



**RESPONSIBLE
AI INNOVATION IN
LAW ENFORCEMENT**
AI Toolkit

Introduction to Responsible AI Innovation



Funded by
the European Union



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 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 in which this publication is used or quoted.

OVERVIEW



What

This document provides a basic overview of what responsible AI innovation means and why it is particularly important in the context of law enforcement. To that end, it explains the basic technical terms used in the AI Toolkit and lays out some of the challenges and opportunities raised by the integration of AI systems into law enforcement. It also provides responses to common questions related to AI systems and their characteristics, and introduces and defines fundamental concepts within ethics and human rights law that offer the basis for the **Principles for Responsible AI Innovation**.

When

This document can serve as an introduction to the topic of responsible AI innovation in law enforcement. It can also be consulted as and when necessary to seek clarification of unfamiliar terms and theories or simply to learn more about the rationale behind its approach to responsible AI innovation.

Who

This document is relevant to any individual or team in a law enforcement agency who is interested in learning more about the importance of a responsible approach to AI innovation and wants to understand the basis for such an approach.

TABLE of CONTENTS



| | |
|--|-----------|
| 1. Understanding the basics of AI innovation in Law Enforcement | 6 |
| 2. The Need for Responsible AI innovation in Law Enforcement | 18 |
| 3. How to carry out AI innovation responsibly | 23 |

| | |
|-----------------------------------|-----------|
| Annex: Want to Learn More? | 33 |
| References | 41 |



1.

Understanding the Basics of AI Innovation in Law Enforcement

In the AI Toolkit, the term **“AI innovation”** is used to refer to the wide range of activities that law enforcement agencies undertake when implementing AI systems in their work. This includes all stages of the AI life cycle, from planning to deployment, use and monitoring, and anything else it may involve.

To fully leverage the benefits of AI innovation, law enforcement agencies need common definitions of key terms and an understanding of basic concepts such as AI and AI systems and how they can be applied in their context. This section covers these topics in a concise and simplified way that is especially aimed at law enforcement agencies and personnel who are new to AI innovation. The aim is to help them to overcome any apprehension and to take their first steps in this area with confidence.

- *FOR A CLOSER LOOK AT THESE AND OTHER TECHNICAL CONCEPTS, SEE THE TECHNICAL REFERENCE BOOK.*

Key Concepts at a Glance

Artificial Intelligence

There is no universally accepted definition of AI but in the AI Toolkit, the term “AI” is used to refer to the field of computer science dedicated to studying and creating technological systems that can imitate human abilities such as visual perception, decision-making, and problem-solving. The products of this field – the technological systems – are called AI systems.

AI Systems

An AI system is essentially a computer system using AI algorithms to achieve specific goals with a certain degree of autonomy. There are different types of AI systems, but all are a combination of *software* and *hardware* built to produce outputs based on the inputs they are given.

More specifically, an AI system is comprised of an *AI algorithm* – the software - and a *computer* where the algorithm is processed – the hardware. There are many 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.

In very simple terms, these are algorithms that learn from *data*. This means most AI systems today are a combination of *software*, *hardware* and *data*. The figure below illustrates these three components:

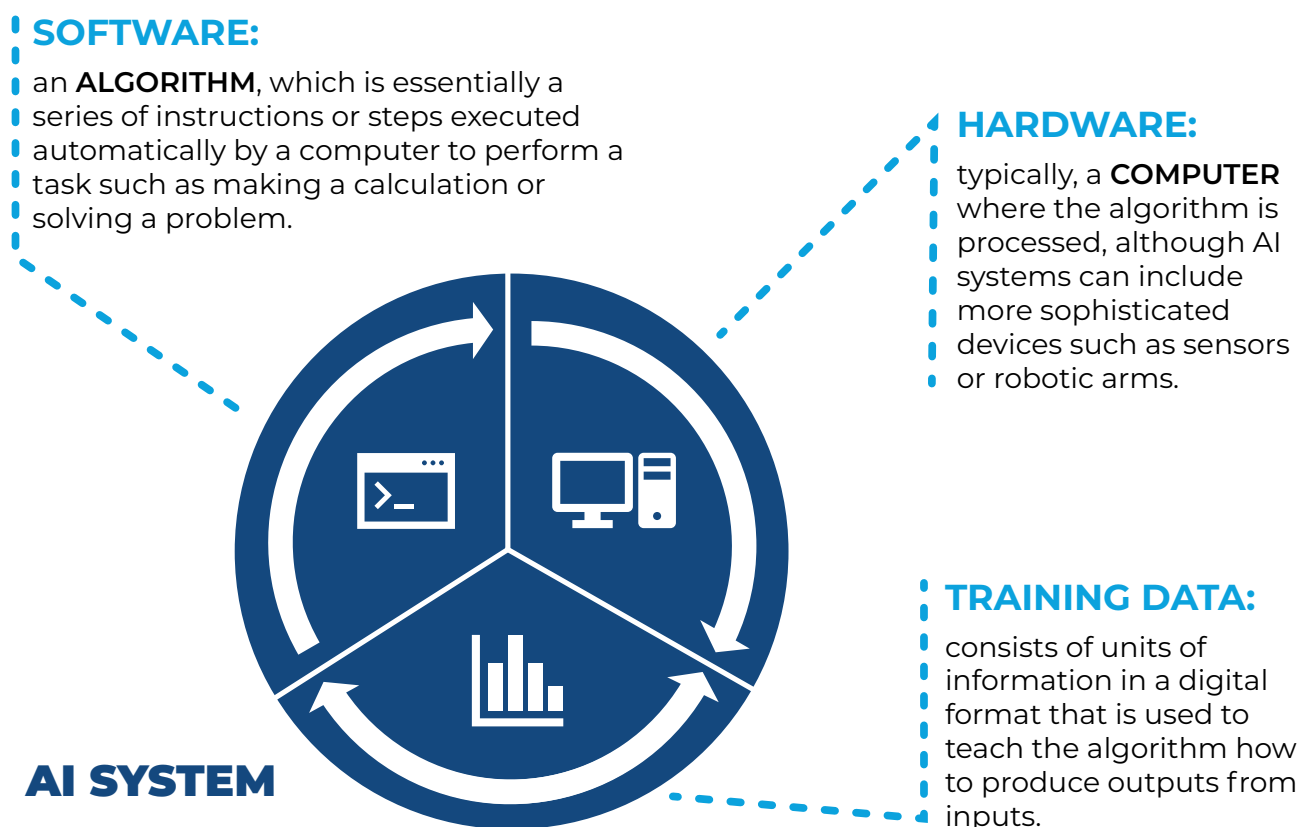


Figure 1 - The Components of an AI System

Although AI systems function with a certain degree of autonomy, they are not independent of humans. On the contrary, humans are an essential element. For example, the data may be collected by humans and/or refer to humans and their behaviour. AI systems are currently developed by humans, who design and build the algorithms and hardware devices and gather the data that is used to train the algorithm. In law enforcement applications, the AI systems' results are followed by either human validation or interpretation, or a human action. Ultimately, it is humans who use the outputs of AI systems and who are affected by them – whether positively or negatively.

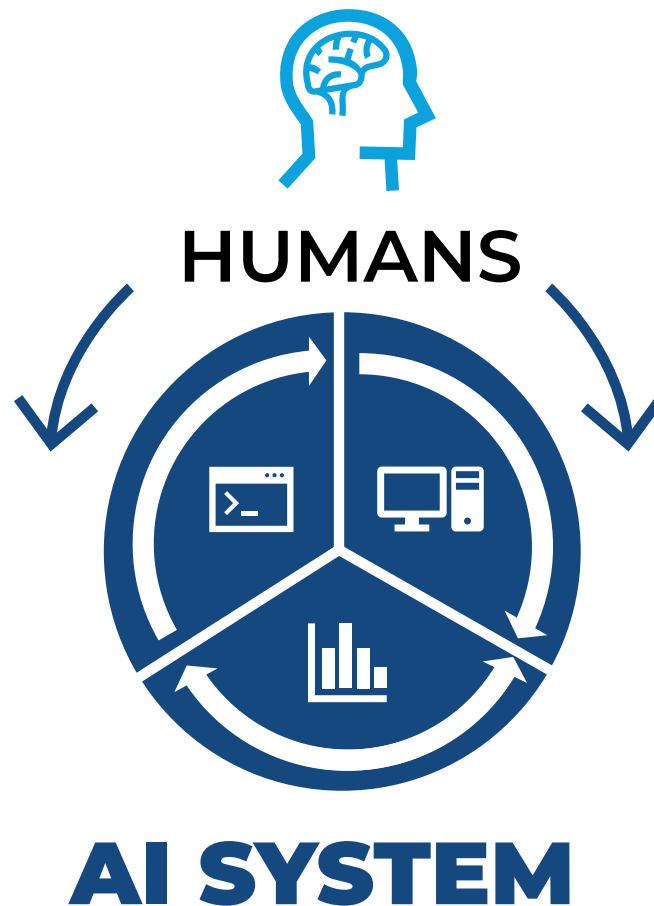


Figure 2 - The role of the human as a fourth element of AI systems

- LEARN MORE ABOUT THE FOUR COMPONENTS OF AI SYSTEMS IN THE [TECHNICAL REFERENCE BOOK](#).

Inputs and Outputs

The *algorithm*, the *computer* and the *training data* are the essential elements to produce a learning model which will then generate outputs based on the input data. These outputs from the AI system can take many forms such as actions, estimates, recommendations, or new content. Typically, the results are produced based on *input data*, which is the information that is fed into the AI system when it is used. This input data can come from various sources – for instance, databases of words, pictures or sounds, sensors, or the actions of the people using the AI system.



Figure 3 - How an AI system works

AI Tools and AI Technology

There are many terms used in the realm of AI to refer to AI systems and related products. In the AI Toolkit, the term ‘AI system’ is most often used as it captures the interaction of software, hardware, and data that makes these products so uniquely powerful.

Other terms that law enforcement agencies may encounter frequently are ‘AI tools’ or ‘AI technology’. These terms do not have universal definitions, but the term ‘AI tool’ is often used to refer to off-the-shelf software programmes that use AI systems. The term ‘AI technology’ is an umbrella term that commonly refers to the application of scientific knowledge of the AI field to practical purposes, such as computer vision.

Machine Learning

Machine learning is a subfield of AI that involves the use of algorithms that learn from examples and not from specific human instructions. When people talk about AI systems, they are usually referring to systems which include machine learning algorithms.

► [LEARN MORE ABOUT MACHINE LEARNING IN THE TECHNICAL REFERENCE BOOK.](#)

Deep Learning and Neural Networks

Deep learning is a subfield of machine learning which focuses on a particular type of machine learning algorithms: neural networks. A neural network is a very powerful type of algorithm whose structure resembles a network of nerve cells. They are more complex than other machine learning algorithms and require higher amounts of training data, but they have been behind some of the most impressive developments in the field of AI.

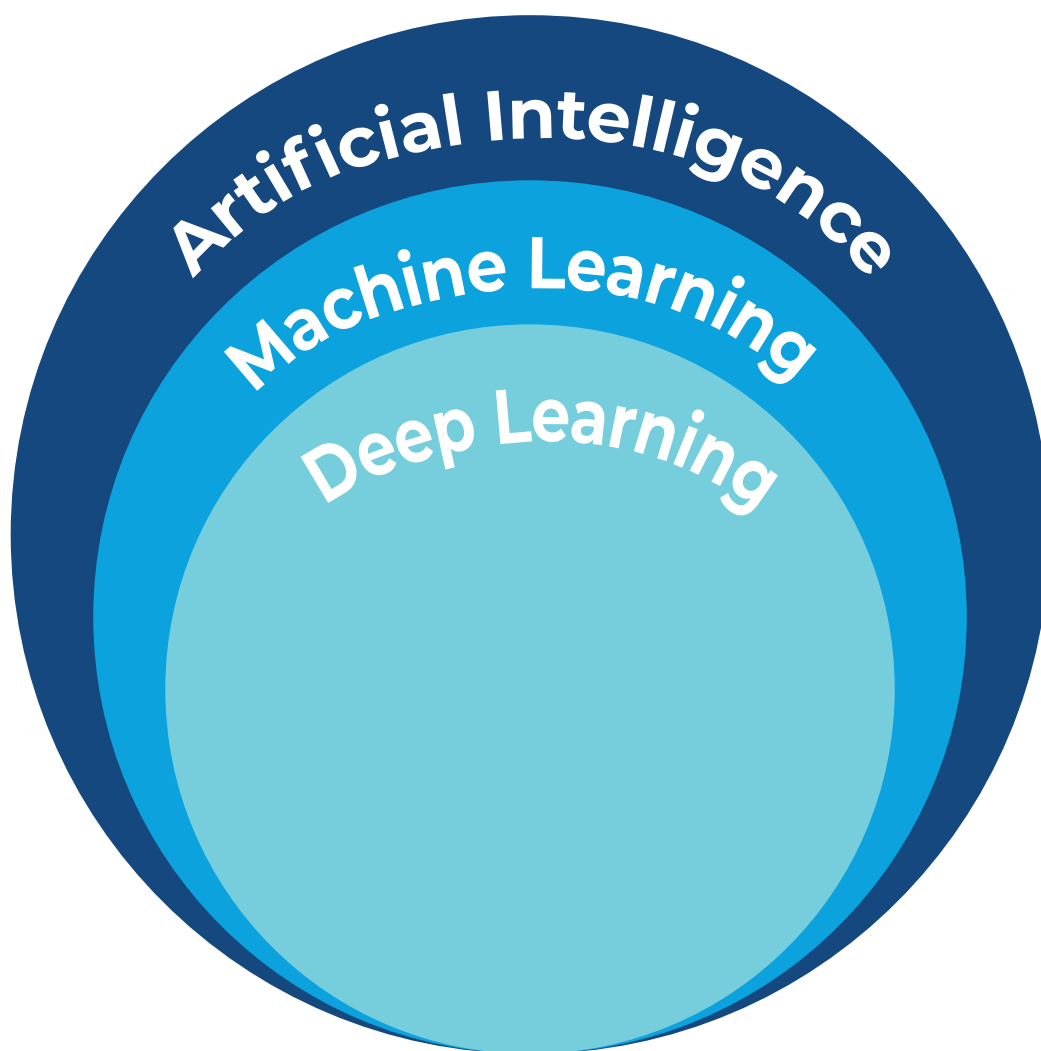


Figure 4 - The relationship between AI, ML and DL

Common Applications of AI in Law Enforcement

AI systems are being implemented in law enforcement for a variety of purposes. These highly versatile systems can be applied in different fields to achieve a wide range of objectives. With the rapid development of AI, algorithms are becoming increasingly specialized and increasingly capable of processing different kinds of data and producing specific types of outputs.

This rapid evolution can also be seen in the law enforcement context. The use of AI in law enforcement is continually on the rise, and law enforcement is making good use of the new types of AI systems and tools that are being developed on a regular basis. Nonetheless, it is possible to classify the most common applications of AI systems in law enforcement according to their main purpose. Currently, AI systems are most frequently applied in law enforcement to:

- analyze images
- analyze text and speech
- assess risks or determine probabilities
- generate content
- optimize processes and automate workflows

These applications are explained in simple terms below to get a sense of the capabilities and limitations of the types of AI systems currently most frequently used in the law enforcement context.

Image Analysis

AI systems can be utilized to analyze very large data sets of photos, videos and other visual information and automatically recognize, classify, and contextualize an image or elements within that image. The AI systems that detect and recognize elements in an image are called **object recognition systems**. Object recognition systems include machine learning algorithms that are built to process pictures, identify geometrical shapes and, ultimately, recognize things, faces and other objects. In the field of AI, the term *object* refers to any identifiable element within an image, including people, animals or other things that would not be called “objects” in everyday language.

PRACTICAL EXAMPLE | How licence plate recognition systems are developed

Licence plate recognition systems are machine learning systems designed to identify numbers and letters contained in pictures. They include algorithms trained with pictures containing numbers and letters which are labelled with the correct identification of the characters depicted. After analyzing thousands of pairs of pictures and labels, the algorithm learns to distinguish characters by extracting rules from 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 licence plate, it can recognize the numbers and letters it contains.

Facial recognition systems – commonly called facial recognition technology or FRT – are a widely-used variation of object recognition systems. A facial recognition system recognizes or supports the identification of specific persons in photos, videos and other visual inputs. 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 technology

Several law enforcement agencies around the world use facial recognition technology in criminal investigations to support the identification of suspects, victims, missing persons, unknown dead bodies and even witnesses.

Law enforcement investigators use facial recognition software for two main purposes: *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 identify them; and *biometric verification*, which is a “one to one” (1 to 1) comparison of two images to verify someone’s identity against, for example, an ID.

Although this technology presents considerable opportunities for socially beneficial uses, mostly through enhanced verification and identification processes, it also creates unique challenges which have led to some negative reactions and distrust from the public. For example, the lack of diversity in training databases has led to facial recognition algorithms showing performance deficiencies based on demographic characteristics such as gender and race.

These performance¹ issues can have negative consequences that often disproportionately affect individuals belonging to disadvantaged groups. For instance, in 2018 an innocent individual from an ethnic minority community was arrested and held in custody as a result of being falsely identified as a suspect in a theft investigation in which facial recognition technology was used. This example² shows how important it is that experts at comparing faces review the outputs of facial recognition systems and avoid automation bias – blindly relying on results generated by automated systems. In fact, although AI systems developers have made efforts to train algorithms using more representative databases, automatic identification is prone to errors and the results should be reviewed by facial experts before any conclusions are drawn or action is taken. This will enable the risk of a false identification – which could lead to harmful consequences such as the wrongful arrest of an individual – to be kept to a minimum.

- [LEARN MORE ABOUT ACCURACY IN FACIAL RECOGNITION TECHNOLOGY IN THE PRINCIPLES FOR RESPONSIBLE AI INNOVATION.](#)

Text and Speech Analysis

AI systems also enable large data sets of text and audio recordings to be analyzed. This is often used to recognize, process, tag and extract meaningful information from text, speech and voice. Text processing is made possible by natural language processing (NLP) capabilities. NLP is a field within AI that combines linguistics and computer science, and which seeks to process and analyze large amounts of natural language data in the form of text or 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.

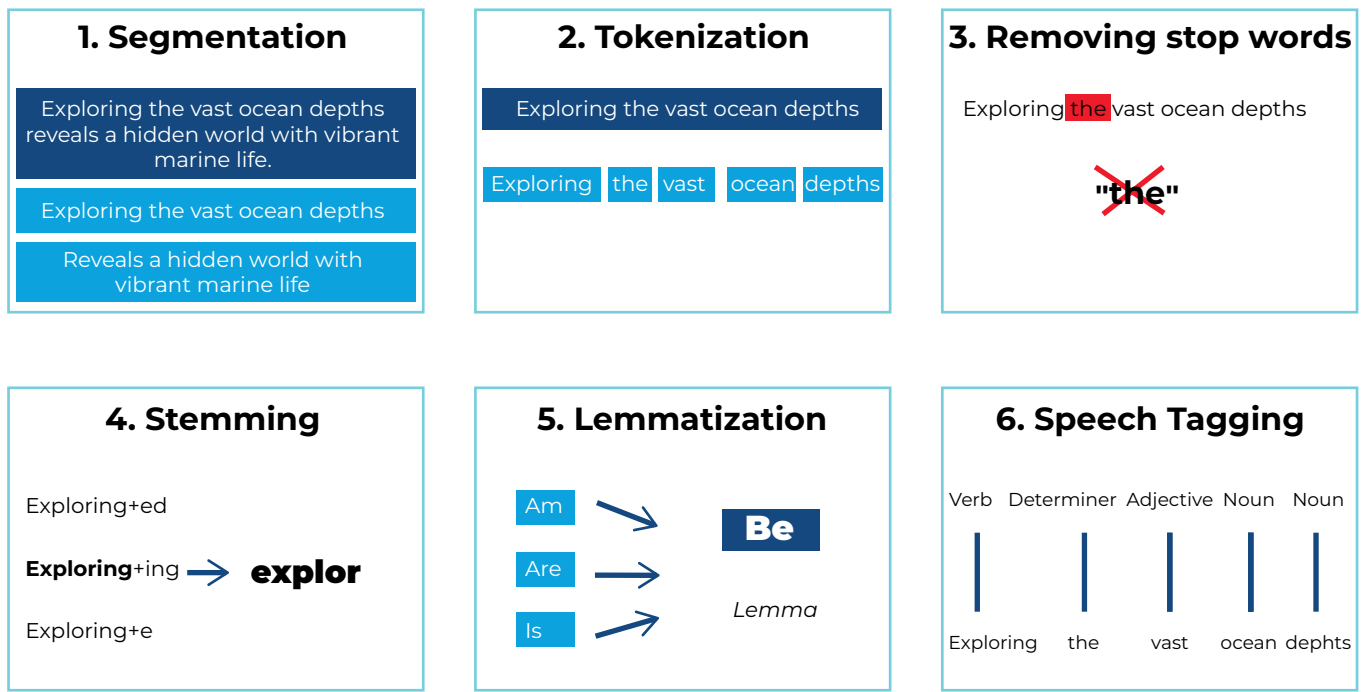


Figure 5 - Pre-processing techniques to clean and break down a text before feeding it into natural language processing models.

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 even predict the next word in a sentence.

Widely-used AI systems which use NLP include systems for speech recognition, translation and generation of natural language.

Risk Evaluation and Predictive Analytics

Pattern identification in very large data sets to recommend or trigger courses of action is another common application of AI systems in law enforcement. These systems use data, statistical algorithms and machine learning techniques to predict the likelihood of future outcomes based on historical data. The “predictions” of these AI systems are not a sure way of looking into the future, but rather extrapolations based on existing past data sets and the present context. In simple terms, these AI systems analyze what has happened in the past and infer likely future outcomes in related contexts. They are frequently used for risk evaluation or predictive analytics as a way of helping agencies with their decision-making processes.

Predictive policing systems are a prominent, although controversial, example of AI systems being used for these purposes. Its risks are explored in the next section. Other examples are AI systems that analyze crowd gatherings and predict the risk of accidents or crowd crushing.

Content Generation

AI systems enable the generation of new content such as images or text. For that purpose, large data sets are analyzed in order to extract patterns and rules and create content for specific contexts. In other words, these AI systems learn how to create new data based on the training data.

Content generation is the base of a widely-publicized phenomenon known as “deepfakes”. Deepfakes are a type of synthetic media which involve the use of machine learning techniques to manipulate or generate fake visual and audio content that humans or even technological solutions cannot immediately distinguish from authentic content.³

WANT TO LEARN MORE?

See the [“A more technical explanation of image generation”](#) section in the annex.

Systems enabling the generation of fake images as well as other content can be used for malicious purposes, including to create fake identities to commit crimes. For example, a general-purpose text generator chatbot, if not appropriately monitored or controlled, may be capable of reducing barriers to criminal activity by giving instructions on how to perform complex criminal activities such as building malware.⁴ However, content generation systems can also be used by law enforcement agencies to carry out their work, such as providing support for law enforcement agencies with undercover operations.

PRACTICAL EXAMPLE | Content generation for undercover operations

In undercover operations, content generation systems can be used to generate fake social media profiles and posts in order to gather information and evidence for an investigation.

For instance, law enforcement officers may create a fake online persona in order to infiltrate a criminal network, posing as a potential buyer or supplier of illegal goods or services. This could include using a deepfake image as a profile picture and a text generation system to create a realistic vendor profile including details of their services or products, as well as information about their background and interests that could help build a rapport with potential targets. The same text⁵ generation system could then be used to create realistic posts, comments, and messages that would make their online identity seem more convincing and trustworthy to targets.

By generating content that aligns with the interests and communication patterns of the criminal network, undercover officers can gain access to valuable information that could be used to build better cases against criminal networks.

Process Optimization and Workflow Automation

Process optimization and workflow automation is another area of implementation for AI systems. This is due to their ability to analyze large data sets and identify anomalies and patterns, predict outcomes, and suggest ways to optimize and automate specific workflows. In the law enforcement context this can be used to make connections between pieces of evidence by correlating certain events that occurred at the same time or place, activities carried out using the same devices, or other similarities.

PRACTICAL EXAMPLE | Responding to cyber-attacks

One common task in responding to a cyber-attack is identifying the source of the attack, the type of attack, and the extent of the damage. This often involves analyzing large amounts of data from various sources, such as network, system, and security event logs. Traditionally, this task would be done manually by security analysts, which can be time-consuming and prone to errors. This process can be optimized and automated to save time and improve accuracy.

For example, a machine learning algorithm can be trained on a data set of known cyber-attacks and normal network activity and can learn to identify patterns and characteristics that are commonly found in cyber-attacks. This model can then be applied to real-time network logs and other security event data to automatically identify potential cyber-attacks based on these patterns, which can help to improve the speed and accuracy of incident detection and response, leading to better protection of critical systems and data.

Furthermore, machine learning algorithms can be used to automate the process of prioritizing alerts and incidents based on their severity or potential impact. By using algorithms to analyze the data and assess the risk of each incident, security teams can prioritize their response efforts and focus on the most critical threats first. Machine learning can also be used to optimize workflows by automatically routing incidents to the appropriate teams or individuals based on their expertise, availability, or other criteria. This can help to streamline the incident response process and ensure that incidents are addressed in a timely and efficient manner.

2.

The Need for Responsible AI Innovation in Law Enforcement

As explained in the previous section, AI systems, and particularly those with machine learning algorithms, are very good at quickly analyzing vast quantities of information which has different origins and formats. They can be designed to perform a wide range of tasks based on information gathered through such a process. Applying these capabilities to law enforcement can have immense benefits, some of which are listed below:

- AI systems can improve the analysis of crime-related data and the detection, prevention and investigation of crimes.
- AI systems can carry out specific repetitive and mundane tasks much faster than any officer ever could. This frees up time for officers to concentrate on other tasks.
- AI systems can help safeguard the well-being of law enforcement officers by reducing their exposure to challenging material such as material related to child sexual abuse.

While their potential is undeniable, AI systems have limitations and can have negative consequences. As with any technology, AI systems are not inherently “good” or “bad”. A car, for instance, can be used for transportation or kidnapping – it is the human behind the wheel that makes the car’s use good or bad. A car can also be badly designed, malfunctioning, and lacking in safety equipment. Such a car, even if it is used with the best intentions, can cause harm because of the way it is designed. The same is true for AI systems. Much like cars, it is the way we, as humans, design and use AI systems that determines whether the outcome will be beneficial or harmful.

To maximize the benefits and minimize the risks associated with AI systems, law enforcement agencies need to take a *responsible* approach to AI innovation. **Responsible AI innovation consists of integrating AI systems into law enforcement work in ways that align with policing principles and which are ethically sound and human rights compliant.** This is a continuous process that requires an understanding of the limitations and risks of AI systems and the implementation of measures to avoid or sufficiently mitigate the negative consequences that can result from their implementation. Most importantly, this process should not be a cause

for concern for law enforcement agencies or personnel seeking to integrate AI system into their work. On the contrary, understanding the limitations and risks of AI systems' empowers individuals and organizations and enables them to move forward with confidence. Such self-assurance is essential in order to counterbalance the tendency people have to rely too heavily on results from automated systems such as AI systems – also known as *automation bias*.

► [LEARN MORE ABOUT COGNITIVE BIASES IN THE TECHNICAL REFERENCE BOOK.](#)

A responsible approach to AI innovation is crucial throughout the AI life cycle, and in all contexts where law enforcement agencies interact with AI. There is a cross-cutting need for responsible AI innovation in law enforcement, **firstly, due to certain characteristics of AI systems that demand increased attention and due diligence** as they may create or exacerbate severe or irreversible adverse impacts on individuals, society, and the environment if they are not understood and dealt with appropriately. These characteristics can be summarized as follows:

- **Value Embedding:** AI systems are susceptible to taking on certain human values which are present in the data used to train them or in certain development decisions. These human values may manifest in the outcomes of the AI systems, leading to the replication of bias and subjectivity from the humans involved in their creation.
- **Scalability:** AI systems can quickly process large amounts of data, allowing for a significant increase in the efficiency of certain tasks. However, this also means that flawed AI systems with inadequate training examples can have a wider impact than if a human was performing the same task.
- **System autonomy:** To a certain extent, AI systems are capable of making decisions without human oversight. This can lead to issues such as the potential for the system to make biased or unethical decisions that can have a direct impact in the real world, and a lack of accountability in the event of any issues.
- **Black Box problem:** Some AI systems are so complex that their inner workings or outputs cannot be understood by humans. It is difficult to interpret and trust the outputs of these so-called black box systems, which can be particularly problematic in applications where people's lives are affected.
- **Invisibility:** AI systems can impact people without their knowledge and understanding, as they often do not know when and where or for what purposes the systems are being used. This can result in accountability issues and can make it difficult for individuals to

contest outcomes if they are produced by errors in AI systems and they are unaware of the involvement of these systems.

As a result of these characteristics, especially when combined, any harm that derives from the use of AI systems can often be more far-reaching and less readily visible than that arising from other forms of technology. Because AI is a relatively new field that is developing rapidly, it is difficult for regulation to keep up. For that reason, anyone seeking to develop, procure, or use AI systems should pay close attention to their potential impact.

Secondly, the law enforcement setting also requires special attention. This is a context where the stakes are high. Given the unique competencies that law enforcement agencies have, the relationship between these agencies and the individuals and communities they serve is often one of power imbalance. For that reason, any wrongful or controversial use of AI systems can potentially have a severe impact on human rights, harm society at large, and undermine the law enforcement agencies' work.

The example below helps illustrate how the interplay of the characteristics of AI systems in the law enforcement context can create critical risks if not addressed carefully.

PRACTICAL EXAMPLE | The risks of predictive policing systems

Predictive policing systems are an example of AI systems applied to predictive analytics. They include algorithms that are trained with pre-existing crime data such as records of past detentions and convictions. Based on these data sets, these algorithms learn how to estimate when, where, and what crimes are likely to occur in certain areas.

The outputs of predictive policing systems can be used to inform decisions about how to police particular geographical areas, including the resources required and the type and nature of policing to be deployed.

The use of these types of algorithms in law enforcement has been controversial because of their susceptibility to reinforce and amplify the prejudices that have informed policing in the past.

Like any AI system, predictive policing systems are not value-neutral and their outputs may reflect the prejudices of law enforcement agencies and officers. The data used to train the algorithm may not accurately represent the current reality: if law enforcement officers are (or were in the past) more prone to policing certain neighbourhoods or arresting people that belong to a certain demographic, the data used to train the algorithm will reflect and reproduce that tendency, which may perpetuate discrimination against certain individuals or groups.

Another potential negative impact of using such systems is the inefficient distribution of police resources. If law enforcement agencies rely on these systems to decide, for instance, to patrol a certain area instead of another, inaccuracies in the systems may result in a waste of often scarce law enforcement resources. There is also evidence that over-policing certain areas may increase community tensions and contribute to an increase in crime. This means that using predictive policing systems could have the unintended consequence of increasing crime rates.

Individuals affected by the use of predictive policing systems may not be aware that such systems have been used, and law enforcement officers may not be aware that the decision to police certain areas has been made with the support of these systems. Even if the various stakeholders are aware of the use of an AI system, the fact that certain algorithms are black boxes can be a significant obstacle to those who wish to challenge the system's output, as it is difficult to understand and explain how such outputs have been reached.

Understanding the limitations and risks of predictive policing tools can help developers, law enforcement users, and society take measures to improve predictive policing tools and reduce the risks associated with them. For instance, developers can use more precise and less biased data to train the algorithms by focusing on victim data rather than just data on arrests. When developing an AI tool that aims to forecast the locations of shootings in a certain area, there is evidence to suggest that focusing on shooting locations leads to more accurate results than using data on arrest locations, as the two may be different and the shooting location data is less susceptible to other variables and potential human prejudices.

Law enforcement officers should also use their expertise and experience to look critically at the results of these AI systems, and understand that these tools do not predict the future, rather they calculate probabilities. This awareness can reduce the risk of automation bias. Another measure which can mitigate risks is to proactively inform or seek a dialogue with the public or those affected by the results of an AI system that has been used, and ensure that the AI system's results can be explained. This can empower stakeholders to better perform their respective roles with regards to responsible AI innovation and, if necessary, enable them to challenge the actions of law enforcement agencies that are based on AI system outputs.

While these measures are not exhaustive, they can help reduce the unintentional discriminatory use of AI systems and enable those who aim to prevent or fight discrimination to critically assess these systems.

3.

How to Carry Out AI Innovation Responsibly

Responsible AI innovation is a continuous process, not a set target. Carrying out AI innovation responsibly means adhering to principles of good policing, following and implementing AI ethics, and respecting human rights law, all with the aim of maximizing the benefits and minimizing the harms resulting from integrating AI systems into law enforcement. It requires involvement and commitment from all relevant stakeholders, as well as appropriate knowledge, structures, procedures, and organizational and technical measures to ensure that the highest standards for good governance, due diligence, and accountability are met. The AI Toolkit includes several resources that aim to guide law enforcement agencies throughout this process. The [Principles for Responsible AI Innovation](#) are the cornerstone of this guidance.

Because of the characteristics of AI systems and the particularities of the law enforcement context, diverse and complex ethical and human rights-related questions arise at each stage of the AI life cycle and regardless of the extent to which a specific agency is engaged with AI systems. The [Principles for Responsible AI Innovation](#) provide law enforcement agencies with a framework that can be used to navigate these issues.

More specifically, these principles help law enforcement – and secondary stakeholders such as external developers – to ensure that the ethical concerns and potential negative impact on human rights arising from AI systems are identified and eliminated or sufficiently mitigated at an early stage. To this end, they establish five core principles by which agencies can guide and evaluate their actions: Lawfulness, Minimization of Harm, Human Autonomy, Fairness and Good Governance.

As human rights law and AI ethics are fundamental pillars of the [Principles for Responsible AI Innovation](#), agencies should understand the essential concepts of ethics and human rights law as well as having a basic knowledge of AI.

Human Rights Law, Law Enforcement and AI

Human rights law imposes obligations on law enforcement agencies, officers and other personnel to protect and fulfil the human rights of individuals and refrain from violating any human rights in all their activities.⁶ While the obligation to pursue their mission in a manner which complies with human rights is not new in law enforcement, the introduction of AI systems adds a layer of risk that makes it even more important to stick to the law.

In fact, the characteristics of AI systems in the context of policing, as explained above, often lead to an interference with human rights. This is why it is fundamental to integrate human rights considerations in the use of AI systems by law enforcement.

- *LEARN MORE ABOUT THIS UNDER THE PRINCIPLE OF LAWFULNESS IN THE [PRINCIPLES FOR RESPONSIBLE AI INNOVATION](#).*

WANT TO LEARN MORE?

See the "[Restrictions and Derogations of Human Rights](#)" section in the annex.

Human rights law can accommodate the features and necessities of the law enforcement context. In certain circumstances and with due regard to a set of essential guarantees, the law allows agencies or officers within those agencies to act in a way that interferes with some of these human rights. This is the case for legitimate purposes such as upholding national security, public order and safety through the investigation and prevention of crimes or other law enforcement tasks that are necessary and proportionate, in order to avert grave harm.

What are Human Rights?

WANT TO LEARN MORE?

See the "[Different types of human rights](#)" section in the annex.

Human rights and freedoms are individual rights endowed on everyone irrespective of their background. They are not granted by States or by national laws but instead derive from the inherent dignity of each person. In other words, all individuals have rights and freedoms simply because they are human beings.⁷ Human rights are characterized as:

- *Universal*, meaning that everyone is equally entitled to them, independent of nationality, sex, age, national or ethnic origin, colour, religion, language, or any other status.

- *Inalienable*, meaning that their enjoyment should not ever be precluded aside from in exceptional and justified circumstances and following due process.
- *Indivisible and interdependent*, meaning that each right can only be fully enjoyed if the remaining rights are also ensured.⁸

There are different types of human rights, but all are equally important.

Human Rights Law and its Sources

International human rights law is based on the Universal Declaration of Human Rights, adopted by the United Nations General Assembly in 1948.⁹ International and regional human rights covenants or treaties are the main instruments which prescribe States' human rights obligations. States agree to be bound by these obligations when they ratify these treaties. There are nine core human rights treaties.¹⁰ All States have ratified at least one of these treaties, which means that all States are bound to human rights obligations.¹¹

To fully understand the importance of incorporating human rights principles into the various stages of AI, it is also useful to look at the question of ethics. This is because human rights are based on the theoretical framework of moral and political philosophy. In other words, ethics forms the basis for the universal justification of human rights. The core ideas of human rights – human dignity and human equality – are ideas formulated and promoted within moral and political philosophy, within the realm of ethics. This means that the theoretical foundations of ethics also provide the theoretical foundations for human rights.

AI Ethics

Ethics provides a body of knowledge with tools, frameworks, and theories to evaluate a given situation, and reason to determine the “right” action. Ethics aims to go beyond descriptive social norms or individual feelings and answer the question of “what is the right thing to do” in a systematic and analytical way where its justification can be understood, analyzed, and shared by any individual.

In the context of AI, ethics involves questions related to AI systems and their impact on individuals, groups of people, society at large and its democratic structures, and the environment, with the

WANT TO LEARN MORE?

See the [“Understanding Ethics through the historical example of women’s right to vote”](#) section in the annex.



goal of determining the best course of action in any given situation involving AI systems. Ethics is therefore at the core of the concept of responsible AI innovation.

AI ethics, as is the case for most principle-based frameworks in ethics today, is based on principlism – an ambitious framework that aims to combine all the main ideas from major philosophical theories under three core values: (1) human autonomy, (2) prevention of harm, and (3) social justice. These core values encompass the philosophical ideas around, among others, individual capability and the right to self-determination, the minimization of suffering, human and societal well-being, equal treatment, and fair distribution of burdens and benefits. The principles of human autonomy, minimization of harm and fairness are directly connected with these core ethical values.

AI's Characteristics and Responsible AI

How can principles help to tackle some of the challenges involved in the use of AI in law enforcement and maximize the potential benefits? The [Responsible AI in Action Workbook](#) can support agencies to translate the [Principles of Responsible AI Innovation](#) into concrete measures, as it sets out the questions that need to be asked and answered throughout the AI life cycle.

WANT TO LEARN MORE?

See the "[Philosophical Theories behind the Core Principles](#)" section in the annex.

This section helps agencies to better understand the principles and the workbook. It provides a closer look at the characteristics of AI systems and explains how the principles can help to directly mitigate the risks and balance out human automation bias. Although specific principles are mentioned in relation to particular features of AI systems, this does not imply that these principles alone can be used to address all the challenges involved. Each situation should be examined individually, and the principles should be considered as a whole in order to determine the most appropriate course of action.

Value Embedding

An AI system is typically composed of three main technical elements: the algorithm, the computer hardware, and the training data and each of these has an impact on the final performance of the system. However, it is useful to discuss the importance of humans as an additional central element in the system. Doing so requires a recognition of the human values which are also embedded in AI, and an awareness of the areas where they come into play.

There are two openings where human values can seep into the system. Firstly, through its training data, where the AI system learns to replicate the patterns in these data to apply to new inputs it will encounter in the real world. These data sets are compiled and curated by humans who may each view and measure the world through their own particular social and cultural lens, and they therefore inject some of their own values and perspectives about the world into the training data, which are often picked up by the AI systems in their learning process.

Secondly, there is the design and tuning of the algorithm, where human involvement is necessary. During this process, the developers adjust the algorithm and make judgments about which outputs are more appropriate than others. This process of judging and ranking some outputs over others is also intended to prompt the algorithm to replicate these preferred results. However, the kinds of judgments made often depend on the social and cultural views, values, and perspectives of the developers.

Through each of these processes, some element of human values is inevitably intrinsically embedded in the patterns which AI systems infer from the world, and results in them replicating any biases and subjectivities of the humans involved in their creation. The potential implications of these biases range from making AI systems inefficient, unreliable and unsafe, to violating human rights such as the right to non-discrimination.

Evaluating this through a principles-based framework involves acknowledging the potential presence of biases and developing appropriate safeguards to address them both during the data collection phase and the design and development phase. In this case, the core principles of *Lawfulness*, *Minimization of Harm* and *Fairness*, and all their underlying instrumental principles, are particularly useful in identifying the recommendations needed to mitigate the risks associated with embedded values.

WANT TO LEARN MORE?

See the [“A more technical explanation of how values are embedded in AI systems”](#) section in the annex.

COMMON QUESTION | How can AI be good or bad if it is just maths?

AI is a technological system and therefore cannot be “good” or “bad”. These are characteristics inherent to humans and not to objects. However, the decisions derived from the outputs of AI systems can impact society and individuals in a positive or negative way depending on how these systems are designed, deployed, and used. As such, AI systems are not value-neutral.

AI systems can have a positive impact on society by automating repetitive tasks, quickly processing vast amounts of data, providing personalized recommendations, and supporting more accurate decisions. For example, AI algorithms can process large amounts of text from chat conversations in a quicker and more efficient way than humans would be able to. This can speed up investigations and help find the relevant evidence to build a stronger investigation case.

Nevertheless, if not designed and used responsibly, AI systems can also have a negative impact. For instance, AI systems can perpetuate biases and discrimination if they are trained with biased data or designed with biased algorithms. Therefore, while AI algorithms are indeed just maths and are based on mathematical models, they are however trained and designed by humans who have values which will then be reflected in the AI outputs.

Scalability

In today's fast-paced law enforcement context, the speed of decision-making is a critical factor for efficiency. Unfortunately, humans are limited in terms of the pace of their decision-making processes by time and cognitive constraints. However, the scalability of AI systems allows them to overcome this limitation by a large margin.

The scalability of AI systems allows them to process and analyze vast amounts of data quickly and efficiently. For example, consider the task of detecting fraud. Humans may be able to manually detect a few dozen fraudulent cases per day, but an automated AI system can perform a few dozen detections per minute. Although this capacity can be extremely beneficial in the context of law enforcement activities, it is important to note that if the detection method is flawed and/or produces discriminatory outcomes, the use of AI systems means that it will impact many more people than if a human was performing the same task.

To avoid issues related to scalability, it is important to consider the core principles of *Minimization of Harm* and *Good Governance*. Following these recommendations ensures that the performance of AI models during training is carefully monitored and documented. If an AI system is adequately tested and monitored to guarantee its accuracy, safety and efficiency, the model will produce more reliable outcomes, which will in turn make its large-scale adoption much more beneficial and secure.

Additionally, introducing the principle of *Fairness* is important in terms of ensuring that the training data is truly representative of the task performed by a model, which helps to prevent

some of the adverse outcomes that can result from scalability, while maximizing the AI system's efficiency and accuracy.

System Autonomy

Many AI systems are designed to have some level of autonomy. Autonomy in this context means a system's capacity to take a series of decisions without human oversight. In some cases, depending on the environment in which the system is used, it may also carry out a series of actions associated with these decisions. Furthermore, the environment in which the system exists may vary widely in terms of its direct impact on humans and the physical world. Consider the following AI systems, which all display some degree of autonomy:

- Email spam filters: able to identify and sort mail autonomously without direct human oversight.
- Financial trading algorithms: able to identify trends and conduct hundreds of small trades within extremely short time frames independently of human oversight; however their behaviour is often dictated by certain pre-set rules.
- Uncrewed drones: able to automatically identify and potentially act upon suspected terrorist establishments; the environment in which this system functions includes the physical world, and it can have direct impact on a large number of individual lives.

Although both AI autonomy and automation are very useful and are a large part of what makes AI systems so efficient, it is important to remember that AI systems are only tools to help law enforcement, and should not be seen as substitutes for decision-making processes. This is particularly true for decisions which have an impact on people's lives. Unchecked systems may be prone to security breaches or produce discriminatory outcomes. This is why it is important to monitor and test AI systems appropriately and ensure that decisions taken in high-stakes contexts are ultimately made by humans.

In this regard, the core principles of *Human Autonomy and Good Governance* can provide guidance on how to ensure that human intervention is introduced appropriately. Adhering properly to these principles enables law enforcement agencies to fully benefit from the many advantages of autonomy and automation while avoiding any unforeseen risks.

WANT TO LEARN MORE?

See the [“Distinguishing System Autonomy from Automation”](#) section in the annex.



COMMON QUESTION | If humans make mistakes, what is wrong with using AI systems that are susceptible to error?

Yes, humans make mistakes. Criminal justice systems, for instance, are not perfect – judges, prosecutors, and law enforcement officers can err, and their mistakes can have negative consequences. Every human system is vulnerable to mistakes as it may be affected by multiple factors. For instance:

- Cognitive bias, such as implicit bias or in-group bias which can favour some individuals over others;
- Emotions such as fear, anger, or anxiety that can lead to hasty or irrational decisions, and emotions such as sympathy or empathy that can lead to biased decision-making;
- Inconsistency, particularly in cases where decisions are made by different people which can lead to unequal or unfair treatment of individuals or groups;
- Limited capacity to process information, which can lead to errors in decision-making processes, particularly when decisions are made under time pressure or with incomplete information.

If humans are susceptible to errors, why should we hold AI systems to a higher standard? Or why should AI decisions be more problematic than human decisions? There are important differences between human errors and AI errors. Mistakes deriving from AI systems can be especially problematic for the following reasons:

- The use of AI is often *invisible*: individuals, even those who interact with the AI systems in question, often do not know that they are using this technology or that they are subjected to it.
- Since humans hold certain values, when AI systems are trained with human-generated data these *values may end up being embedded* in the system outputs.
- If errors or biases are embedded in AI systems, these errors may be amplified, causing harm on a *faster and broader scale* than humans alone would.
- If AI systems are *black boxes* and their decisions cannot be fully explained, these decisions are difficult to contest.

These characteristics outline another important difference between AI systems' mistakes and human mistakes: as it is usually easier to detect errors made by humans and attribute them to the individual in question, there are more established systems in place to ensure humans are held accountable for those mistakes.

The Black Box Problem

When an AI system's inner workings cannot be understood by humans and/or its outputs are unexplainable to any human, it is considered a black box. A *black box* system is often defined as one which is too complex for any human to comprehend. This complexity arises from at least two aspects of the system: (1) the data on which the system is trained may be vast and highly varied, with a dataset that may contain hundreds of thousands of examples over hundreds of categories/features, and (2) the number of computations made using the input to arrive at any given output may be in the order of millions.

Not all machine learning systems are black boxes. However, neural networks are usually black boxes. They have become increasingly widely used in AI systems due to their huge potential, but the use of black box algorithms carries the risk of limiting humans' ability to apply a critical analysis to the outputs of AI systems.

To avoid or minimize the risks associated with the black box problem, it is important to consider the core principles of *Lawfulness*, *Human Autonomy* and *Good Governance*. Ensuring that these principles are respected involves making algorithms explainable whenever possible, and providing clear documentation regarding a system's inputs and outputs as well as its capabilities and limitations, especially when such systems are used in the context of criminal investigations. Adhering to these principles also involves ensuring human oversight during an AI system's design and testing phases, as well as regularly monitoring the model's performance and making adjustments as required.

Explainability practices also rely heavily on fostering collaboration between the relevant stakeholders. This may include encouraging cooperation between policy-makers, technology developers and end-users to ensure that systems are designed in a way that covers the needs of all stakeholders and aligns with ethical principles and human rights.

Invisibility

Organizations' use of AI systems tends to lack transparency for a variety of reasons, such as the complexity of the systems or the use of black box models, as well as legitimate or alleged proprietary or confidentiality reasons. As a result, AI systems often affect individuals in ways that are imperceptible to them. This issue is known as AI invisibility, and it derives from the fact that individuals frequently do not know when and where the AI systems that concern them are being used, even when these systems are collecting and analyzing their personal data. This is problematic as it makes it difficult for external parties to analyze and contest the use of AI systems and their results.

These shortcomings can be addressed by implementing the core principles of *Human Autonomy*, *Good Governance* and *Fairness* when using, or preparing to use, an AI system. Following these recommendations allows for human monitoring and oversight of the functioning of the AI system, thus enabling the relevant stakeholders to evaluate and participate in the system's design choices and assess whether they are accurate or not.

Adhering to the above-mentioned principles can further make data collection and processing *Lawful* as it ensures that breaches of privacy are limited by the instrumental principles of *Legitimacy*, *Necessity* and *Proportionality*. In practice, this entails employing privacy-enhancing technology when developing, procuring, deploying or using AI systems with intrusive potential, for instance. It also involves informing the public when and how an AI system is being used. By making the design choices and the use of AI public, individuals and organizations can provide feedback and recommendations on how to improve a system's performance.

COMMON QUESTION | Why should humans review the decisions of AI systems?

Although, in some cases, AI systems can make decisions that are more accurate, consistent, and objective than human decisions, there are situations where human decision-making is essential.

First, human judgement is necessary to identify and correct errors or biases in AI systems. If we cannot guarantee that AI systems are free of biases in the training data or during algorithm design, AI outputs need to be reviewed, particularly when it comes to sensitive and high-stake situations such as in the criminal justice environment.

Secondly, in complex situations and those that affect people's lives, humans are better equipped to understand the context or nuances of a situation and are able to make moral judgements and consider the ethical implications. Human review can help ensure that the final decision takes into account the relevant contextual factors such as individual circumstances, cultural differences, and social norms.

Thirdly, AI systems are unable to deal with uncertainty, unexpected events, or situations where there is incomplete data – or no data at all – about a real-life situation. In these circumstances, humans can rely on past experiences, prior knowledge, or even their intuition to deal with new situations.

Finally, humans are also better at creative problem-solving, which involves thinking outside the box and generating novel solutions to problems. AI systems are limited to the data they have been trained with and cannot come up with completely new, creative solutions.

Annex:

Want to Learn More?

1. 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.

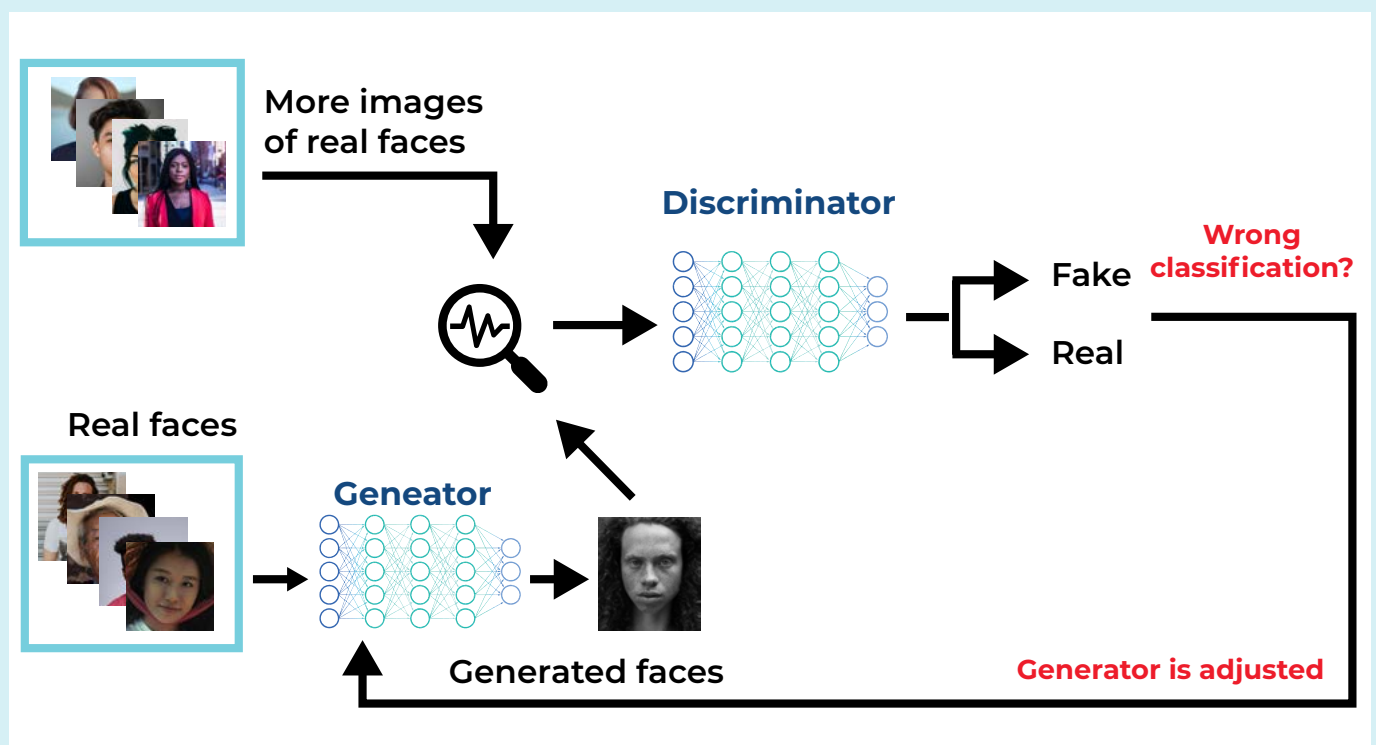


Figure 6 - Training method of generative adversarial networks.

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.¹²

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 2.¹³

2. A more Technical Explanation of How Values are Embedded in AI systems

The datasets that are used to train machine learning algorithms may contain value judgements. For example, in the case of a recidivism algorithm that analyzes data of previous crimes.

During the training, if there are two examples in the data with exactly the same features (same age, gender, education level, number of arrests, sentences etc.) but different values on the variable “race”, and these two examples have a different label (let’s say one committed a crime and the other not) the system will value “race” because it was the only difference between the two data points. If similar examples are repeated, even with small variations, the system will learn that the variable “race” plays a role in committing a crime. The system is not selecting “race” arbitrarily, or motivated by discriminatory beliefs, (it is just a piece of technology), the system is programmed to learn the best model that can represent the current data with its embedded human values.

During the learning process the internal parameters (weights) of the model are automatically adjusted so that the predicted label matches the real label, which in this case, would consist of increasing the weight associated with the feature “race”.

Also, for the system, characteristics such as “race” and “number of arrests” are not intrinsically different – both are just data points, represented by numbers with no intrinsic value associated. Humans are the ones who give more value to sensitive features. It is the developer that decides which features are better to characterize a certain type of task.

At the same time, some decisions by the developer can also carry value judgements that are then reflected in the system's outputs. For instance, the way developers of AI set a decision threshold is often imbued with values that will be reflected in the outputs of the AI system.

► *LEARN MORE ABOUT DECISION THRESHOLDS IN THE [TECHNICAL REFERENCE BOOK](#).*

As a result, AI systems may produce results that are perceived as value-neutral but in fact reflect specific values – regardless of whether the human developers intended or were even aware of it.

3. Restrictions and Derogations of Human Rights

Law enforcement and policing tasks may require intruding into people's private sphere, for example as part of body searches, to obtain blood or DNA samples for evidence, observation, surveillance of their movements and whereabouts, seizure of property, or dispersal of assemblies. Such measures interfere with human rights and must comply with a set of criteria in order to be lawful. Similarly, using AI systems as part of investigations or crime prevention measures may give rise to human rights compliance issues.

International human rights law recognizes that States may restrict a specific human right if it is in pursuit of a legitimate aim – for example, to uphold national security or public order – and if the restriction of human rights is deemed necessary and proportionate and is provided for in domestic law.

Restrictions or limitations to some human rights are allowed if they fulfil the following requirements:

- **Legality:** restrictions must be provided for in national law. This law must be in force at the time the restrictions are applied, and must be publicly accessible, clear and precise so that anyone can understand that such restrictions are allowed and under what circumstances.
- **Legitimate aim:** human rights treaties determine the purposes that allow for restrictions. They include, among others, public order, public health, national security, and the rights of others,
- **Necessity:** restrictions are only allowed if they are necessary to achieve the legitimate aim, and proportionate to that specific aim.

- **Proportionality:** restrictions must be the least intrusive means appropriate to pursuing such aims.
- **Non-discrimination:** restrictions must not be applied in a way that discriminates against certain individuals or groups based on their race, ethnic origin, religion, ability, gender, or other such characteristics.

Such restrictions must never be so extensive that they render the essence of the right meaningless. In other words, States must ensure that any human rights restriction is the exception, not the norm.

Under exceptional circumstances, such as in times of emergency that “threaten the life of the nation”, **derogations from human rights obligations** (in other words, the suspension of certain human rights) may be authorized by the State, but only to the extent strictly required in the situation. This only applies to so-called derogable rights. Human rights treaties specify which rights are derogable. For example, the rights to freedom of movement and freedom of expression are not absolute and can be subject to derogation in times of emergency. Conversely, derogations cannot lawfully be made to the right to life and the prohibition against torture.

4. Different Types of Human Rights

International human rights instruments contain sets of civil and political rights and economic, social, and cultural rights. The rights and freedoms of some groups – including women, members of ethnic minority groups, children, persons with disabilities, and migrants – are spelt out in relation to their specific challenges and needs. Other thematic conventions underline the need to protect basic rights such as the right to be free from torture, slavery, and human trafficking.

Thematic conventions addressing the impact of new technology such as AI on human rights or defining digital rights are yet to be adopted at an international or regional level, although recommendations and proposals have been issued to provide guidance for the development and application of algorithmic technology.¹⁴

Human rights can be divided into three main groups:

- **Non-derogable rights that must not be limited.** These rights include the right to life, the prohibition against torture and other inhumane or degrading treatment or punishment, the prohibition against slavery, the prohibition against imprisonment

merely on the grounds of inability to fulfil a contractual obligation, the prohibition against retrospective crimes and punishments, the right to be free from the retroactive application of criminal laws, the right to recognition as a person before the law, and the right to freedom of thought, conscience and religion.

- **Civil and political rights**, including respect for privacy, personal liberty, freedom of expression, freedom of assembly and association, freedom of thought and opinion, freedom of movement, and the right to a fair trial.
- **Economic, social, and cultural rights**, including the right to health, education, social security, and an adequate standard of living, access to work and labour rights, and the right to fair and just working conditions.

Also within these thematic areas are the principles of non-discrimination and equality that apply to the exercise of all human rights.

5. Understanding Ethics Through the Historical Example of Women's Right to Vote

Until the late 19th century, all around the world, women were not allowed to vote. National laws considered them at best as second-class citizens and at worst men's property (in some places, this still continues today). These laws were in line with the social norms which reflected most individuals' "moral feelings" at the time – written documents from this period often show that people considered the idea of women being equal to men as outrageously immoral. Any breach of these laws would therefore be met with a response from law enforcement. The question is, was this inequality and discrimination against women ever ethical? Was it ever "right" to treat women as a property?

This is a good illustration of the way laws, social norms, and individual "moral compasses" can diverge dramatically from ethics. In contrast to the prevailing argument at the time, it is now indisputable that women and men share the same "humanity" and thus, their autonomy and self-determination is as inviolable as that of any other person. Any infringement of their autonomy is and always has been ethically wrong. These arguments were already presented by prominent British philosophers Mary Wollstonecraft and John Stuart Mill long before the United Kingdom passed the laws that allowed women to vote in 1928. These philosophers were not writing based on their own opinions and preferences. Rather, they were making a universal

argument from an ethics standpoint, stating that the social norms and laws of the time were indeed unethical.

We can apply this historical illustration of ethics, laws, and social norms to our discussion on responsible AI innovation: an AI system that was fed data on voting from before the 20th century would probably learn that women and women's opinions do not count. The AI system itself does not hold a moral judgement about women, it simply learns from the social norms embedded within the data and replicates the same approach. If we had to use this data to understand society's priorities or needs, the voting data would simply not have the relevant information about half the population. Such a project would therefore not only be ethically wrong but also factually incorrect. Responsible AI innovation would require finding the relevant data for such a project and not proceeding with the project in the absence of such data. If the AI system were to decide who should vote based on pre-20th century data, it would result in dire discrimination.

6. Philosophical Theories Behind the Core Principles

Individual Autonomy and Kantian Ethics

The concept of autonomy, or individual self-determination, serves a critical role in all moral and political theories, but the theory that highlights it most is Kantian ethics. At the heart of Kantian ethics is the idea that one should never treat another individual as an instrument but rather should always acknowledge, respect, and even encourage their self-determination. Practically, this means that one should not lie, defraud, manipulate, or coerce another person even if it may result in a higher overall benefit. One's self-determination, one's right to decide for oneself, cannot be violated for the sake of others.

This Kantian approach to autonomy is echoed in the human rights framework. Human dignity and human freedom are rooted in autonomy, because this capacity for self-governance is what gives humans their moral agency and justifies their rights.

The Kantian emphasis on autonomy and human dignity is also echoed in the obligations imposed on States by international human rights instruments. At the core of these instruments is the obligation to respect the dignity of all human beings. With the adoption of the Universal Declaration of Human Rights in 1948, the foundation for human rights protection was defined by Article 1 which states that "all human beings are born free and equal in dignity and rights".

It is understood as safeguarding the self-determination and equal dignity of all humans, and is seen as a crucially important and cross-culturally unobjectionable normative premise.

Prevention of Harm and Utilitarianism

Utilitarianism is the philosophical theory that defines the right action or the right policy as the one that minimizes the harm and maximizes the benefits. According to this theory, to determine whether an action or a policy is “good”, we should only focus on its consequences and calculate its overall impact in terms of the suffering and happiness it brings.

Prevention or minimization of harm can also be seen as a more nuanced version of the well-known “first, do no harm” principle. In many cases, not taking any action in fear of causing harm may result in even bigger harm or loss of benefits. For example, in the case of an armed kidnapping, any action that the police take may result in harm. However, inaction might result in even bigger harm. Therefore, instead of a strict “do no harm” policy, a more nuanced utilitarian guideline would be to calculate the risks and benefits of each option and choose the one that would result in the least foreseeable harm for everyone involved.

While utilitarianism, with its focus on the value of minimizing suffering and harm, has an undeniable appeal, it is also open to criticism. At the extreme, utilitarianism can justify actions that violate human autonomy or equality, if the consequence minimizes the overall suffering caused when we take into account everyone who is affected by these actions. In other words, minimization of harm is not always compatible with respecting individual autonomy; a logical consequence of having two theories that prioritize two different values.

Fairness and Theories of Justice

There are multiple theories of justice, which argue for competing ideas regarding fair distribution. For example, egalitarianism is based on the idea that equality is paramount: for something to be fair it must be equally distributed. One of the most influential theories of justice is Rawlsian theory, named after the philosopher John Rawls. Rather than equality, Rawlsian theory focuses on the well-being of the most vulnerable groups in society. For Rawls, inequality is only justified if it benefits those who are least well off. Fairness in this case means that these groups are protected.

In the context of AI, the fact that there are different definitions of fairness depending on which theory is used is particularly important, because it means that the “fairness” of an AI system cannot be defined and determined without first determining which definition of fairness is and should be applied. In other words, one cannot simply talk about an AI system being fair.

First one must establish in what particular way it is fair: is it fair because it protects the most vulnerable, or is it fair because it protects and promotes equal distribution? One could argue the merits of pursuing both aims, but we must accept that they may be mutually exclusive: sometimes it is not possible to treat everyone equally and protect the most vulnerable at the same time. The relevant question in AI ethics is therefore: what are the most appropriate criteria for establishing whether the use of AI is fair?

7. Distinguishing System Autonomy From Automation

Despite their intertwined nature, autonomy differs from *automation* in one crucial aspect. Automation refers specifically to the use of computational systems (algorithms and/or robots) to perform a task previously carried out by humans, for example, picking up items on an assembly line. However, automated systems are not necessarily autonomous, as they may be programmed with a strict set of rules for a specific environment (such as a factory assembly line) with no capacity for independent decision making. An autonomous system in this same factory scenario may have the capacity to make decisions about resource allocation and workflow organization, whereas a robotic arm is unable to carry out these tasks.

References

- 1 Patrick Grother, Mei Ngan and Kayee Hanaoka. (2019). Face Recognition Vendor Test (FRVT) Part 3: Demographic Effects. NISTIR 8280. Accessible at: <https://nvlpubs.nist.gov/nistpubs/ir/2019/nist.ir.8280.pdf>
- 2 Bobby Allyn. (2020). 'The Computer Got It Wrong': How Facial Recognition Led to False Arrest of Black Man. Accessible at: <https://www.npr.org/2020/06/24/882683463/the-computer-got-it-wrong-how-facial-recognition-led-to-a-false-arrest-in-michig>
- 3 Oscar Schwartz. (2018). You thought fake news was bad? Deep fakes are where truth goes to die. The Guardian. Accessible at: <https://www.theguardian.com/technology/2018/nov/12/deep-fakes-fake-news-truth>
- 4 Europol. (2023). ChatGPT - the impact of Large Language Models on Law Enforcement. Europol. Accessible at: <https://www.europol.europa.eu/publications-events/publications/chatgpt-impact-of-large-language-models-law-enforcement>
- 5 Europol. (2023). ChatGPT - the impact of Large Language Models on Law Enforcement. Europol. Accessible at <https://www.europol.europa.eu/publications-events/publications/chatgpt-impact-of-large-language-models-law-enforcement>
- 6 United Nations Office of the High Commissioner for Human Rights (OHCHR). (n.d.). What are human rights?. Accessible at <https://www.ohchr.org/en/what-are-human-rights>
- 7 Daniel Moeckli, Sangeeta Shah, Sandesh Sivakumaran, and David Harris (eds). (2017) International Human Rights Law. (3rd edition, OUP). p. 559.
- 8 United Nations Office of the High Commissioner for Human Rights (OHCHR). (n.d.). What are human rights?. Accessible at <https://www.ohchr.org/en/what-are-human-rights>
- 9 Accessible at: <https://www.un.org/en/about-us/universal-declaration-of-human-rights>
- 10 United Nations Office of the High Commissioner for Human Rights (OHCHR). (n.d.) The Core International Human Rights Instruments and their monitoring bodies. Accessible at <https://www.ohchr.org/en/core-international-human-rights-instruments-and-their-monitoring-bodies>
- 11 United Nations Office of the High Commissioner for Human Rights (OHCHR). (n.d.). What are human rights?. Accessible at <https://www.ohchr.org/en/what-are-human-rights>
- 12 Stability. (n.d.). Stability AI. Accessible at <https://stability.ai>
- 13 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>
- 14 For example: Council of Europe, Committee of Ministers, Recommendation on human rights impacts of algorithmic systems, adopted 8 April 2020, CM/Rec(2020)1 and the EU Proposal for a Regulation of Artificial Intelligence (AI Act) of 21 April 2021.

How to cite this publication: UNICRI and INTERPOL. (June 2023). Toolkit for Responsible AI Innovation in Law Enforcement: Introduction to Responsible AI Innovation

© United Nations Interregional Crime and Justice Research Institute (UNICRI), 2023

© International Criminal Police Organization (INTERPOL), 2023.



www.interpol.int
www.unicri.it



INTERPOL_HQ



@INTERPOL_HQ
@UNICRI



INTERPOL HQ
UNICRI



INTERPOL
UNICRI



@INTERPOL
@UNICRIHQ