine learning model analyses purchasing data from several clients along with the consumer's browsing history on the website to identify and suggest similar products that the consumer may be likely to purchase.

How does this work? For instance, consumer A buys a flashlight and a packet of 10 batteries. Consumer B buys a flashlight, a walkie talkie and two packets of 20 batteries. Consumer C buys a cell phone charger, two flashlights, a power bank, and four packets of 4 batteries. After analysing millions of examples of shopping carts such as these, the model learns that consumers who buy a flashlight tend to also buy batteries. There were, of course, other products in the shopping carts, but the frequency of their appearance is much lower than the combination of *flashlights* and *batteries*. As a result, the model recommends *batteries* to future purchasers of flashlights. Recommender systems like this could improve the quality of the acquired products, improve buying power while reducing storage costs and delivering financial savings.

DEEP LEARNING

As seen before, deep learning is a subfield of machine learning. It deals with a smaller family of algorithms known as neural networks. Inspired by the biological networks of neurons that constitute animal brains, these algorithms are composed of several layers of artificial neurons that progressively extract higher-level features from the raw input data. The adjective "deep" in deep learning refers to the use of multiple layers in the network.

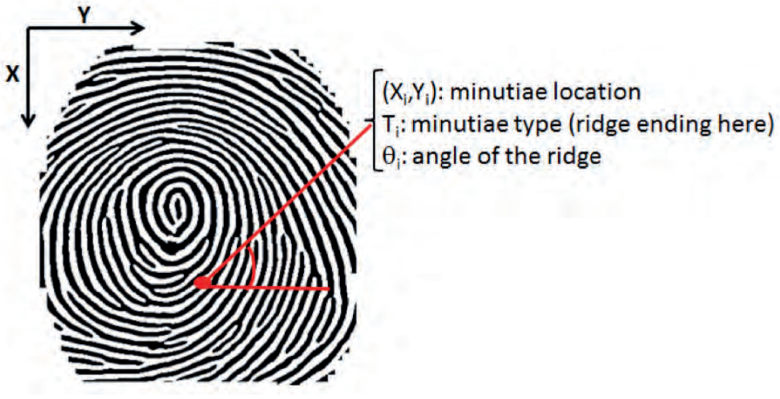
A key aspect of deep learning algorithms is that their performance scales with the amount of training data. While less advanced machine learning methods plateau at a certain level of performance when more training data is added, deep learning networks often continue to improve as the size of the data set increases.

Advances in deep learning have driven a great deal of progress and research in recent years in terms of image and video processing, text analysis, and speech recognition, boosting the development of the field of AI as a whole. In fact, recent deep learning models can achieve extremely high levels of accuracy, sometimes exceeding human performance.[7]

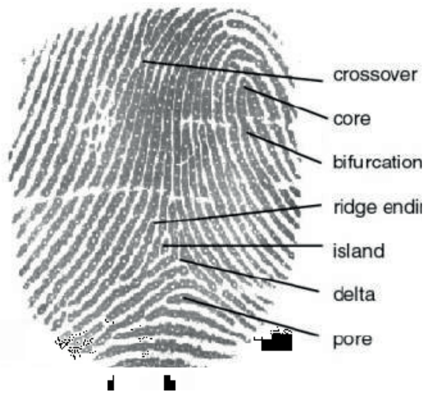
PRACTICAL EXAMPLE

Deep learning in fingerprint matching

Fingerprint matching seeks to identify or verify a person's identity based on patterns in the epidermal ridges on an individual's fingers. Although there are many different algorithms that can perform this task, most are based on the same features: minutiae coordinates, angle and type.


Y
X
(X_i,Y_i): minutiae location
T_i: minutiae type (ridge ending here)
θ_i: angle of the ridge

A



B

Figure 3-A. Features for fingerprint recognition include minutiae coordinates, angle of the ridge and type of minutiae such as ridge ending, bifurcation, core, delta, island and crossover. Copyright (2015) by B. Vibert and others. Reprinted with permission [8]

Figure 3-B. Image on the right depicts the different types of minutiae as well as sweat pores that allow for a higher accuracy on fingerprint matching. By Chander Kant & Rajender Nath. CC BY-NC 4.0.[9]

Fingerprint matching is a challenging task due to image noise, distortions, rotations and displacement which can result into large variability in impressions of the same finger and similarities between impressions of different fingers.

Traditional systems have proved to be effective in fingerprint matching and have been widely used by law enforcement agencies for a number of years. However, while these algorithms perform well on rolled and plain fingerprints, they often fail to match partial and latent fingerprints, i.e., fingerprints that have been unintentionally left on a surface.[10]

In recent years, deep learning techniques have helped to overcome some of these limitations. By enhancing latent fingerprint images, it is possible to improve fingerprint matching. In fact, the most recent deep learning models, with the support of high-resolution scanners, can even recognize sweat pores, improving fingerprint recognition even further.

The Four Components of AI Systems

Having clarified some of the key terminology, we will now seek to improve our understanding of what AI is. We will do so by breaking **AI systems** down into their most basic components. From a technological perspective, all AI systems are made up of two elements: **algorithms** and a **computer**. There are several types of AI algorithms, but machine learning algorithms are the most prominent and the most commonly used within the AI field, so these will be discussed in detail. Machine learning algorithms learn from **data** which means most AI systems today are a combination of algorithms, a computer and data.

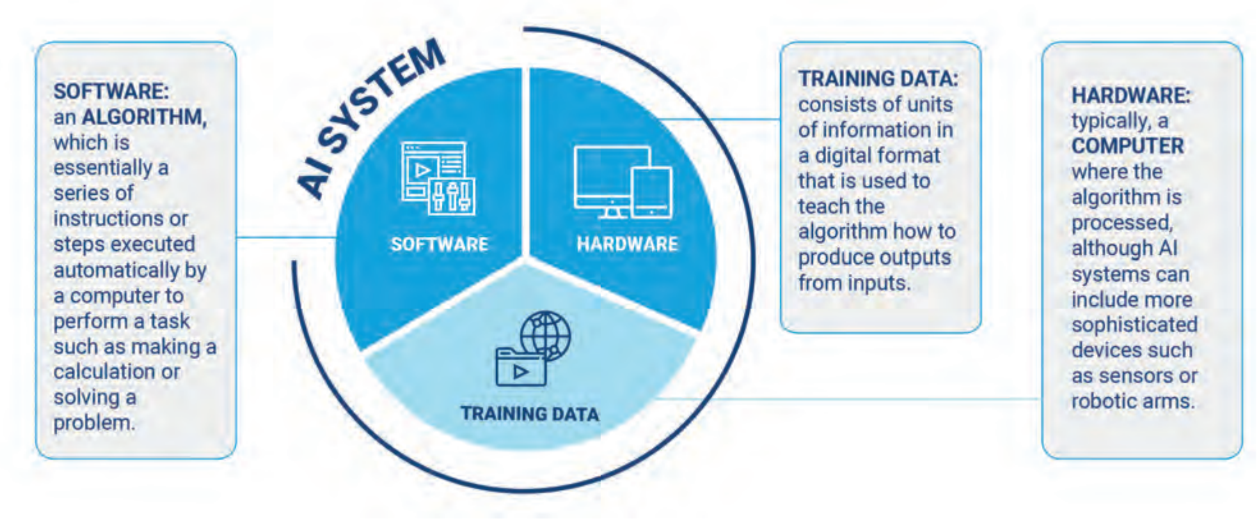

AI SYSTEM
SOFTWARE
HARDWARE
TRAINING DATA
SOFTWARE: an ALGORITHM, which is essentially a series of instructions or steps executed automatically by a computer to perform a task such as making a calculation or solving a problem.
TRAINING DATA: consists of units of information in a digital format that is used to teach the algorithm how to produce outputs from inputs.
HARDWARE: typically, a COMPUTER where the algorithm is processed, although AI systems can include more sophisticated devices such as sensors or robotic arms.

Figure 4. The components of an AI system.

However, there is a fourth element, and one which plays an integral role in ensuring the responsible use of AI innovation: the **human being**. Without people, the potential of AI could never be unlocked in law enforcement – or in any other domain for that matter.

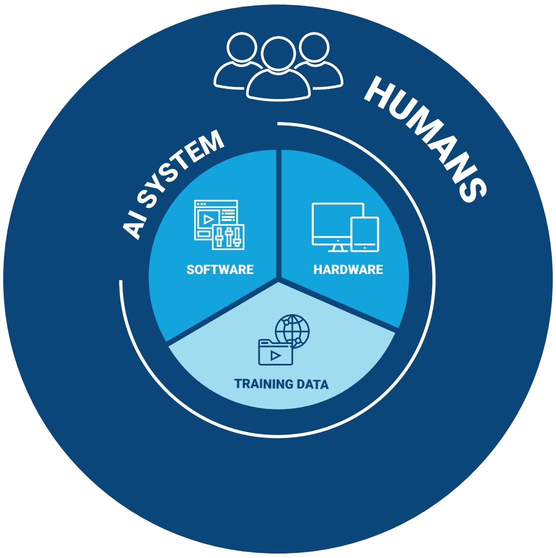


Figure 5. The role of the human as a fourth element of AI systems.

THE ALGORITHMS

An algorithm is a series of instructions to perform a calculation or solve a problem that is executed automatically by a computer. Algorithms form the basis of everything a computer does and are consequently also a fundamental aspect of all AI systems. An algorithm that has been trained on data for a specific problem is usually called a **model**, as it already models reality with a specific mathematical function.

Algorithms can be written in several **programming languages**. In AI research and development, the most common of these languages are Python, Lisp, C++, Java, and R. Several private entities working in the AI domain have also developed machine learning frameworks and libraries that support and facilitate the implementation of learning algorithms, for instance TensorFlow, Keras, NumPy, PyTorch, Scikit Learn, Pandas, Spark, and Apache MXNet.

In a typical learning problem, the algorithm is usually fed with a set of *variables* or *features,* that might be called **input** and returns a set of **outputs**, also called *responses, outcomes, labels* or *predictions.* These terms will be used interchangeably. In the statistical literature, the respective terms *independent variables* or *x* and *dependent variables* or *y* are also used instead of input and output.

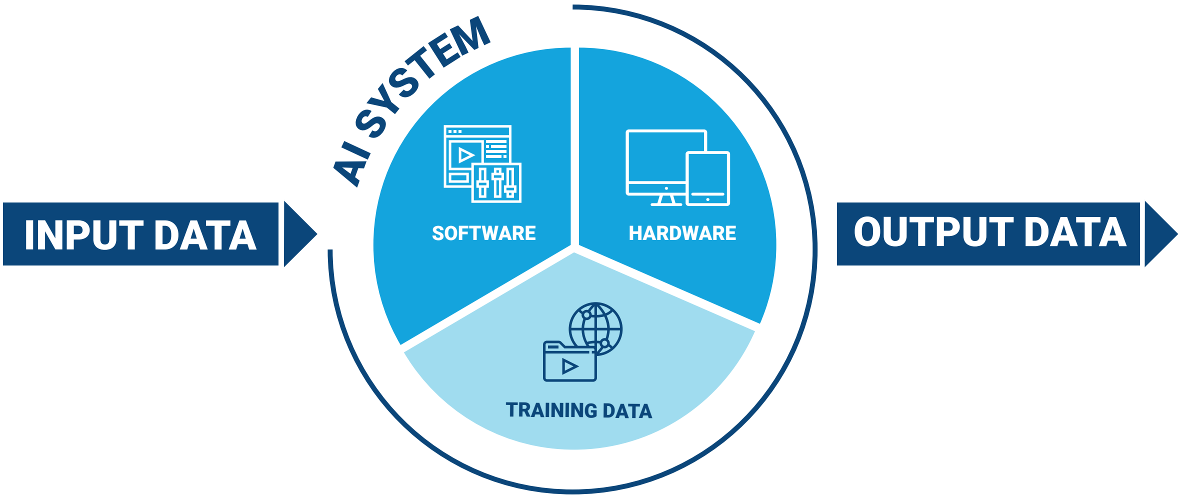


Figure 6. How an AI system works – input and output terms.

Classification, recognition, generation (of content), recommendation, etc., are goals that describe the types of tasks that an AI system can perform in the place of a human. In order to carry out these tasks, the AI system needs to “learn”. This learning depends on the type of data available and the nature of the task, and can be divided accordingly into at least three different learning methods. For each task, data, and learning method, different types of algorithms can be implemented, some with better performance than others. It is important to understand all of these levels of possibilities in order to use the best AI model for a specific problem.

TYPES OF LEARNING TASKS

Algorithms are powerful tools for analysing data and deriving new insights in the form of *outputs* from data points or *inputs*.

The **inputs** in question will vary in nature according to the problem and the type of measurement. They may be **quantitative** measurements such as age or height, where some measurements are bigger than others, and measurements which are close in value are close in nature; or they may be **qualitative** measurements where values belong to a finite set of options such as gender or colour, where there is no specific order. These differences in the input nature have led to distinctions in the types of methods used: some methods are naturally better for quantitative inputs, some for qualitative, and some for both.

The **outputs** can also vary, and we can try to predict **quantitative** or **qualitative** measurements. These different output types have led to a naming convention for prediction tasks: the term *regression* is used when the goal is to predict quantitative outputs, and *classification* when the goal is to predict qualitative outputs. As we will see, these two tasks have a lot in common, and regression can be also used for classification.

|► *Learn more about quantitative and qualitative data in The Data section.*

Classification

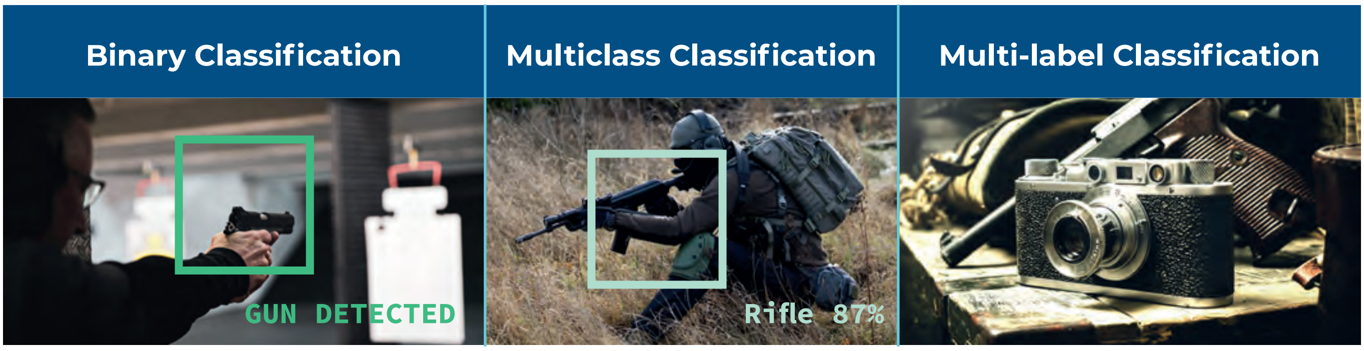
Classification is the most common task and consists of assigning a class label to a certain example. In other words, a classification algorithm is trained to identify which category an object belongs to from a finite set of values or classes. Algorithms used for classification are also known as **classifiers**.

There are many different types of classification tasks:

- **Binary classification** refers to predicting one of two classes: yes/no or 0/1. For example, to detect if there are guns in an image the system would classify it as “positive” when guns are present or “negative” if they are not.

- **Multi-class classification** involves predicting one of more than two classes. For instance, in the gun detection example, the system aims to identify which kind of gun is depicted in the image, by predicting one class among several classes.

- **Multi-label classification** involves predicting one or more classes for each example, for example when identifying numerous objects in one picture.



Binary Classification

- Positive
- Negative

Multiclass Classification

- Handgun
- Automatic
- Rifle
- Revolver
- Carbine

Multi-label Classification

- Camera;
- Gun
- Table
- Vintage
- ...

Figure 7. Different types of classification tasks.

Regression

When the goal is to predict a continuous value or a real number, i.e., any number within a range, it is conventionally called a *regression* task. This can be achieved using a set of statistical and machine learning processes for estimating the relationship, or correlation, between *variables*: a numerical *dependent variable*, the *label*, and one or more *independent variables*, the *features*. The most common form of regression analysis is linear regression, in which the goal is to find the line (or a more complex linear combination) that most closely fits the data according to a specific mathematical criterion. Other techniques such as principal component analysis or linear discriminant analysis can also produce a continuous output.

PRACTICAL EXAMPLE **Analysing fear of crime**

In the example below, regression analysis is used to correlate *fear of crime* with *income inequality* in 29 countries.[11] We start by plotting in a graph all the data points with *income inequality* index as a feature or independent variable, and *fear of crime* as a label or dependent variable.

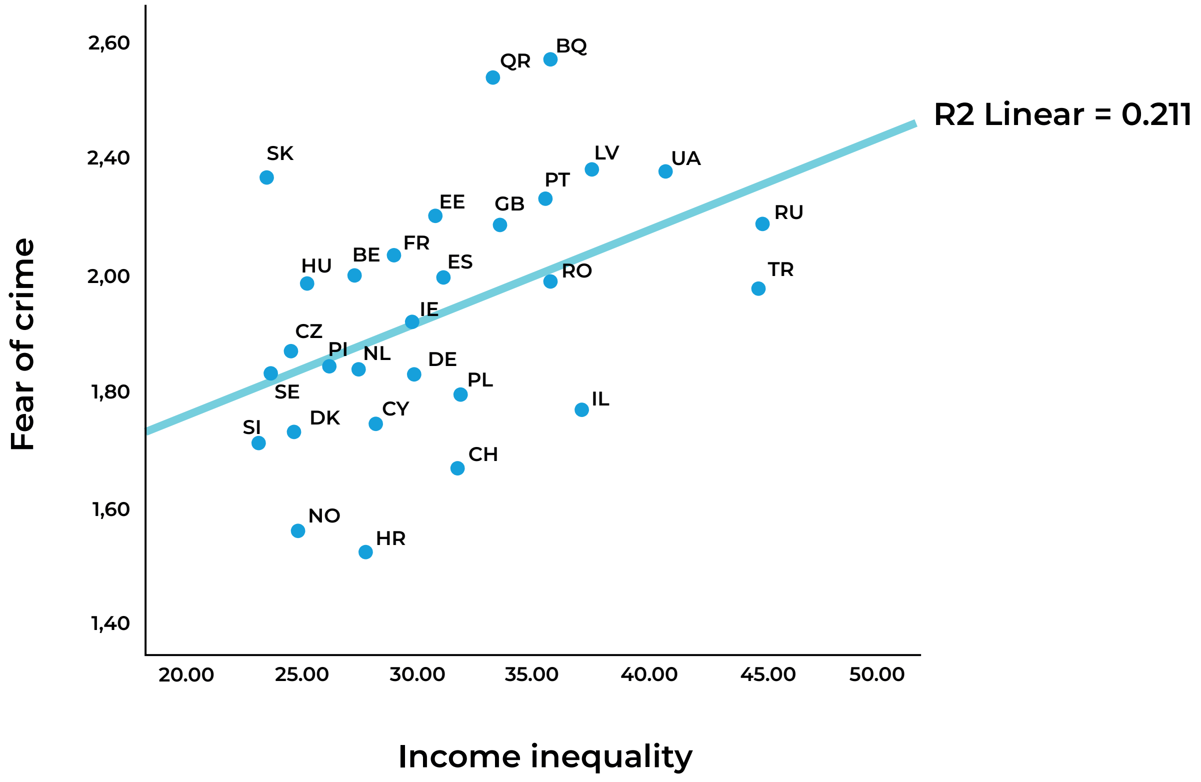


Figure 8. Regression analysis correlating income inequality with fear of crime in 29 world countries. By Christin-Melanie Vauclair and Boyka Bratanova. CC BY 3.0.[12]

The graph shows that these two variables are positively correlated, meaning that when *inequality* increases, the *fear of crime* also increases. However, when looking at the distribution of the points, we can see that the data points do not follow the regression line exactly, which means that the correlation is not very strong. Even if the correlation were strong, we cannot assume that *income inequality* is the cause of *fear of crime*. Other factors would have to be investigated to understand whether one is directly affecting the other or if the two are independent (see section *Correlation vs Causation*).

Based on the regression line obtained, we can estimate the fear of crime when the income inequality index is exactly 44.00 for example. Regression models help us to make predictions and speculate about possible scenarios, but we must be careful to not jump to false conclusions.

Using Regression for Classification - Classification Thresholds

While regression is used to predict continuous values, it can also be used for a classification task by setting a classification threshold (also called a decision threshold) for the predicted values. For example, binary classification is often used to sort predicted values into either *positive* or *negative* classes, for example, *true* or *false*, *pass* or *fail*, etc.

A regression technique (such as logistic regression) can also be used to estimate the probability of a predicted value belonging to the positive class. The easiest method for converting a regression model into a classification model would be to set a threshold of 50%, below which values are classified as negative and above which values are classified as positive. In fact, not all examples are that easy to define. Classification thresholds are problem-dependent, and these values must be adapted to the situation.

Imagine an email priority classifier that flags emails containing important information such as INTERPOL Red notices.[13] Based on keywords from the email text, a classifier computes the probability that the email contains important information. Any value above a certain threshold indicates *priority* (positive class); a value below the threshold indicates *not priority* (negative class). One aspect of setting a threshold is assessing how much a mistake will cost. For example, mistakenly labelling a non-priority message as *priority* (false positive) is inconvenient, but not problematic. However, mistakenly labelling a priority message as *non-priority* (false negative) is extremely serious and could lead to failing to arrest a suspect. In these cases, the straightforward option would be to lower the threshold so that there are fewer false negatives.

|▶ *See the Evaluation of classification section to understand the difference between false positives and false negatives.*

PRACTICAL EXAMPLE

Risk matrix – classification using regression algorithm

As explained above, classification can be implemented through regression with a **classification threshold** applied. One example is a risk matrix, an AI system used to identify and risk-assess subjects' predisposition to using violence. The risk matrix measures the potential risk subjects pose by scoring them based on evidence that they have committed violent acts in the past, any previous convictions for weapons offences, plus any police intelligence indicating that they may have access to weapons, or that they have been involved in violent incidents. A threshold of 0.7 is then established to convert the continuous score into a classification which aims to predict one of two classes: positive or negative evidence of a predisposition to using violence.

Regression (violence score)	0 – 0.70	0.71 – 1
Classification (predisposition to using violence)	Negative	Positive

Figure 9. Using regression for a classification task. While for regression the goal is to predict the violence risk score, for classification the aim is to predict one of the two classes, positive or negative evidence of a predisposition to using violence.

It should be noted that this example and its application can produce **bias** and **discriminatory outputs** and should be considered with due care.

|▶ *Learn more about the impact of bias in the Introduction to Responsible AI Innovation*

Other types of learning problems

Although regression and classification are the most common tasks, there are other types of learning problems, including:

- *Clustering:* the task of grouping a set of data points so that elements in the same group, or cluster, are more similar to each other than to those in other groups.

- *Ranking*: consists of ordering objects according to relevance, for example ranking web pages in response to user queries.

- *Sequence labelling*: when the input is a sequence of elements, and the output is a corresponding sequence of labels. For example, if we want to label words in a sentence with their syntactic category, the input will be a sequence of words such as "*this flower is beautiful*", and the output will be the sequence "*determiner noun verb adjective*".

- *Content generation*: when the output consists of creating a new example based on the training database.

LEARNING METHODS

In contrast with classic algorithms, machine learning algorithms do not require explicit instructions to perform a specific task, but rather extract patterns and learn implicit rules from a large number of examples. There are three ways of "learning", i.e., extracting information from the data in order to predict future events.

Supervised Learning

Supervised learning is the most common learning mechanism and involves training a model using human-labelled data. In this method, the algorithm uses a number of examples of input-output pairs to extract implicit rules, i.e., a mathematical model that can transform an input into a certain output. Once it has learned these rules, the model can be used to predict an output for an input that it has never seen before. In other words, supervised learning consists of learning how to map an input to an output label, based on a number of examples of input-output pairs.

Imagine a group of developers who start by creating a simple image processing system that is able to recognize police badges. During training, several input-output pairs are fed into the system (for instance, star-shaped badge = Las Vegas; circle-shaped badge = Stockton, etc.). In the beginning, the system gives a random probability of being each label. If the label with the highest probability is not the correct one (as shown in the image), the model weights are adjusted to give a correct output. And every time the system gives the highest probability to the correct label, nothing is changed in the model.

After ingesting several different examples, the learning model is able to predict the classification of unlabelled data, as long as it correlates to the training data set (for instance input: star-shaped badge? – Las Vegas).

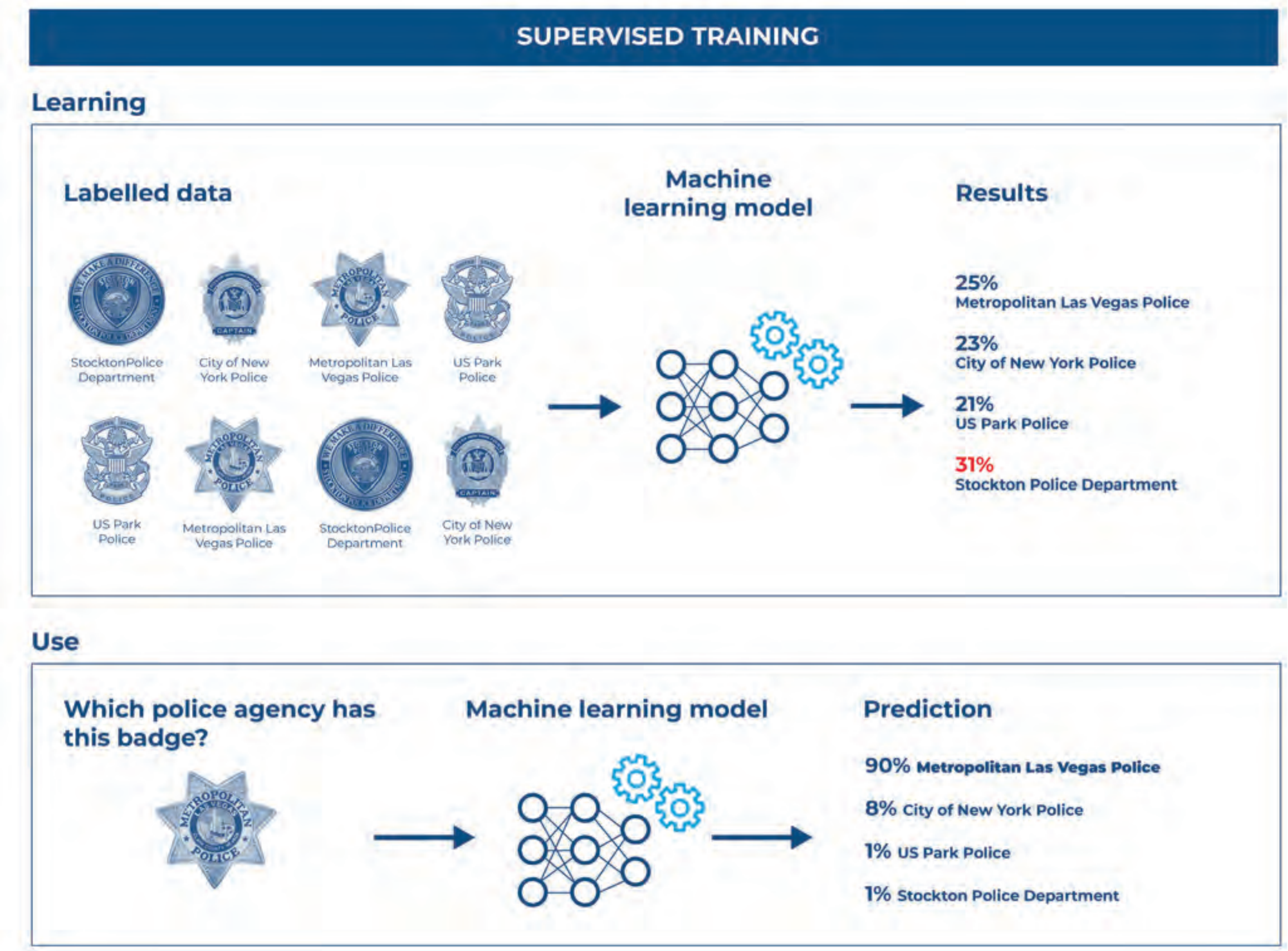

SUPERVISED TRAINING
Learning
Labelled data
StocktonPolice Department
City of New York Police
Metropolitan Las Vegas Police
US Park Police
US Park Police
Metropolitan Las Vegas Police
StocktonPolice Department
City of New York Police
Machine learning model
Results
25% Metropolitan Las Vegas Police
23% City of New York Police
21% US Park Police
31% Stockton Police Department
Use
Which police agency has this badge?
Machine learning model
Prediction
90% Metropolitan Las Vegas Police
8% City of New York Police
1% US Park Police
1% Stockton Police Department

Figure 10. Supervised Learning method: After training with labelled training data, the model extracts information about the characteristics of the data that allow it to classify each example. Once unlabelled testing data is inserted into the model, it can analyse this input and will output a label for that example. By NYPD Captain's Association. CC BY-SA 4.0. By US Park Police Badge (USPP). CC BY-SA 4.0.

Unsupervised Learning

Unlike supervised learning, with unsupervised learning the algorithm has to find patterns without having the corresponding labels. In this case, the goal is to find patterns in data that have not been labelled, classified or categorized. One example of its use is for clustering, which consists of grouping data elements with similar features.

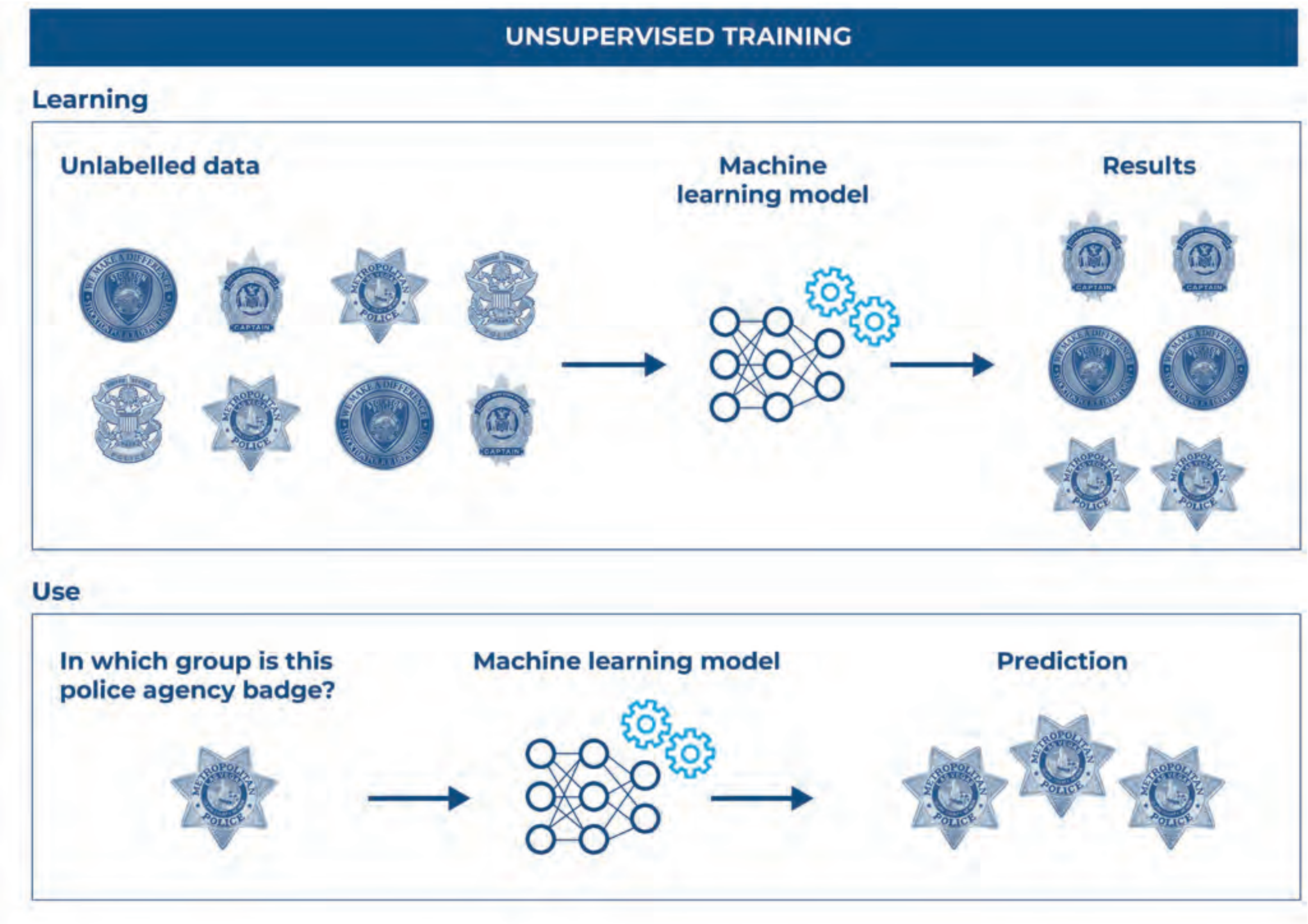


Figure 11. Unsupervised Learning method: After training the model with unlabelled training data, the model extracts information about the characteristics of the data that allow it to group examples based on similar characteristics. Once another unlabelled testing data point is inserted into the model, it can analyse this input and will output a classification indicating which group this example fits best into. By NYPD Captain's Association. CC BY-SA 4.0. By US Park Police Badge (USPP). CC BY-SA 4.0.

PRACTICAL EXAMPLE

Child sexual abuse and exploitation investigations

Clustering is used in the field of child sexual exploitation investigations where **image processing** systems can group pictures of the same child at different ages from materials collected from seized devices. This allows investigators to determine the duration of the abuse to support the prosecution case against the perpetrator.

Allowing the algorithm to identify its own patterns, rather than training it on human-labelled data, highlights details that developers might not have thought to look for. For example, also in child abuse investigations, a **case management** AI system that connects pieces of evidence between several investigations could help locate an offender. By recognising that a room in the background of a picture from a child sexual abuse investigation is the same as that in another picture from a case of human trafficking, the system allows investigators to connect the two cases and gather the right evidence to prosecute the suspect.

Reinforcement learning

The third learning method is reinforcement learning, whereby the model does not learn through data pairs of input-output examples, but through trial and error in order to achieve a specific goal. This involves the machine's interaction with an environment in order to maximize its performance in a specific task. Robotic applications typically use this type of learning method.

WANT TO LEARN MORE?

How does Reinforcement Learning work?

Controlling traffic and rush hour to ensure safety is not easy, and reinforcement learning can help with this task. Monitoring traffic continuously in complex urban networks allows a map of traffic patterns and vehicle behaviour to be built, with information such as the times when traffic is heaviest, the directions it is coming from, and how quickly cars move through each light colour. Traffic signals can be controlled by reinforcement learning models which analyse these patterns and adapt accordingly depending on the time of day, climate, and season.

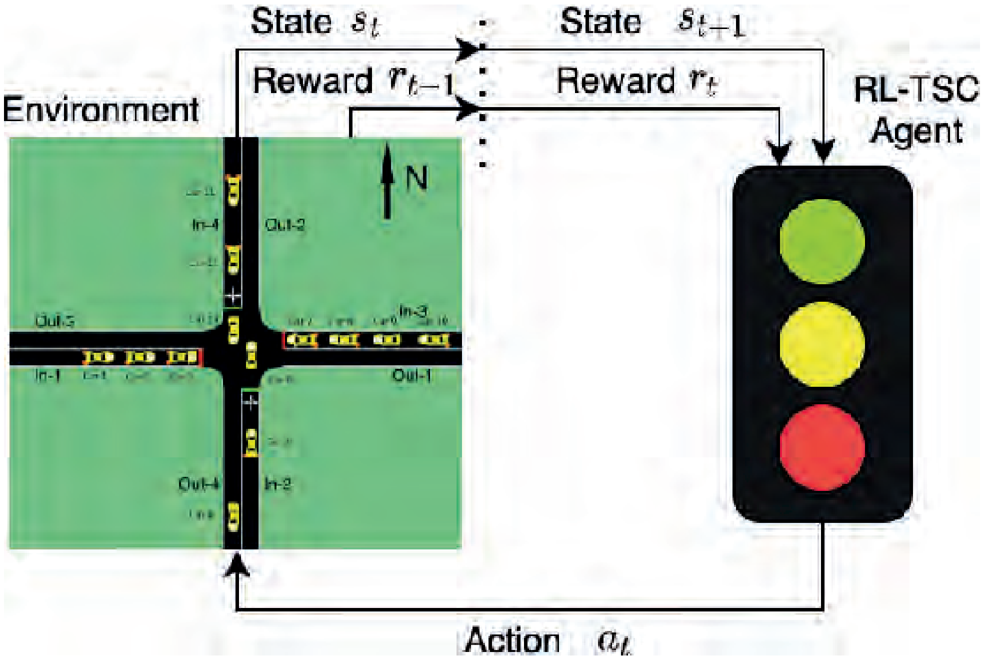


Figure 12. Reinforcement Learning method: During training, the model is given a goal to achieve and tries all possibilities to achieve this goal. By Mobin Zhao, Wangzhi Li, Yongjie Fu, Kangrui Ruan and Xuan Di. CC BY 4.0. [14]

A typical reinforcement learning scenario is composed of three elements: an **agent** (the traffic lights system with reinforcement learning) takes *actions* in an **environment** (an intersection). These actions will influence the **state** of the environment (the traffic), which is fed back into the agent together with a **reward** that can be positive or negative according to the effect the action had on the *state* or traffic.

TYPES OF ALGORITHMS

The three previous learning methods allow the model to "learn" or to extract patterns from data, but there are several algorithms that can then make use of these methods. Below, there is an explanation of some of the most common algorithms that can be used for Machine Learning tasks. This field is continuously evolving, and therefore, the list below is not exhaustive and new types of algorithms can emerge and replace those below.

Neural Networks

Neural networks are a class of algorithms whose significance has been profound, to the extent that they have laid the groundwork for a distinct domain in machine learning known as deep learning. Inspired by the biology of nerve cells, neural networks consist of a set of *neurons* or *nodes* (i.e. computational units containing a number or a *weight*) that receive input, process this input through a computational function, and transmit it to the next neuron.

By stacking layers of neurons together, neural networks can progressively extract higher-level features from an input point to be able to perform complex tasks. For example, in object recognition tasks, lower layers may identify edges and corners, while higher layers may identify digits, letters or faces.

WANT TO LEARN MORE?

A more technical explanation of neural networks

The architecture or algorithm structure of neural networks is inspired by the brain cells' morphology. A typical neural network may consist of thousands or even millions of simple processing **nodes** (the coloured circles in figure below) that are densely interconnected and organized in **layers** (vertical columns of coloured circles). These nodes include numbers called **weights** that are multiplied in every iteration by numbers along the outgoing **edges** (black lines) to which they are connected. The multiplicative values obtained from each incoming edge at a node are then added together to assign a value to the node. The typical neural network has an **input layer** (blue circles), one or many **hidden layers** (green circles), and an **output layer** (red circle).

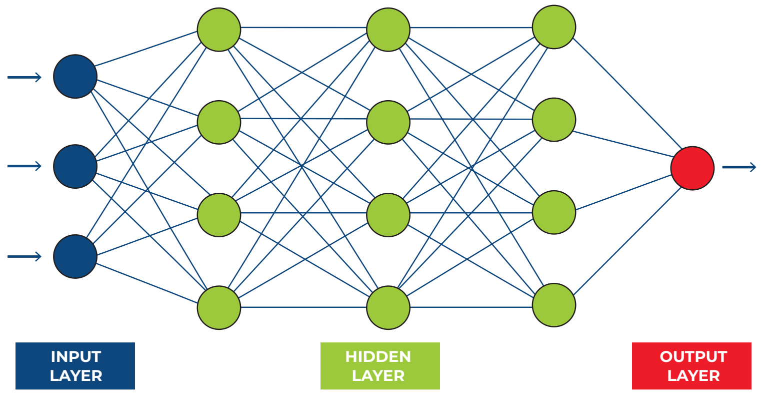


Figure 13. The basic structure of a neural network. This example has an input layer with 3 nodes, 3 hidden layers with 4 nodes each and an output layer with a single node.

The number of **layers** varies with the complexity of the model, the type of data and the intended performance level.

The number of **nodes per layer** also varies from model to model and according to the type of input (if it is an image, it will have many more input nodes than if it is a word) and the type of output (a yes/no result requires a single output node, but a gun recognition model will require one output node per type of gun to be identified).

The first layer of neurons receives and processes the data, and the subsequent layers continue to process the data received from the previous layer. The network is trained by adjusting the weights based on the training predictions. For example, if the network classifies an image correctly, the weights contributing to the correct answer are increased, while other weights are decreased. This is called the **back-propagation** learning method.

Within the family of neural networks there are different algorithm architectures that are specialized in different tasks. For example, while convolutional neural networks are widely used for image processing, recurrent neural networks and transformers are better suited for text processing.

The immense and complex architecture of neural networks, sometimes including as many as 150 layers and millions of neurons, means that they are not easy to interpret. In fact, it is hard for developers to understand how the output has been reached for a given input. This is why neural networks are often described as **black boxes**: their results are not immediately explainable, and this may limit their use in the law enforcement context. There are, however, new methods and tools which can be used to explain the outputs of neural networks.

|▶ *Learn more about the black box problem and explainability tools in the section Understanding AI systems.*

Decision Trees and Random Forests

Decision trees are a supervised learning model used for classification and regression, which means they can predict a category or a real number. These algorithms use a branching structure that allows to illustrate the results of a decision.

Decision trees can be used to map the possible outcomes of a decision, depicting clearly the reason for a specific outcome, which is useful for transparency and explainability purposes.

|▶ *Learn more about the importance of transparency and explainability of AI systems in the* ***Introduction to Responsible AI innovation*** *and in the* ***Principles for Responsible AI Innovation.***

As depicted in the image below, each end node of a decision tree represents a possible outcome. Percentages are assigned to nodes based on the likelihood of that outcome occurring.

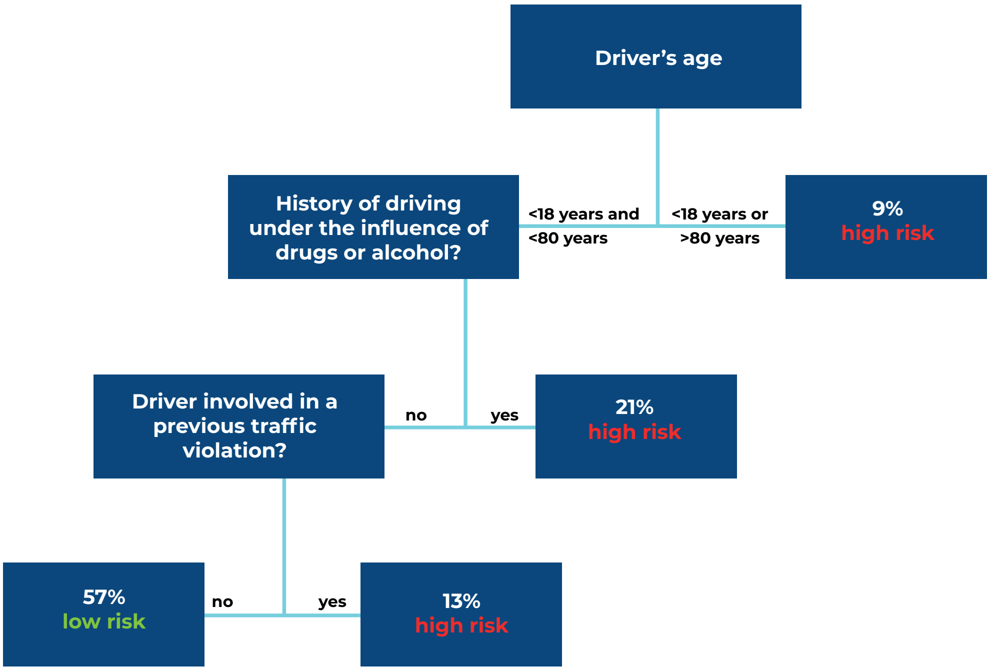


Figure 14. Decision tree of the influence of offender characteristics on the decision to confess or not during police interrogation.

Random forest is a variation of a decision tree which combines the output of multiple decision trees to reach a single result. It can also be used to solve both classification and regression problems. The initial data set is divided into smaller data samples and each one is analysed by a decision tree. The final result is the majority of the results obtained by all the decision trees combined. Random forests consist of multiple decision trees which form an ensemble. This architecture may produce more accurate results, although the method is more time-consuming, and the outputs are less easy to interpret than single decision tree.

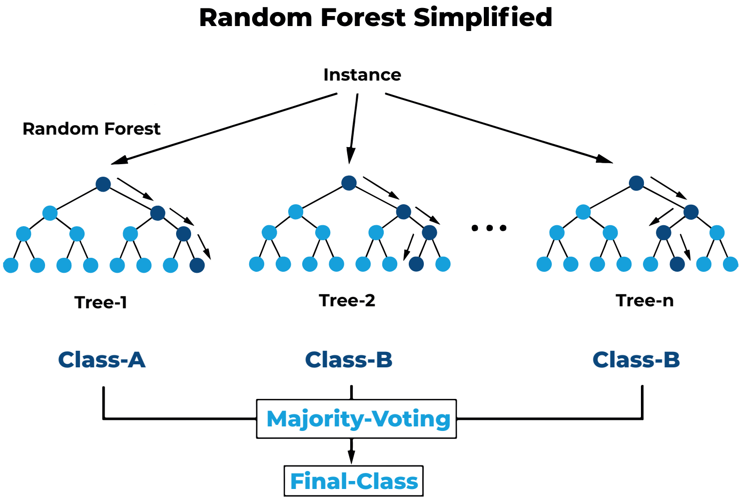


Figure 15. Example of a standard random forest.

Linear Regression

Linear regression is a commonly used supervised algorithm that predicts continuous values. This algorithm is used to find the best fitting line, plane (or a more complex shape) that describes two or more variables.

Linear regression algorithms can quantify the strength of a correlation between variables in a data set and can be used for predicting future data values based on historical values. In the example below we can correlate the size of several cities in kilometers square with the amount of requests received per year by the police. As shown, there is a positive correlation between the two variables and the linear regression model allows to estimate the number of requests for the police in a city with an area of 600 km^2, which would be approximately 30000 per year.

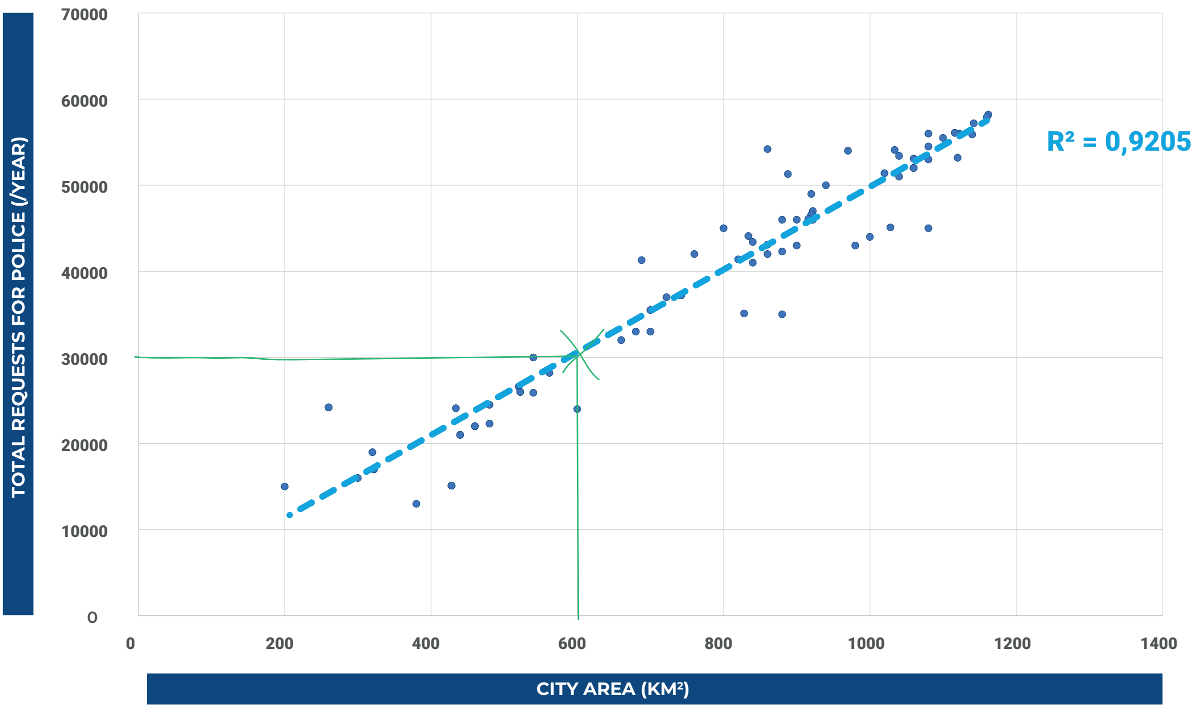


Figure 16. A linear regression model correlating city area in kilometers square with number of requests for the police per year. These models can be used to predict or estimate new values based on a new input (green cross).

However, it is essential to remember that correlation does not necessarily mean causation and it is important to understand the context around the data before inferring a causal relationship between variables.

Logistic Regression

While linear regression is used for regression - predicting a real number using a given set of data features - logistic regression is used for classification - predicting the probability of belonging to a certain class. The figure below depicts a logistic regression model for binary classification. The algorithm estimates the probability of an event occurring (1) or not (0) by fitting an S-shaped curve to the data.

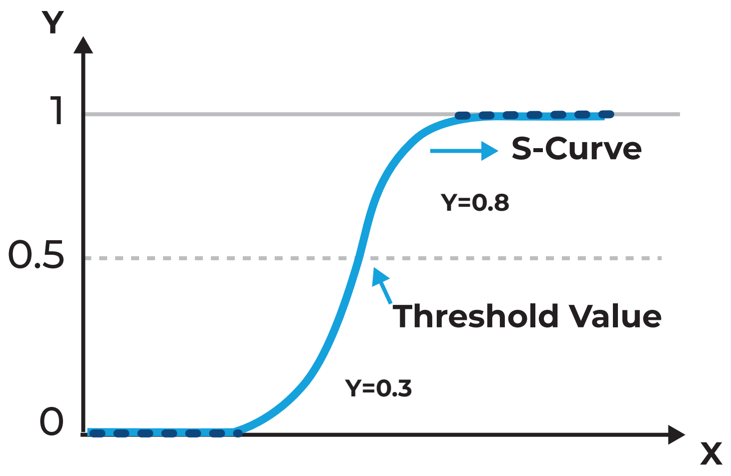


Figure 17. Example of a logistic regression standard model.

Naïve Bayes

Naïve Bayes is another technique used for classification which allows the probability of a hypothesis (for instance, a class label in a classification task) to be computed based on certain observed features. It is not a single algorithm, but a family of algorithms based on the common “naïve” assumption that the features are independent of each other, which is usually not the case.

For example, a Naïve Bayes algorithm could be used to identify potentially fraudulent financial transactions. In this case, the model would be trained on a data set of labelled transactions, where each transaction is labelled as either *legitimate* or *fraudulent*. The algorithm would use the features associated with each transaction, such as the *transaction amount*, the *merchant location*, the *transaction type*, and so on, to learn a probability distribution of the features associated with legitimate and fraudulent transactions.

K-Nearest Neighbour

Also known as KNN, K-Nearest Neighbour is used for regression or classification, and it can learn in a supervised or unsupervised way. Looking at the image below, we can see that given a new data point (green dot), the algorithm returns the closest (or two closest or k closest) neighbouring points. In a supervised learning task, the points are labelled, and we can, therefore, assign a value to the input data point. In this example, the green dot should be classified either as a blue square or a red triangle. If k = 3 (solid line circle) it is assigned to the red triangles category because there are 2 triangles and only 1 square inside the inner circle. If k = 5 (dashed line circle) it is assigned to the blue squares category (3 squares vs. 2 triangles inside the outer circle).

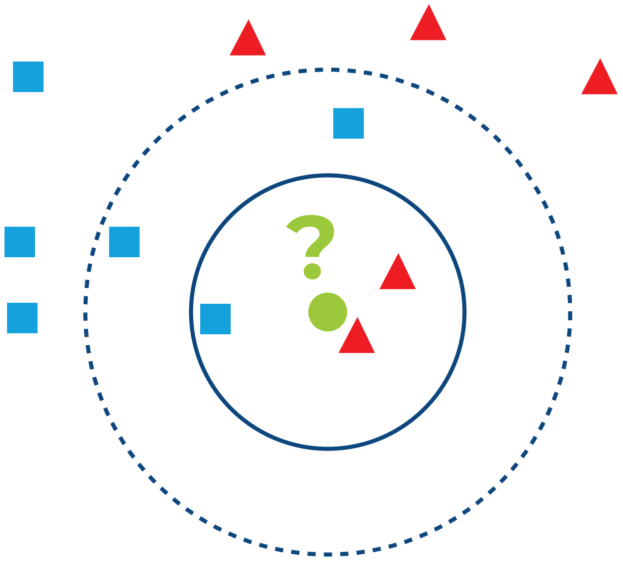


Figure 18. Example of k-NN classification.[15]

KNN is the simplest of all machine learning classifiers. It differs from other machine learning techniques in that it does not produce a model, but it simply stores all available cases and classifies new instances based on an instant measure of similarity.

MODEL PERFORMANCE

To know if a model is performing as expected, its behaviour must be assessed. This is important from both a technical functionality perspective and to ensure responsible use of AI systems. Evaluating performance includes setting and measuring against specific metrics such as *accuracy*, *precision* and *recall*, and analysing for potentially harmful behaviours, such as *feedback* *loops*. There are online tools that can help evaluate models' behaviour such as *What-If Tool*[16], and others that detect errors such as *Error analysis*[17].

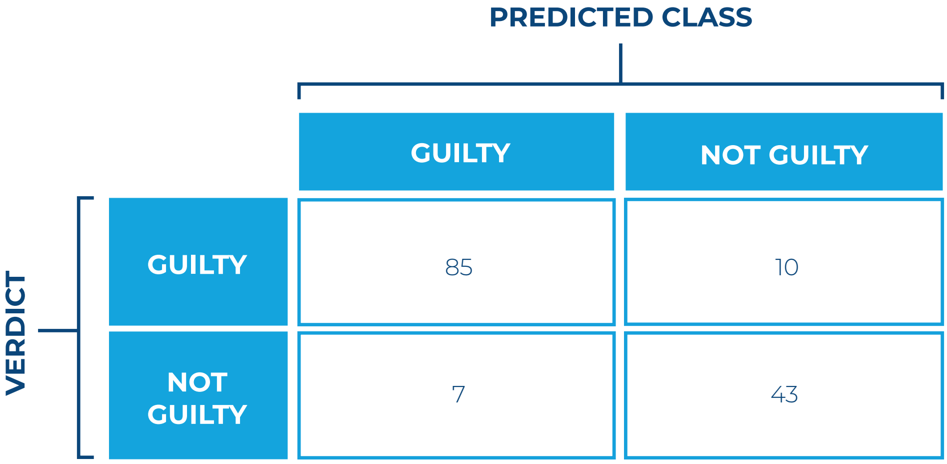
Evaluation of regression

To establish whether a model is performing the proposed task well, we need to evaluate its performance. In regression tasks, the goal is to predict a numeric value such as the number of robberies that will occur next summer. An evaluation of the model's performance is not based on whether the model predicted the value exactly or not. Instead, it is based on how close the predictions were to the actual, measured values. For that we can use the term **error**, which summarizes on average how close predictions were to their actual expected values. There are three error metrics which are commonly used for evaluating and reporting the performance of a regression model:

- Mean squared error (MSE)
- Root mean squared error (RMSE)
- Mean absolute error (MAE).

Evaluation of classification

In classification tasks, the goal is to establish whether the predicted class or label is correct or not, and how many times the model predicted the right class. For simplicity's sake, we will mostly discuss a binary classification problem where the goal is to predict one from two classes, *positive* or *negative*. Imagine a classifier that estimates if the person is *guilty* (positive class) or *not guilty* (negative class) based on the evidence. This is, of course, a simplistic approach used for the sake of explaining the consequences of different types of errors and should not be considered as a viable tool to be used in the criminal justice setting.

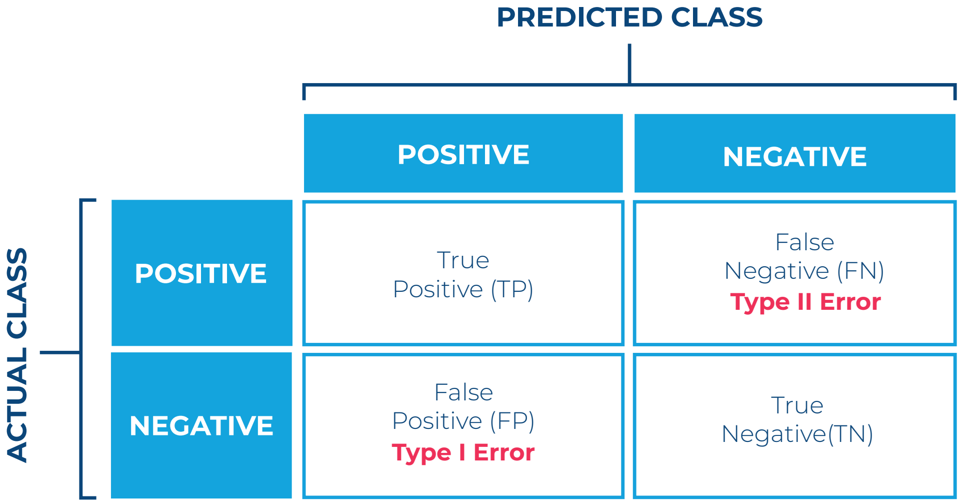


		PREDICTED CLASS	
		GUILTY	NOT GUILTY
VERDICT	GUILTY	85	10
	NOT GUILTY	7	43

Figure 19. Model performance.

The table above shows that 85 individuals were correctly classified as *guilty* and 43 innocent individuals were correctly classified as *not guilty*. However, we can see that the model is not 100% correct, since it classified 10 criminal individuals as *not guilty* and 7 innocent individuals as *guilty*. How can we analyse this performance?

These two types of errors have a completely different impact, so they must be analysed differently. In the first case, we have individuals who may be wrongly considered innocent, which ultimately could lead to failing to prosecute, convict or arrest them. In the second case, we have innocent individuals who are wrongly considered guilty, which ultimately could lead to wrongfully prosecuting, convicting or arresting them. This clearly demonstrates the need for responsible AI innovation. The table below (called a confusion matrix) shows how these different cases are analysed:



		PREDICTED CLASS	
		POSITIVE	NEGATIVE
ACTUAL CLASS	POSITIVE	True Positive (TP)	False Negative (FN) **Type II Error**
	NEGATIVE	False Positive (FP) **Type I Error**	True Negative(TN)

Figure 20. Confusion matrix allows a model's performance to be assessed.

- **True positives** (TP): Predicted *positive* and are *positive*, i.e., the 85 criminal individuals who were classified as *guilty*.

- **False negatives** (FN): Predicted *negative* and are actually *positive*, i.e., the 10 criminal individuals who were classified as *not guilty*.

- **True negatives** (TN): Predicted *negative* and are *negative*, i.e., the 43 innocent individuals who were classified as *not guilty*.

- **False positives** (FP): Predicted *positive* and are actually *negative*, i.e., the 7 innocent individuals who were classified as *guilty*.

The most commonly used metric for judging a model is **accuracy** which focuses on the "good performance" of the model. This metric is defined as the percentage of correct predictions, which is calculated by dividing the number of correct predictions (the *true positives* and the *true negatives*) by the number of total predictions:

$$Accuracy = \frac{true\ conditions}{all\ cases} = \frac{TP + TN}{TP + FN +\ TN + FP} = \frac{85 + 43}{85 + 10 + 43 + 7} = 88\%$$

Unfortunately, accuracy can be a flawed metric, especially with unbalanced classes. For instance, if a data set of 1000 officers from a law enforcement agency contains 5 individuals who misuse agency funds, a classifier that predicts that everyone is honest would have an accuracy of 99.5% without ever finding a single bad employee. For these reasons, there are other evaluation metrics that allow for a better and more complete idea of the model performance including, precision, recall, and F1 score.

Precision shows how precise the model is, i.e., from the positive predictions, how many were actually *positive*. This is a good measure to use when the costs of *false positives* are high, which is the case when there is a risk of classifying innocent individuals as *guilty*.

$$Precision = \frac{true\ positive}{positive\ predictions} = \frac{TP}{TP + FP} = \frac{85}{85 + 7} = 92\%$$

The second metric is **recall**, which calculates how many of the actual *positives* the model detected and classified as *positive*. Recall should be used when there is a high cost associated with *false negatives*, which is the case when there is a risk of classifying criminal individuals as *not guilty*.

$$Recall = \frac{true\ positive}{actual\ positive} = \frac{TP}{TP + FN} = \frac{85}{85 + 10} = 89\%$$

F1 score is used when we seek a balance between precision and recall. This is particularly relevant in cases where there is an uneven class distribution (for example, a large number of negative data points).

$$F1\ score = 2\ \times \frac{precision \times recall}{precision + recall} = \frac{2 \times TP}{2 \times TP + FP + FN} = \frac{2 \times 85}{2 \times 85 + 7 + 10} = 0{,}91$$

Depending on the usage and the type of database, different metrics can better characterize the model's performance. It is important not to rely too much on accuracy, which is often quoted by suppliers, and to investigate all types of errors – in other words, to have a good understanding of the algorithm as a whole.

Feedback loops

After being trained, a machine learning model can make predictions about an event. If the actions based on those predictions are then fed into the model, the system enters into a feedback loop. A feedback loop is a process that repeats itself through multiple iterations, where the output of the process serves as input for the next iteration. Feedback loops can play a significant role in boosting machine learning models' performance because we have more and updated data to continuously train the model. However, they can also lead to some ethical and legal issues in certain AI systems, such as recommendation systems or decision engines.

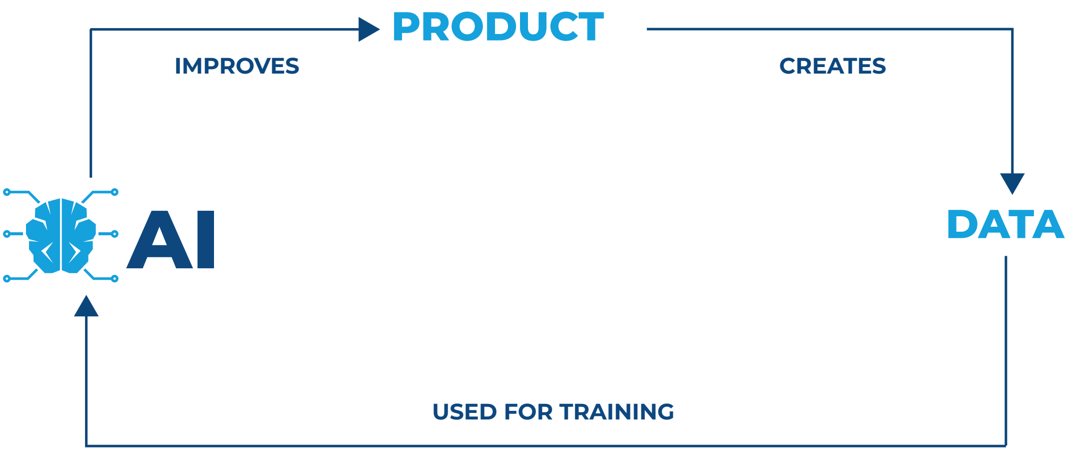


Figure 21. Example of a feedback loop where outputted data is fed back into the AI system to improve its performance recursively.

PRACTICAL EXAMPLE **Hotspot predictive policing**

One example of a harmful feedback loop is in **hotspot predictive policing** algorithms. These algorithms can be used to inform decisions about how to distribute law enforcement officers across neighbourhoods to allow for a more efficient use of limited resources.

If more officers are sent to one location, they tend to make more arrests there. If the number of arrests in that location is then fed back into the system – **feedback loop** – the system will conclude that the crime rate is very high in that neighbourhood, which will lead to even more officers being sent to that area. This means the software ends up overestimating the crime rate in one neighbourhood, without taking into account the possibility that more crime is observed there simply because more officers were patrolling in that area.

In addition to the problem described above, feeding the system with real-time data may not be recommended in a law enforcement context because this data may be sensitive and confidential. As such, this should be analysed on a case-by-case basis.

THE DATA

Data is the second of the four fundamental components of AI systems using machine learning algorithms. Data plays a central role because it is the raw material which machine learning models use to operate. In other words, machine learning models process large amounts of data in order to carry out functions such as recognising patterns in the data, patterns that can then be used to make predictions about new data points. To obtain accurate predictions, however, the quality and quantity of the data must be sufficient.

Data can be defined as units of information describing quantity, quality, facts, statistics or other basic units of meaning, collected for reference or analysis.[18] For the purposes of AI, data is a set of values of variables about persons, objects, or any other entity.[19] A data set, or database, is a collection of several units of data that are measured and reported, producing data visualizations such as graphs, tables, or images.

Generally speaking, data can be categorized into two types:

- **Quantitative data** refers to any information that can be quantified, counted or measured, and given a numerical value, such as age, height, income, etc.

- **Qualitative data** is descriptive in nature, expressed in terms of language rather than numerical values, such as gender, education level, name, etc.

Data can also be classified according to two main categories:

- **Structured data,** when it is stored in a predefined format such as a table or spreadsheet. Common formats for structured data include CSV, Excel and JSON. In this format, data is usually more accessible and can be used directly with minimal pre-processing.

- **Unstructured data**, when it is a conglomeration of many varied types of data that are stored in their native formats, such as in an email or a social media post. This data normally needs to be transformed into a structured format in order to be processed by a machine learning model. Unstructured data accounts for the large majority of all data produced.

DATA COLLECTION

Data collection is the process of gathering and measuring information using standard validated techniques. Depending on the required information and the field of study, the approach to data collection is different.

To reduce the likelihood of errors and ensure that the data gathered is both defined and accurate, a formal data collection process including appropriate data collection instruments and delineated collection standards is necessary.

PRACTICAL EXAMPLE — Data collection requirements for facial recognition

In the field of facial recognition, law enforcement agencies should establish standards and thresholds for **image quality** when collecting images for reference databases in order to mitigate the risk of errors. In the law enforcement context, it is common to use a reference database of known suspects, composed of photos and mugshots of criminals, missing persons and persons of interest. When taking mugshots, officers should be aware of the image resolution to fulfil the standards of image quality.

A common method of evaluating image quality is to measure the number of pixels between the eyes or pupils (interpupillary distance). Recommendations vary between 20 and 60 as a minimum for interpupillary distance. However, other experts state that using this measure as a threshold is often insufficient to confirm image quality.[20]

Depending on the context, these standards should be defined by the law enforcement agency, taking into account the nature of the case in question, the established legal frameworks, the results of internal testing, the strength of the algorithm, and any recommendations from the technology provider regarding their system in particular.

DATA REQUIREMENTS

A good machine learning model requires that both data **quantity** and **quality** are ensured. Generally speaking, simple models with large data sets are more accurate and effective than complex models with small data sets. Data sets that are too large or complex to be dealt with by traditional data-processing application software are usually referred to as **big data**. Big data is traditionally understood as having at least three distinct dimensions: *volume*, *velocity*, and *variety*, which are also known as “the 3 Vs”.

Machine Learning models usually require training with large data sets. However, it is also not useful to have a large amount of data if it is:

- **inaccurate** – it does not correctly represent the situation, for example there are duplicate entries, missing or incorrect data,
- **unrepresentative** – it does not correctly represent the diversity of the population, for example in terms of gender, race and ethnicity, or
- **outdated** – it does not correctly represent the current state of the situation.

For this reason, some authors then go on to add more Vs to “the 3 Vs” list, also including variability, veracity and value. The quality of the outcome highly depends on the quality of the training data. Even if the algorithm is very advanced and sophisticated, if the data is bad the results will also be bad.

Figure 22. How to select a representative sample of the population to have good quality data.

Furthermore, the data available on the internet is mostly representative of North America and European contexts, so it is difficult to obtain data quality and quantity on certain parts of the world.

In order to guarantee the quality of data, some organizations propose standardized processes for documenting databases. Similar to electronics datasheets, databases may be accompanied with a datasheet that documents its motivation, composition, collection process, recommended uses, etc.[21] This is a good practice to ensure transparency and accountability.

|▶ *Learn more about the importance of transparency and accountability on the* ***Principles for Responsible AI Innovation.***

Algorithmic bias

Algorithmic bias describes systematic and repeatable errors in an AI model that create "unfair" outcomes, such as discriminating against one category over another in ways that differ from the intended function of the algorithm.

Bias can result from many factors, including the design of the algorithm or the way data was collected, selected or used to train the algorithm. Because algorithms are often considered to be neutral and unbiased, they can inaccurately project greater authority than human expertise (in part due to the psychological phenomenon of automation bias). However, bias can enter algorithmic systems as a result of pre-existing cultural, social, or institutional expectations, the technical limitations of their design, or by being used in inappropriate contexts or by users who are not considered in the software's initial design.

Algorithmic bias has been cited in cases of gender, racial, and other types of **discriminatory decisions** made by machine learning systems. One widely discussed case was denounced by Propublica, in which they claimed that a tool to score the risk of recidivism by criminal defendants was biased against African Americans.[22] Some predictive policing systems have also come under close scrutiny, and their use has been curtailed due to the biases that were discovered.[23] The principal source of bias in these two cases was the **bias in historical decisions** that was reflected in the data points used to train the machine learning models.

There are also other possible sources of bias in data sets, such as **unbalanced classes**. For example, if one class has fewer data points than the other, the algorithm has fewer opportunities to learn the nuances of that class and may, therefore, produce more errors for that class. Currently, since most open-source face databases contain images of lighter-skinned individuals, facial recognition models trained on those databases can have higher error rates when recognising darker-skinned faces.[24]

One very important step to mitigate and exclude bias in databases is exploratory data analysis, which allows for investigating if the selected database is accurate and representative of the considered classes. Additionally, there are open-source tools and libraries that help with the detection and elimination of errors and bias, including *AI Fairness 360,*[25] *FairML*[26] and *Fairlearn*[27].

Artificial data

Besides the challenge of obtaining accurate and meaningful data, machine learning algorithms require vast quantities of data which may not always be available or accessible.

One possible solution for overcoming the absence of sufficient quantities of real-world data is to use augmented or artificial data. Artificial data can be produced for the purposes of training algorithms by adding slightly modified copies of already existing data or newly created synthetic data from existing data.[28] This can be done using content generation models to generate text or images based on similar databases.

Artificial data can also be used to create bias-free balanced databases, which can be used to train and generate fairer models, or to avoid the use of personal data and thus increase privacy.

EXPLORATORY DATA ANALYSIS

The first step in getting to know your data set is exploratory data analysis (EDA). The main purpose of EDA is to help understand the data before making any assumptions, and to ensure that the results produced are valid and applicable to the desired goal. This is essential in the early stages, to understand which data are most suitable for solving the problem at hand.

EDA includes a set of methods for analysing and investigating the data and summarizing their main characteristics. It helps determine how best to clean, select and transform data points to get the answers you need. It can help identify obvious errors and detect outliers, anomalous events or bias, as well as better understand patterns within the data and find useful connections between the variables.

EDA comprises several different methods, such as:

- **Data visualization**, a data science method to observe the data set values and range for each feature. This method also allows to understand visually the interactions between features in the data and the relationship between each feature and the label.

- **Dimension reduction** methods to reduce the number of features to be analysed.

DATA PRE-PROCESSING

After defining a data set and analysing it using EDA, data normally needs to be cleaned and transformed into a format that the model can read. This is usually the most time-consuming, but also the most important, step in terms of guaranteeing that a machine learning model performs to a high level. Several different methods are used for pre-processing data:

- **sampling**, which selects a representative subset from a large population of data;

- **transformation**, which combines and manipulates raw data to produce a single input;

- **denoising**, which removes noise from data;

- **imputation**, which synthesizes statistically relevant data for missing values;

- **normalization**, which organizes data for more efficient access; and

- **feature extraction**, which extracts a relevant feature subset that is significant in a particular context.

These tools and methods can be used on a variety of data sources, including data stored in files and databases or streaming data. Given its nature, unstructured data requires more work to pre-process, as it is not stored in a standard format such as a spreadsheet and has to be converted into a similar format. to be converted into a similar format.

Law enforcement agencies must take care to ensure that relevant data is not deleted or mistakenly transformed during this stage. Over-cleaning the data can nuance or compromise important evidence or even result in inaccurate outputs.

DATA ANONYMIZATION

If the model includes the processing of **sensitive** or **personal** data, techniques such as anonymization and pseudonymization should be implemented in order to protect the **right to privacy**.

Data anonymization consists of transforming personal data into anonymous data so that the individuals or groups of individuals to which the data belongs are no longer identifiable in the data.

Pseudonymization consists of the removal or substitution of all direct identifiers with other unique identifiers in such a way that unique individuals are still distinguishable in a data set, but their identity cannot be traced back without access to additional information.

There are libraries and platforms online that support data privacy and anonymization. Presidio ensures sensitive data is properly managed and governed, and SmartNoise injects noise into data to prevent the disclosure of sensitive information and manage exposure risk.

While there may be merits to anonymization, it should be noted that anonymized data are still personal data if the anonymization can be reversed – meaning that, with appropriate skills or technology, the data could be linked back to individuals or groups.[29]

PRACTICAL EXAMPLE — **CCTV footage anonymization**

Anonymization may be helpful if law enforcement is investigating a series of robberies that have occurred in a particular area, and they have obtained CCTV footage from several stores in that area. By anonymizing the footage, any identifying information about individuals who are not involved in the investigation can be removed, such as their faces, license plates, or other identifiable features. This would allow law enforcement to use the footage to identify potential suspects without compromising the privacy of innocent individuals who may have been captured in the footage.

THE COMPUTER

The third of the four fundamental components of AI systems is the computer. When developing an AI system, selecting the most appropriate computer resources can greatly increase the chances of success of the task at hand.

The first question to answer is: *where will the model run and be stored?* Several factors must be taken into account, including the processing needs, the available budget, and security concerns.

If the decision is to run it locally, the next question should be: *which computer processing unit is required?* This will depend on the complexity of the model and the size of the database to be processed.

This section tries to support you to answer both of these questions and make the best decision regarding the most appropriate hardware.

ON-PREMISES VS CLOUD

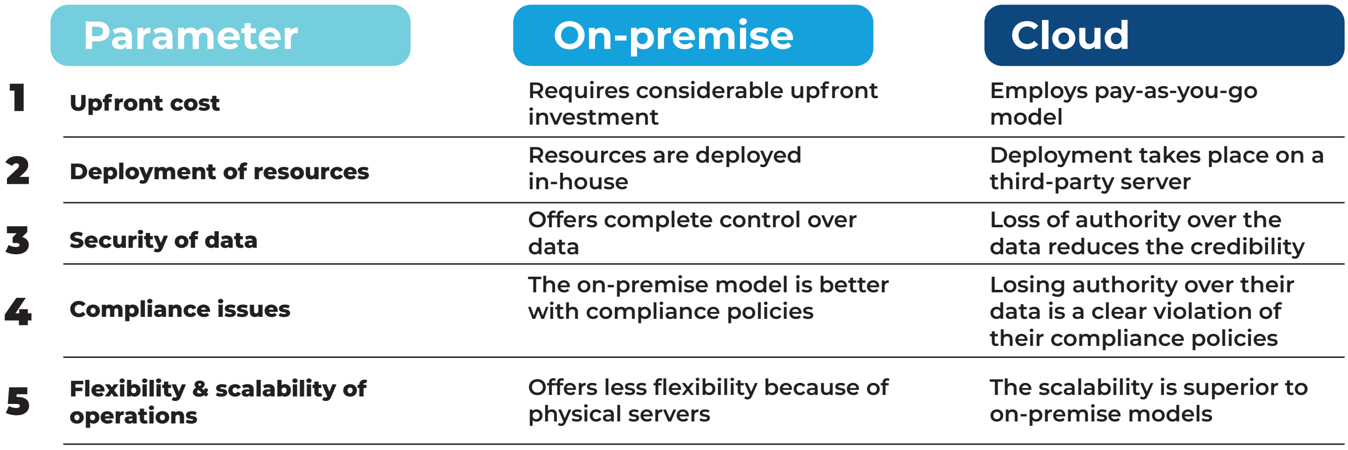
A standard laptop is enough to start experimenting with machine learning, explore the data set, or develop small models, as not much computation power is required to perform these tasks. However, as machine learning models get bigger and more complex, more advanced hardware with higher computation power is needed. Such computation power can either be located **on-premises** or accessed remotely via the **cloud**.

Organizations can create IT infrastructures to store the software within the organization's physical office and host it on-site. This is called **on-premises** software. In this case, all applications are hosted and processed locally, and the agency has to purchase servers, software licenses, and office space, and hire dedicated employees to manage the infrastructure. This means that the IT staff have more control over the server hardware and the data configuration, security, and management because they have physical access to the data. All data and important information are accessed locally, and no third party can access them remotely, which creates a safer environment. This may be necessary when confidential and sensitive information is being stored and processed.

The **cloud** refers to the software, servers, and services that run over the internet rather than locally on the organization's computers. Cloud services can provide easy but paid access to applications (i.e. software as a service – SaaS) and resources (i.e., infrastructure as a service – IaaS), without the need for internal infrastructure or hardware. Cloud servers with the latest processing units (which are often unaffordable otherwise) deliver advanced features which allow machine learning models to run faster. As they are not bound to a specific piece of hardware, cloud services allow users to experiment with different resources to select the most appropriate one for the case in question. However, this configuration carries more risks, as covert data may be accessed by third parties, and it should, therefore, be avoided when dealing with confidential or sensitive data.

In many cases, a **hybrid solution** may also be viable, where preliminary model testing to establish the computation power needed and the most appropriate hardware is carried out on cloud machines with synthetic data, and implementation is developed on-premises to guarantee that the input data is safe and processed using the best resources. If the data is not confidential, an alternative hybrid solution may also be adequate, in which the preliminary model is tested on-premises and is then moved to the cloud to widen the scale.

The image below shows the differences between on-premises and cloud servers.



	Parameter	On-premise	Cloud
1	**Upfront cost**	Requires considerable upfront investment	Employs pay-as-you-go model
2	**Deployment of resources**	Resources are deployed in-house	Deployment takes place on a third-party server
3	**Security of data**	Offers complete control over data	Loss of authority over the data reduces the credibility
4	**Compliance issues**	The on-premise model is better with compliance policies	Losing authority over their data is a clear violation of their compliance policies
5	**Flexibility & scalability of operations**	Offers less flexibility because of physical servers	The scalability is superior to on-premise models

Figure 23. Five key differences between on-premises and cloud servers.[30]

PROCESSING POWER

Depending on the complexity of the model and the size of the database, different processing capacities may be required. The processing capacity of a computer depends on the type of processing units it has, as they are the core element of all computers.

The main processor, the **central processing unit** (CPU), is the principal element in the computer system. It can perform complex mathematical calculations quickly, as long as they are sequential. When required to perform multiple tasks simultaneously, the CPU performance begins to slow down.

This differs from specialized processors such as **graphics processing units** (GPUs), which are specifically designed to speed up the rendering of images. GPUs are best suited for parallel processing. As training data sets continue to get larger, most AI applications require parallel processing to enable tasks to be performed efficiently, which means that in most cases GPUs are the preferred option for training AI models, especially neural networks.

However, CPUs are still useful for tasks that require sequential algorithms, or for performing complex statistical computations such as real-time training or recurrent neural networks that rely on sequential data.

WANT TO LEARN MORE?

Next generation processors

Because of the importance of computer processing power in deep learning applications, there has been considerable research into the development of more powerful chips. As deep learning projects advance and require more and more processing power, new more powerful custom-built processors are appearing on the market. One example is tensor processing units (TPUs), specifically made to run projects using *TensorFlow*, an open-source framework and library developed by Google[31]. In addition to these advanced processing units, researchers are using Field-Programmable Gate Arrays (FPGAs) to successfully run machine learning workloads which can outperform even GPUs when running neural networks.

It is important to understand the advantages and disadvantages of each option and to procure hardware or use cloud processing platforms which are powerful enough to efficiently process the AI system that the law enforcement agency wishes to use.

In the event that the agency purchases its AI system, the supplier will usually provide information about the hardware specifications required to run their software. Keep in mind that, the considerations that need to be taken into account before acquiring new software go beyond its cost.

THE HUMAN

Last but not least, the fourth fundamental component of an AI system is the human – a particularly important component in terms of responsible AI innovation. Machine learning models are developed by humans and rely on data that has been collected based on human decisions, prepared by humans, which often relates to humans. Since humans are not impartial and are susceptible to several types of cognitive bias, models and data sets can inadvertently reflect and reproduce these biases.

To prevent this from happening or minimize the negative effects, we first need to be aware of these common human biases. Secondly, we need to proactively analyse the model and explore the data to identify sources of bias before training the model. Finally, we must assess and monitor model predictions to check for bias. Since all of these steps are susceptible to human error, the final output produced by the system, which will be probabilistic and not factual, will naturally not be free of errors. This is why decisions that have an impact on society or individuals should always be reviewed and ultimately made by humans.

COGNITIVE BIASES

When building AI models, it is important to be aware of common human biases that can manifest in the data or in the model design. Awareness of the various types of cognitive biases can help humans take proactive steps to mitigate their occurrence and effects. There are hundreds of different types of bias that can affect human judgments and decisions,[32] but we will only discuss the types that are commonly seen in the AI field. For each type of human bias, an example related to facial recognition technology is provided.

- **Automation bias:** tendency to favour results generated by automated systems over those generated by non-automated systems, irrespective of the error rates of each. *Example: using the output of a facial recognition system as a final decision without having it checked by a face examiner.*

- **Implicit bias:** tendency to make assumptions based on one›s own mental models and personal experiences that do not necessarily apply more generally. *Example: defining camera brightness based on people with light skins when creating a facial recognition database.*

- **Selection bias:** tendency to select a data set that is not reflective of the real-world distribution. *Example: facial recognition data sets that have fewer examples of one specific ethnic group but will then be used to recognize individuals from that group.*

- **Group attribution bias**: tendency to generalize something which is true of some individuals in a group to the entire group to which they belong. Two key manifestations of this bias are:
  - **In-group bias**: tendency to prefer members of a group to which you also belong, or who display characteristics that you also share. *Example: the face examiner reviewing the outputs of the facial recognition system might favour members of the same ethnicity.*

  - **Out-group homogeneity bias**: tendency to stereotype individual members of a group to which *you do not belong*, or to see their characteristics as more uniform. *Example: the difficulty of the face examiner in distinguishing faces from other ethnic groups because "they look more similar".*

- **Reporting bias**: occurs when the frequency of events, properties, and/or outcomes captured in a data set does not accurately reflect their real-world frequency. This bias can arise because people are more likely to document circumstances that are unusual or especially memorable. *Example: only evaluating facial recognition systems when the system provides incorrect outputs.*

CORRELATION VS CAUSATION

Another common human tendency is to look for patterns in nature. When two events appear to be closely associated, humans tend to try to find a causal association between the two, even when it does not exist.

As explained in the regression section, **correlation** describes the relationship between two variables and determines whether the value of one variable moves *with* another variable (in the same direction or in the opposite direction). **Causation** can exist simultaneously and occurs when one variable impacts the other. If a causal relationship between two variables exists, making adjustments to one variable will affect the other variable. However, if the relationship is one of correlation only and not causation, the adjustment will have no effect.

It is very important to distinguish correlation from causation to avoid reaching inaccurate conclusions. When there is a correlation between two variables – event A and event B – we cannot simply assume that event A caused event B, because other options may also be viable:

- The correlation between the two variables might be a coincidence.

- The variables might be associated in other ways:
  - The opposite is true: event B actually caused event A.
  - The two events were caused by another event: A and B are correlated, but they are actually caused by C.
  - There is a chain reaction: A causes C, which leads C to cause B.
  - There is a conditional cause: A does cause B, as long as C happens.

PRACTICAL EXAMPLE

Causation and correlation

Here is an example to gain a better understanding of the difference between causation and correlation. The figure below shows the relationship between two variables: *house breaking index* and *ice cream sales*.

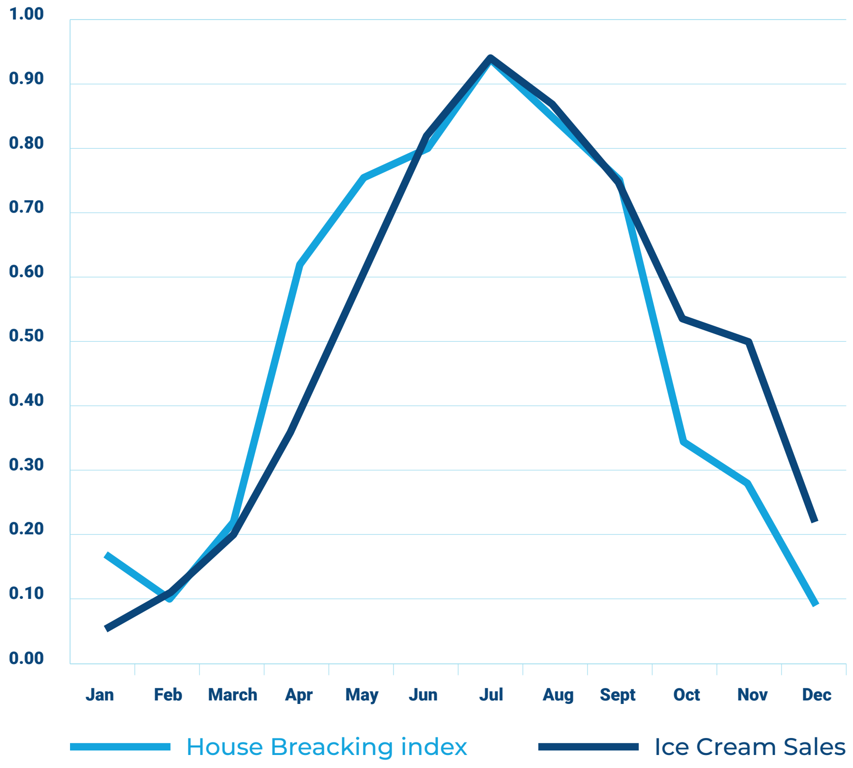


Figure 24. Correlation between house breaking index and ice cream sales along the year.

According to the figure, these two variables are correlated in time: both variables reach their highest value during the months of June and July. However, this does not necessarily mean that there is a **causal** link between the two. In other words, it does not mean that an increase in ice cream sales results in an increase in house breakings, or that an increase in house breakings results in an increase in ice cream sales.

After some investigation, it can be concluded that there is a third factor involved, also known as a **confounder**, that impacts both variables – Summer. As the weather is hot, people buy ice cream and also go on vacation, and have their windows and doors open to keep the house cool, which leaves their houses vulnerable to house breaking. The weather is the *hidden factor* that brings both variables together.

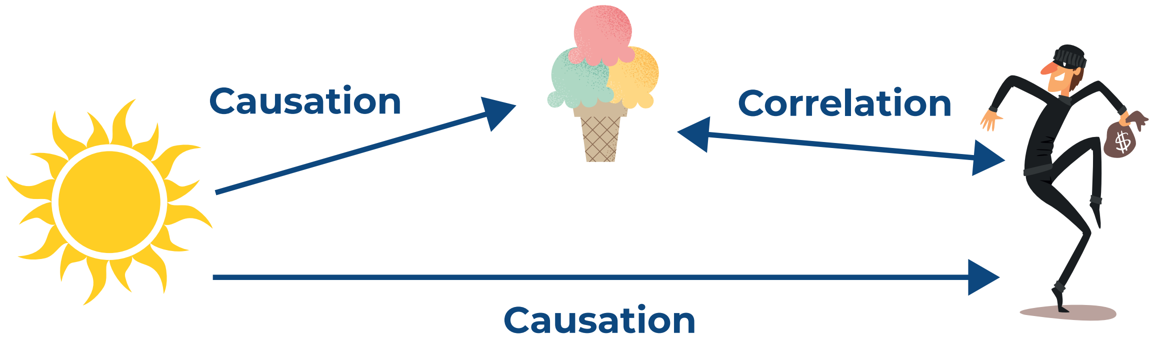


Figure 25. Correlation between the house breaking index and ice cream sales, and causation between Summer and ice cream sales and Summer and the house breaking index.

As expected, there is no causal link between *ice cream sales* and *house breakings*, only a correlation. Ice cream sales and house breakings both have a causal relationship with the weather. This example also shows how important it is to select the appropriate features for each task and not to simply use all types of data available.

Correlation does not always imply causation, but causation always implies correlation. Once a correlation is detected, we can test for causation by running controlled experiments that track each variable while also excluding other variables that might interfere. In the law enforcement context, this test is not always possible, so it is important to obtain more information about the context in order to verify if there is a genuine causal relationship between the variables.

UNDERSTANDING AI SYSTEMS

Machine learning algorithms, particularly deep learning models, can sometimes be composed of millions of internal parameters that are used to compute a single output for a given input (see neural networks section). This can make extremely difficult to understand the internal mechanisms that lead the system to produce the output, and it may not be clear how the model produced the result – this is also known as the **black box problem**. For this reason, it is difficult to trace and verify results, which can raise issues in terms of human rights, particularly the right to a fair trial and due process for an accused individual or individuals. One essential requirement for the responsible use of AI systems is the notion of explainability. **Explainability** focuses on ensuring that there is a certain level of human understanding of how the model reaches its outputs.

|▶ *Learn more about the importance of explainability of AI systems in the* ***Principles for Responsible AI innovation.***

Certain classes of algorithms, including more traditional machine learning algorithms, tend to be more readily explainable, although they may not perform as well. Others, such as neural networks, while performing better, remain much harder to explain. Improving our ability to explain AI systems remains an area of active research, with tools being developed to explain local and global decisions,[33] such as InterpretML,[34] SHAP[35] and LIME.[36]

However, good explanation systems should also be balanced, and should explain both why a forecast may be correct, and why it may be wrong. Tools for generating balanced explanations are also being developed, such as BEEF: Balanced English explanations of forecasts.[37] These tools can help explain and present, in plain terms, the features in the data that were most important for the model and the effect of each feature on any particular output.[38] The example below[39] shows a tool that creates a heatmap of the pixels that were more relevant (marked in red) to the classification produced.

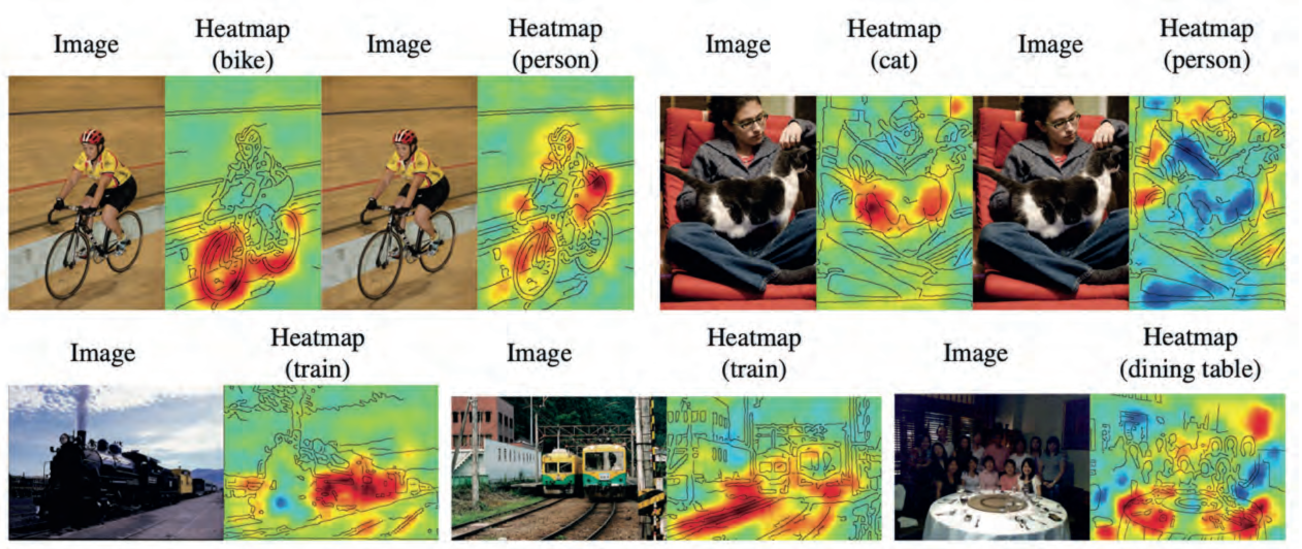


Figure 26. Images shown next to the heatmaps of an explainability tool when considering the prediction score for a particular classification. Copyright (2016) by Lapuschkin and others. Reprinted with permission.[40]

We can see that the model works well in most cases, but there is one instance where the train rails actually play a bigger role in a “train” classification than the train itself. This explainability method helps the developers to understand and debug their models to be able to achieve responsible AI.

THE APPLICATIONS

AI has found application in several fields, resulting in the emergence of different machine learning algorithms, which have been optimized for the type of data they process. For instance, images and text are structured differently and therefore different architectures of algorithms can be employed.

This section addresses the main applications of AI that can be used by law enforcement agencies, covering image processing, text and speech processing, risk evaluation and predictive analytics, content generation and process optimization and workflow automation.

COMPUTER VISION – IMAGE ANALYSIS – OBJECT RECOGNITION

Computer vision is a broad and interdisciplinary field that enables machines, particularly computers, to interpret the visual world, much like human vision. It involves the development of algorithms, techniques, and systems that allow machines to extract meaningful information from digital images, videos or other visual input.

Computer vision techniques often incorporate **image analysis** methods as part of their processes. Image analysis algorithms are trained on large data sets to automatically recognize, classify, or contextualize an image or elements within that image.

This is possible thanks to **object recognition,** which specifically deals with identifying and classifying objects within images or scenes. It involves algorithms and techniques that enable computers to process pictures and identify geometrical shapes and, ultimately, objects.

PRACTICAL EXAMPLE

How license plate recognition systems are developed

License plate recognition systems are machine learning systems designed to identify numbers and letters contained in pictures or videos. They include algorithms trained with pictures depicting numbers and letters which are labelled with the correct identification of the characters depicted. (In videos, the process is similar since each video frame is processed individually as a single picture.)

After analysing thousands of pairs of pictures and labels, the algorithm learns to distinguish characters by creating rules extracted from repetitive patterns found in the training data. For example, it learns that the number "1" normally has a shorter line at the top than the number "7" and that the number "7" can have one extra line in the middle.

After learning all these specific rules, the AI system can apply them to new circumstances, so that when it receives a picture of a license plate without any label, it can recognize the numbers and letters it depicts.

Image analysis is a complex task that requires large amounts of training data – images – as well as advanced algorithms – neural networks. Object recognition is a common task within image analysis. It consists of identifying specific objects in images – i.e., any distinguishable element within an image, including people, animals or other things.

How do these AI systems identify these elements in images? Read the next box to learn more.

WANT TO LEARN MORE?

A more technical explanation of image processing

Image analysis systems often use a type of algorithms called **convolutional neural networks** (CNN). This is a type of deep learning algorithm that has multiple layers of nodes that analyse and correlate adjacent pixels (locally connected nodes). This allows the algorithm to make connections and extract meaning from neighbouring pixels.[41]

CNN models use filters to identify features like edges, textures, or shapes within an image, just like how detectives search for specific clues in a crime scene. When moving from input to output, CNN layers progressively learn more complex and abstract features from simple shapes to high-level representations.

For example, if the input is a picture of a face, the first layer of the algorithm may identify lines and curves. The following layers may then identify a combination of lines and curves which lead to the identification of facial features, and the last layer may identify different faces. The image below illustrates this.

The CNN learns to recognize entire objects or scenes by combining all the learned features, similar to how an officer puts together all the gathered information to solve a case.

The final layer makes the decision. In this example, it finds the corresponding **name** – label or output – to a certain **face** – or input image.

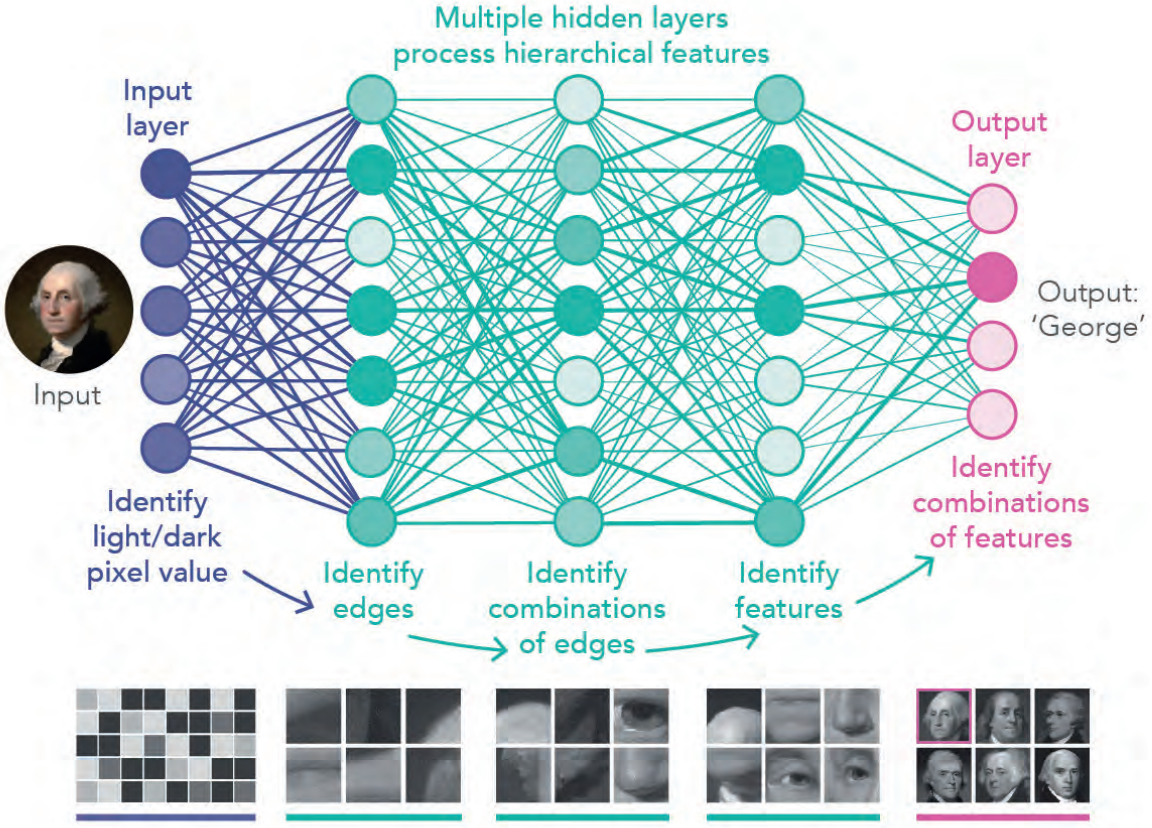


Figure 27. Neural network architecture for image processing. While the input layer identifies pixel intensity grey levels, the following layers identify higher-level facial features. Copyright (2019) by Lucy Reading-Ikkanda/PNAS. Reprinted with permission.[42]

CNN models can assist in analyzing visual data relevant to law enforcement, such as applications in forensic investigations, like enhancing images, recognizing faces, identifying objects, or analyzing CCTV footage.

Facial recognition

Facial recognition technology is a widely used variation of object recognition within the law enforcement context. This technique consists of recognizing or helping the identification of persons of interest in photographs or videos. It does so by comparing and analyzing the patterns, shapes and proportions of an individual's facial features and contours with images of faces in a database.

PRACTICAL EXAMPLE

Facial recognition for biometrics

Law enforcement investigators use facial recognition software in biometrics for two main purposes:

- ***Biometric verification***, which is a "one to one" (1 to 1) comparison of two images to verify someone's identity against, for example, an identification document (ID).
- ***Biometric identification***, which consists of a "one to many" (1 to n) comparison of an image of a person against a database of images in order to search for their identity.

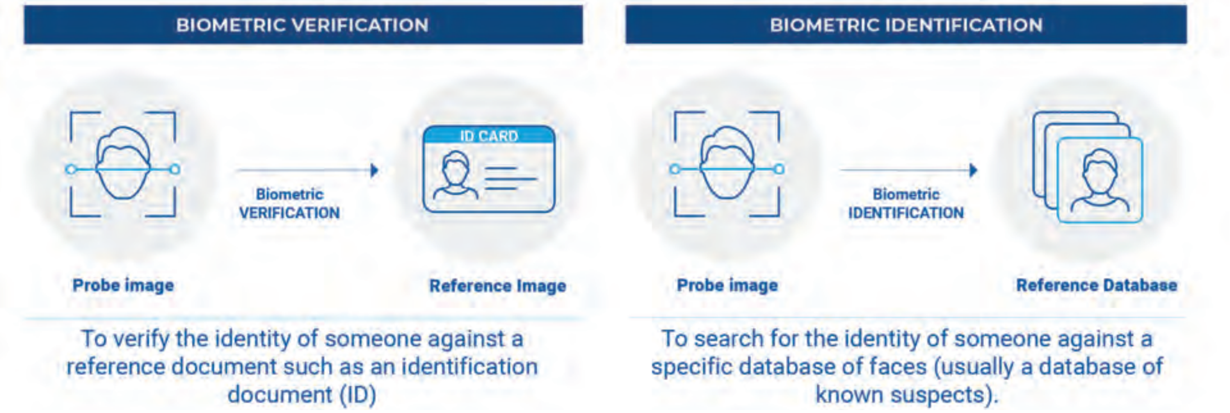

BIOMETRIC VERIFICATION
BIOMETRIC IDENTIFICATION
Biometric VERIFICATION
ID CARD
Probe image
Reference Image
To verify the identity of someone against a reference document such as an identification document (ID)
Biometric IDENTIFICATION
Probe image
Reference Database
To search for the identity of someone against a specific database of faces (usually a database of known suspects).

Figure 28. Main purposes of facial recognition in biometrics.

To identify an unknown person of interest, investigators work with:

- A **probe image**, representing a person of interest to be submitted for biometric verification or identification. Investigators can already have the probe image of the person of interest or they can collect it from CCTV footage, for example.
- A **reference image**, usually an official document or card with the name and photograph or other information on it that can be used to prove someone's identity.
- A **reference database**, which consists of the repository of images against which a probe image is compared. In the law enforcement context, it is common to use a reference database of known suspects, composed of photos and mugshots of criminals, missing persons and persons of interest.

These models are trained with pairs of images (two pictures of the same person or two pictures of different persons) associated with a binary label ("same person" or "not the same person"). After reviewing large amounts of examples, the algorithm learns to identify the key features that characterize an individual. By applying this comparison process, the facial recognition model can then be used for biometric verification (comparing a suspect's image with an identification document) or for biometric identification (comparing a suspect's image with each image of the known suspects database).

To avoid performance issues, it is important that users of facial recognition systems are experts at performing a comparison of faces image-to-image, and review the outputs of facial recognition systems to avoid blindly relying on results generated by automated systems, mitigating the risk of automation bias.

|▶ *Read more about the risks and concerns of facial recognition technology in the **Introduction to Responsible AI Innovation.***

TEXT AND SPEECH ANALYSIS – NATURAL LANGUAGE PROCESSING

These AI techniques consist of analyzing large data sets to recognize, process, and tag text or speech in order to extract meaningful information. The processing of text is possible thanks to natural language processing (NLP). NLP is a field that crosses both linguistics and computer sciences and seeks to process and analyze large amounts of natural language data either in the form of text or in voice recordings.

NLP starts by applying linguistics to pre-process the text, separating sentences into words, simplifying word variations, and removing stop words, i.e., words that are not relevant to the meaning of the sentence. Below are a few examples of the pre-processing usually applied to any NLP task.

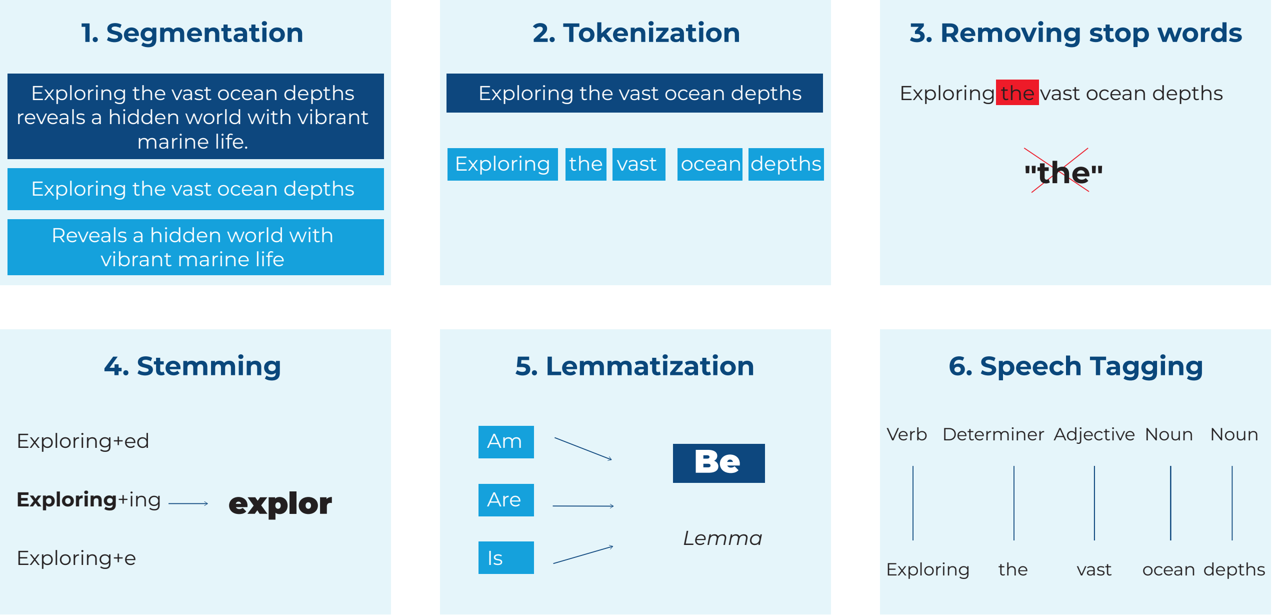


Figure 29. Pre-processing techniques to clean and break down a text before feeding it into natural language processing models.

Once the text is clean, the algorithm can extract information about the topic and meaning of the text, identify details such as locations, people's names etc., or predict the next word in a sentence.

Tasks in NLP frequently involve speech recognition, translation or natural language generation.

WANT TO LEARN MORE?

A more technical explanation of natural language processing

Natural language processing is often carried out using models like Recurrent Neural Networks (RNN), Long Short-Term Memory Networks (LSTM), or more recently, Transformer models.

Recurrent neural networks are neural networks with nodes connected along a temporal sequence. They rely on an internal memory that processes sequences of inputs to extract morphosyntax and semantic functions based on a sequence of words.[43]

The most advanced models, such as the GPT, use other types of algorithms called **transformers**. Transformers are also designed to handle sequential input data but do not necessarily process the data in order. For example, if the input data is a sentence, the transformer does not need to process the beginning of the sentence before the end. Rather, it identifies the context that confers meaning to each word in the sentence. This reduces the time needed to train the algorithm by allowing it to perform multiple tasks in parallel.[44]

The image below illustrates the different architectures of recurrent neural networks and transformers.

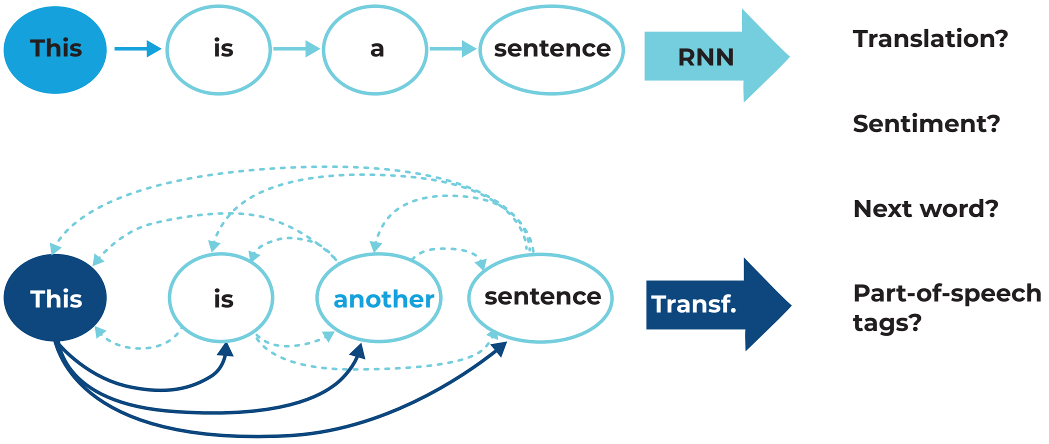


Figure 30. Differences between text analysis performed by two different natural language processing models: a Recurrent Neural Network (RNN) and a Transformer. Copyright (2021) by Chaitanya K. Joshi. Adapted with permission.[45]

RISK EVALUATION AND PREDICTIVE ANALYTICS

Risk evaluation algorithms or predictive analytics allows to analyze large data sets to identify patterns that can recommend courses of action or trigger specific actions. In other words, these models use statistical algorithms and machine learning techniques – such as classification and regression – to identify the likelihood of future outcomes based on historical data.

Predictive models analyze what has already happened and extrapolate likely outcomes in related contexts. These outputs can be presented using simple charts, graphs, and scores that indicate the probability of certain events happening in the future. These predictions can support and guide decision-making processes but should never be seen as facts.

PRACTICAL EXAMPLE

Predictive policing

Predictive policing is an example of predictive analytics and consists of using algorithms to analyze pre-existing crime data in order to predict and help prevent potential future crimes. While law enforcement work is traditionally rather reactive by nature, predictive policing could, in some instances, be used to forecast when, where and what type of crimes are most likely to occur.

These models are trained on historical crime data, including details like types of crimes, locations, dates, times, and any other relevant contextual information. This data often spans several years.

Machine learning algorithms, such as classification and regression models, like random forests or neural networks, can be trained to recognize patterns and correlations between the identified features and instances of crime.

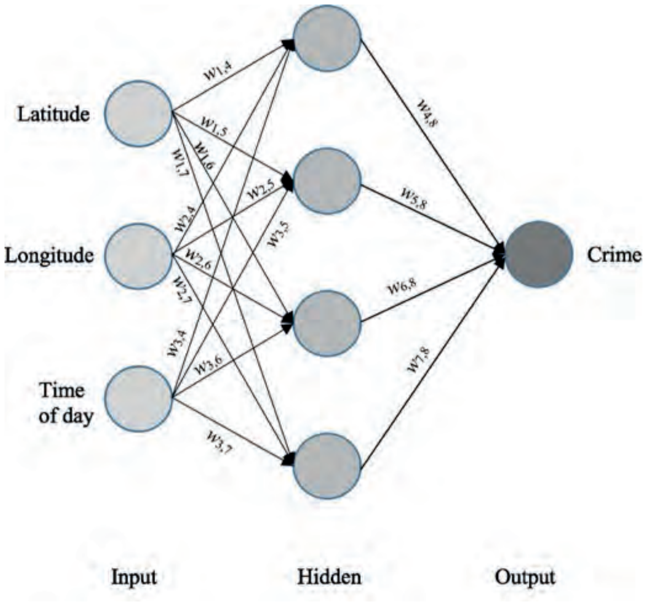


Figure 31. Predictive policing schematical neural network with three input nodes collecting information about latitude, longitude and time of the day and one binary output node predicting if a crime will occur or not.

Predictive policing is intended to supplement traditional policing methods by providing data-driven insights to help law enforcement agencies make informed decisions about resource allocation and crime prevention strategies. It is a tool aimed at optimizing law enforcement efforts, reducing response times, and potentially deterring criminal activity. However, concerns have been raised about the potential for biases in data used for predictions, as well as the ethical implications (such as feedback loops) and privacy concerns associated with the use of predictive algorithms in law enforcement. Therefore, predictive policing might be banned in some jurisdictions. It is therefore required to guarantee that predictive policing is used within the limits of the law and responsibly – i.e., in a way that is aligned with ethics and compliant with human rights.

|▶ *Read more about the risks of predictive policing in the Introduction to Responsible AI Innovation.*

CONTENT GENERATION

Content generation using machine learning techniques involves sophisticated algorithms that create new text or images based on patterns learned from existing data.

In text generation, the goal of the algorithm is to predict the next word in a sequence, leveraging contextual information from previous words. This enables the generation of coherent and contextually relevant text, applicable in various domains such as language generation for chatbots, automated content creation, and even storytelling. The most common algorithms for text generation are NLP architectures such as Recurrent Neural Networks (RNN), Long Short-Term Memory Networks (LSTM), or Transformer models (such as BERT, GPT).

Image generation is also possible, using machine learning models, particularly Generative Adversarial Networks (GANs) or Variational Autoencoders (VAEs), to create new images based on learned representations of existing data. These models are trained on large data sets of existing images, providing the model with patterns, features, and representations of the visual information. Once trained, the models can produce new images by taking random noise as input and transforming it into images that resemble the patterns learned from the training data.

WANT TO LEARN MORE?

A more technical explanation of image generation

Image generation systems often include a type of algorithm called a Generative Adversarial Network (GAN). These algorithms include two neural networks that compete with one another, thereby improving their respective performances. One is called the Generator and the other is called the Discriminator. The image below illustrates how GANs work to create realistic pictures of human faces.

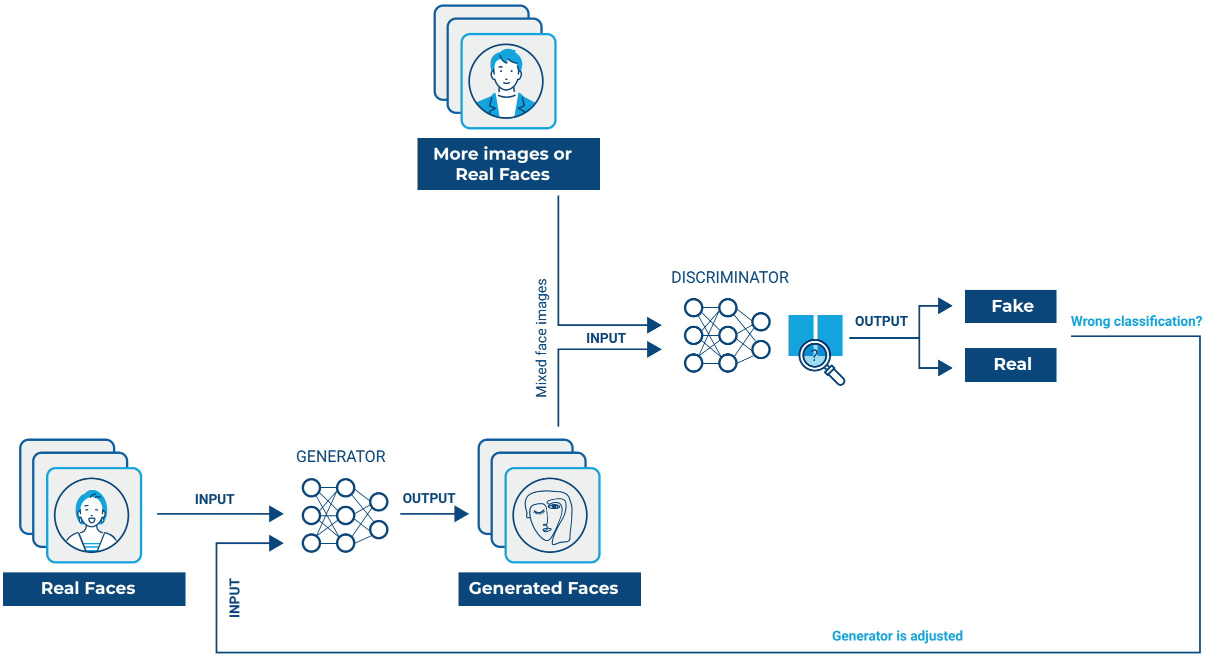


Figure 32. Training method of generative adversarial networks.

The Generator is trained on a data set of pictures with faces to produce new samples that mimic these pictures. The generated images are included in a data set mixed with “real” images, i.e., pre-existing pictures of human faces that were not created by the Generator. This mixed data set is then used to train the Discriminator to distinguish between real and fake samples.

At first, the generated pictures barely resemble real pictures, but as the Generator and the Discriminator keep running against each other, the Generator improves until it is able to create fake pictures that the Discriminator cannot distinguish from real pictures. These pictures are completely original fabrications of the AI system, though they essentially mimic the data set it was trained with.[46]

Besides GANs, more advanced algorithms such as diffusion models have emerged recently and have been included in image generation systems that have gained prominence such as Midjourney AI and DALL-E 2.[47]

Challenges persist in content generation, particularly in ensuring the quality and diversity of the generated content. Maintaining coherence and realism in text remains a significant challenge. In image generation, issues include generating visually consistent images without artifacts or distortions. Ethical considerations are also paramount, as biases present in the training data could be perpetuated in the generated content, impacting fairness and inclusivity.

Despite these challenges, the applications of content generation are far-reaching in the law enforcement context. Generating incident reports, case summaries, or legal documentation can streamline administrative tasks for law enforcement officers. In forensic investigations, generating visual reconstructions or simulations based on collected data can also assist in analyzing and presenting evidence to support investigations and legal proceedings.

Content generation models can also be used to create artificial data and synthetic data to aid in training robust and bias-free machine learning algorithms.

PROCESS OPTIMIZATION AND WORKFLOW AUTOMATION

Process optimization and workflow automation consists of analyzing large data sets to identify anomalies, patterns, predict outcomes or ways to optimize and automate specific workflows.

In the law enforcement context, this can be applied to streamline investigations, making connections between pieces of evidence and finding patterns correlated with events, time and places.

PRACTICAL EXAMPLE | **Link analysis in drug trafficking network investigations**

Link analysis in the law enforcement context involves mapping connections and relationships between individuals, organizations, or entities based on various types of data. This method helps investigators understand complex networks, uncover patterns, and identify key actors or nodes within criminal activities.

Let’s consider a case of investigating a drug trafficking network.

Law enforcement gathers data from various sources such as phone records, financial transactions, surveillance footage, arrest records, and informants. Using machine learning systems, investigators create a visual representation of the network. Each node represents an individual or entity, and the links between them signify relationships, communications, transactions, or associations.

Investigators analyze the network map to identify patterns, such as frequent communication between certain individuals, common locations, financial transactions, or hierarchical structures. Focusing on specific nodes or individuals within the network, they can identify key players like the leaders or primary suppliers, as well as peripheral members or facilitators involved in the network.

Analyzing the connections and activities within the network helps in making predictions or anticipating future moves. For example, understanding how changes in communication patterns might indicate an upcoming drug shipment. The insights gained from link analysis assist in building a stronger case by establishing connections between suspects, providing evidence of their involvement, and supporting the prosecution’s arguments.

Link analysis in this context enables law enforcement to comprehend the structure of criminal networks, identify crucial targets for investigation, and disrupt illegal activities. It helps in unraveling complex webs of connections and provides a clearer picture of how different individuals or entities are involved in criminal enterprises.

ENDNOTES

1 Andreas Kaplan & Michael Haenlein. (2019). "Siri, Siri, in my hand: Who's the fairest in the land? On the interpretations, illustrations, and implications of artificial intelligence". Business Horizons.

2 Daniel Crevier. (1993). AI: The Tumultuous Search for Artificial Intelligence. New York, NY: BasicBooks.

3 Henry Shevlin, Karina Vold, Matthew Crosby and Marta Halina. (2019). The limits of machine intelligence. Accessible at: *EMBO reports*, 20(10). doi:https://doi.org/10.15252/embr.201949177.

4 Hal Hodson. (2019). *DeepMind and Google: the battle to control artificial intelligence*. [online] The Economist. Accessible at: https://www.economist.com/1843/2019/03/01/deepmind-and-google-the-battle-to-control-artificial-intelligence.

5 Nick Bostrom. (2014). Superintelligence: Paths, Dangers, Strategies. Oxford University Press.

6 Rory Cellan-Jones. (Dec. 2, 2014). Stephen Hawking warns artificial intelligence could end mankind. BBC. Accessible at https://www.bbc.com/news/technology-30290540

7 Jon Russel. (May 25, 2017). Google's AlphaGo AI wins three-match series against the world's best Go player. TechCrunch. Accessible at: https://techcrunch.com/2017/05/24/alphago-beats-planets-best-human-go-player-ke-jie/amp/?guccounter=1

8 Benoît Vibert & Christophe Charrier & Jean-Marie Lebars & Christophe Rosenberger. (2015). Comparative Study of Minutiae Selection Algorithms for ISO Fingerprint Templates. Proceedings of SPIE - The International Society for Optical Engineering. 9409. 10.1117/12.2080795.

9 Chander Kant & Rajender Nath. (2009). Reducing Process-Time for Fingerprint Identification System. International Journal of Biometric and Bioinformatics.

10 Paolo Contardoa, Paolo Sernania, Nicola Falcionellia and Aldo Franco Dragonia (2021). Deep learning for law enforcement: A survey about three application domains. Accessible at: http://ceur-ws.org/Vol-2872/paper06.pdf

11 Christin-Melanie Vauclair & Boyka Bratanova. (2017). Income inequality and fear of crime across the European region. European Journal of Criminology, 14(2), 221–241. Accessible at: https://doi.org/10.1177/1477370816648993

12 Christin-Melanie Vauclair & Boyka Bratanova. (2017). Income inequality and fear of crime across the European region. European Journal of Criminology, 14(2), 221–241. Accessible at: https://doi.org/10.1177/1477370816648993

13 INTERPOL. About Red Notices. Accessible at: https://www.interpol.int/How-we-work/Notices/About-Red-Notices

14 Mobin Zhao, Wangzhi Li, Yongjie Fu, Kangrui Ruan, Xuan Di. (April 2021). CVLight: Decentralized Learning for Adaptive Traffic Signal Control with Connected Vehicles. https://arxiv.org/abs/2104.10340

15 Wikipedia. (2023). k-nearest neighbors algorithm. Accessible at: https://en.wikipedia.org/wiki/K-nearest_neighbors_algorithm#/media/File:KnnClassification.svg

16 Github. What-If Tool. Accessible at: https://pair-code.github.io/what-if-tool/

17 Error Analysis. Accessible at: https://erroranalysis.ai

18 OECD. (2008). Glossary of Statistical Terms. Accessible at https://stats.oecd.org/glossary/
Oxford Dictionary.

19 *Ibidem*

20 WEF, INTERPOL, UNICRI, The Netherlands Police. (Nov. 2022). A Policy Framework for Responsible Limits on Facial Recognition Use Case: Law Enforcement Investigations. https://unicri.it/A-Policy-Framework%20-for-Responsible-Limits-on-Facial-Recognition

21 Timnit Gebru, Jamie Morgensen, Briana Vecchione, Jennifer Wortman Vaughan, Hanna Wallach, Hal Daumé III & Kate Crawford. (Dec. 2021). Datasheets for Data sets. Communications of the ACM, Vol 64(12): pp. 86-92. Accessible at: https://www.microsoft.com/en-us/research/publication/datasheets-for-data sets/

22 Julia Angwin, Jeff Larson, Surya Mattu, and Lauren Kirchner. (May 2016) "Machine bias: There's software used across the country to predict future criminals, and it's biased against blacks". ProPublica.

23 Rashida Richardson, Jason Schultz & Kate Crawford. (2019). Dirty Data, Bad Predictions: How Civil Rights Violations Impact Police Data, Predictive Policing Systems, and Justice. SSRN Scholarly Paper No. 3333423. Accessible at: https://papers.ssrn.com/abstract=3333423

24 NISTIR 8280. (2019). Face Recognition Vendor Test (FRVT) Part 3: Demographic Effects. Accessible at: https://nvlpubs.nist.gov/nistpubs/ir/2019/nist.ir.8280.pdfhttps://www.nist.gov/publications/face-recognition-vendor-test-part-3-demographic-effects

25 AI Fairness 360. IBM Research Trusted AI. Accessible at: https://aif360.mybluemix.net

26 J. Adebayo. (2023). FairML: Auditing Black-Box Predictive Models [Python]. Accessible at: https://github.com/adebayoj/fairml (Original work published 2016).

27 Improve Fairness of AI systems. Fairlearn. Accessible at: https://fairlearn.org

28 Pierluigi Casale. (Feb. 13, 2020). How does Artificial Intelligence improve map making, Tomtom. Accessible at https://www.tomtom.com/blog/maps/artificial-intelligence-map-making/

29 United Nations Development Group. (2017). Data Privacy, Ethics and Protection: Guidance Note on Big Data for Achievement of the 2030 Agenda. Accessible at https://unsdg.un.org/resources/data-privacy-ethics-and-protection-guidance-note-big-data-achievement-2030-agenda

30 Toolbox. (Aug. 31, 2021). Cloud vs. On-Premise Comparison: Key Differences and Similarities. Spice Works. Accessible at: https://www.spiceworks.com/tech/cloud/articles/cloud-vs-on-premise-comparison-key-differences-and-similarities/

31 Google. Introduction to Cloud TPU. Accessible at: https://cloud.google.com/tpu/docs/intro-to-tpu

32 Wikipedia. (2023). List of cognitive biases. Accessible at: https://en.wikipedia.org/wiki/List_of_cognitive_biases

33 Andrew Burt. (Dec. 13, 2019). The AI Transparency Paradox. Harvard Business Review. Accessible at https://hbr.org/2019/12/the-ai-transparency-paradox.

34 InterpretML. (2023). Understand Models. Build Responsibly. Accessible at: https://interpret.ml/

35 Welcome to the SHAP documentation—SHAP latest documentation. Accessible at: https://shap.readthedocs.io/en/latest/index.html

36 Marco Tulio. (Aug 9 2016). arXiv,"Why should I trust you?": Explaining the Predictions of Any classifier. Accessible at https://arxiv.org/abs/1602.04938

37 Sachin Grover, Chiara Pulice, Geraldo Simari and V.S. Subrahmanian. (2019). Beef: Balanced english explanations of forecasts. IEEE Transactions on Computational Social Systems, 6(2), pp.350-364. Accessible at: https://www.scholars.northwestern.edu/en/publications/beef-balanced-english-explanations-of-forecasts

38 Harmanpreet Kaur, et al. (2020). Interpreting interpretability: Understanding data scientists' use of interpretability tools for machine learning. CHI 2020 Paper. Accessible at http://www-personal.umich.edu/~harmank/Papers/CHI2020_Interpretability.pdf

39 Sebastian Lapuschkin, Alexander Binder, Gregoire Montavon, Klaus-Robert Muller, Wojciech Samek. (2016). [IEEE 2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR) - Las Vegas, NV, USA (2016.6.27-2016.6.30)] 2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR) - Analyzing Classifiers: Fisher Vectors and Deep Neural Networks. 2912–2920. doi:10.1109/CVPR.2016.318

40 Sebastian Lapuschkin, Alexander Binder, Gregoire Montavon, Klaus-Robert Muller, Wojciech Samek. (2016). [IEEE 2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR) - Las Vegas, NV, USA (2016.6.27-2016.6.30)] 2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR) - Analyzing Classifiers: Fisher Vectors and Deep Neural Networks. 2912–2920. doi:10.1109/CVPR.2016.318

41 Kunihiko Fukushima. (1980). Neocognitron: A Self-organizing Neural Network Model for a Mechanism of Pattern Recognition Unaffected by Shift in Position. Biological Cybernetics. 36(4), 193–202. Accessible at https://link.springer.com/article/10.1007/BF00344251.

42 M. Mitchell Waldrop. (Jan. 22, 2019). What are the limits of deep learning? PNAS. Accessible at: https://www.pnas.org/doi/10.1073/pnas.1821594116

43 UNOCT-UNCCT & UNICRI. (2021). Algorithms and Terrorism: The Malicious use of Artificial Intelligence for terrorist purposes.

44 Ashish Vaswani, Noam Shazeer, Niki Parmar, et al. (2017). Attention Is All You Need. Accessible at: https://arxiv.org/abs/1706.03762

45 Chaitanya K. Joshi. (2021). Transformers are Graph Neural Networks. Accessible at https://www.chaitjo.com/post/transformers-are-gnns/

46 Stability. (n.d.). Stability AI. Accessible at https://stability.ai

47 Michael Stephenson. (2023). Exploring Midjourney AI Art Generator: A Guide to Stable Diffusion Generative Creation. Medium. Accessible at https://ai.plainenglish.io/exploring-midjourney-ai-art-generator-a-guide-to-stable-diffusion-generative-creation-7480c60d14be

How to cite this publication: UNICRI and INTERPOL. (Revised February 2024). Toolkit for Responsible AI Innovation in Law Enforcement: **Technical Reference Book.**

2025 Update: This concerns a correction to minor errors contained in the previous version.

www.interpol.int
www.unicri.it

INTERPOL_HQ

@INTERPOL_HQ
@UNICRI

INTERPOL HQ
UNICRI

INTERPOL
UNICRI

@INTERPOL
@UNICRIHQ

www.ai-lawenforcement.org