



ILNAS

White Paper

ARTIFICIAL INTELLIGENCE

TECHNOLOGY, USE CASES AND APPLICATIONS,
TRUSTWORTHINESS AND TECHNICAL STANDARDIZATION

Version 1.1 - February 2021





White Paper

ARTIFICIAL INTELLIGENCE

TECHNOLOGY, USE CASES AND APPLICATIONS,
TRUSTWORTHINESS AND TECHNICAL STANDARDIZATION

Version 1.1 · February 2021

ILNAS

Institut Luxembourgeois de la
Normalisation, de l'Accréditation, de la
Sécurité et qualité des produits et services

 **ANEC**

Agence pour la Normalisation et
l'Economie de la Connaissance

Foreword

The attention given to Artificial Intelligence (AI) has been increasing worldwide, with numerous national AI strategies being published all over the world. Luxembourg is no exception: the strategic vision on AI's development in Luxembourg¹ was published in 2019 (along with the data-driven innovation strategy²). This document aligns with the European objectives of gaining a leadership position in the digital market in general and AI in particular, objectives that can be achieved by putting forward European values and placing the focus on trustworthiness and human centrality.

Technical standardization plays an important role in the adoption of technologies and in building trust, and as such is an important tool towards achieving these national and European goals. Within this frame of reference, the "Institut Luxembourgeois de la Normalisation, de l'Accréditation, de la Sécurité et qualité des produits et services" (ILNAS) leads the "Luxembourg Standardization Strategy 2020-2030"³, signed by the Minister of the Economy, which identifies the ICT sector as one of the most relevant for national economic growth, along with the Construction and Aerospace sectors. Directly linked, ILNAS has also developed the "Luxembourg's policy on ICT technical standardization 2020-2025"⁴, which it carries out with the support of the Economic Interest Group "Agence pour la Normalisation et l'économie de la Connaissance" (ANEC GIE – Standardization Department). This policy embraces Smart ICT technologies such as Artificial Intelligence, Blockchain, Cloud Computing, and the Internet of Things. It aims to promote and strengthen the use of technical standards by the national market, to reinforce the position of Luxembourg in the global ICT standardization landscape - particularly through a stronger involvement of national stakeholders in the relevant standardization technical committees - and to pursue the development of research and education programs in the Smart ICT standardization area.

In this context, among the most recent developments, ILNAS has created, through a fruitful collaboration with the University of Luxembourg, a new Master's degree "Technopreneurship: mastering smart ICT, standardisation and digital trust for enabling next generation of ICT solutions", that will start in 2021. This diploma will allow national stakeholders to gain familiarity with Smart Secure ICT technologies, notably from a standardization and Technopreneurship point of view, in order to seize the future business opportunities offered in this innovative area. ILNAS has also launched different research activities in the Smart Secure ICT domain, which are directly contributing to the success of its program of education about standardization. As a result of these research activities, various documents have been published. In collaboration with the University of Luxembourg, a White Paper "Data Protection and Privacy in Smart ICT"⁵ and three technical reports on the gaps between scientific research and technical standardization in Cloud Computing, Internet of Things and Big Data/Artificial Intelligence⁶ were produced in October 2018 and October 2019, respectively. More publications were realized with the support of the ANEC GIE in order to inform the market about technical standardization developments in Smart ICT, such as the "National Technical Standardization Report on the Internet of Things"⁷ and the White Papers on "Internet of Things"⁸, "Blockchain and Distributed Ledger Technologies"⁹ or "Digital Trust for Smart ICT"¹⁰.

1 <https://digital-luxembourg.public.lu/stories/luxembourgs-strategic-vision-ai>

2 <https://gouvernement.lu/dam-assets/fr/publications/rapport-etude-analyse/minist-economie/The-Data-driven-Innovation-Strategy.pdf>

3 <https://portail-qualite.public.lu/dam-assets/publications/normalisation/2020/strategie-normative-luxembourgeoise-2020-2030.pdf>

4 <https://portail-qualite.public.lu/dam-assets/publications/normalisation/2020/policy-on-ict-technical-standardization-2020-2025.pdf>

5 <https://portail-qualite.public.lu/dam-assets/publications/normalisation/2018/White-Paper-Data-Protection-Privacy-Smart-ICT-october-2018.pdf>

6 <https://portail-qualite.public.lu/dam-assets/publications/normalisation/2019/TR-Smart-ICT-Gap-Analysis-SR-TS-ILNAS-UL.pdf>

7 <https://portail-qualite.public.lu/dam-assets/publications/normalisation/2020/national-technical-standardization-report-iot-june-2020.pdf>

8 <https://portail-qualite.public.lu/dam-assets/publications/normalisation/2018/white-paper-iot-july-2018.pdf>

9 <https://portail-qualite.public.lu/dam-assets/publications/normalisation/2018/white-paper-blockchain-june-2018.pdf>

10 <https://portail-qualite.public.lu/dam-assets/publications/confiance-numerique/white-paper-digital-trust-september-2017.pdf>

The present white paper on Artificial Intelligence is part of this education and research effort. It is intended to inform the national market about relevant AI standardization activities in the context of the AI technology landscape, market overview and trustworthiness challenges. With this white paper, ILNAS aims to provide a general survey of the opportunities offered by AI, with a view towards encouraging the national market's future involvement in the standards development process, for the benefit of Luxembourg's economy.

Jean-Marie REIFF

Director

ILNAS

Jean-Philippe HUMBERT

Deputy Director

ILNAS

Acknowledgements

The working group involved in the preparation of this white paper is:

Name of the contributor	Institution/Organization
Mr. Jean-Marie REIFF	ILNAS
Dr. Jean-Philippe HUMBERT	ILNAS
Mr. Nicolas DOMENJOUR	ILNAS
Dr. Jean LANCRENON	ANEC GIE
Mrs. Leslie FOUQUERAY	ANEC GIE
Mrs. Natalia CASSAGNES	ANEC GIE

The working group appreciates the support it received from external participants and would like to thank all those who contributed, in different ways, to the development of this white paper:

- Mr. BOVY Philippe, KPMG Luxembourg SCOP
- Mr. CADY Vincent, Tarkett GDL SA
- Mr. DANN John, Ministry of State, Central Legislation Service (SCL)
- Mr. DELOGNE Christophe, Everis Spain SLU
- Mr. GINDT Max, Ministry of State, Department of Media, Telecommunications and Digital Policy (SCM)
- Dr. SABETZADEH Mehrdad, University of Luxembourg, Interdisciplinary Centre for Security, Reliability and Trust (SnT)
- Dr. TREFOIS Christophe, University of Luxembourg, Luxembourg Centre for Systems Biomedicine (LCSB)

Table of Contents

Foreword	4
Acknowledgements	7
Abbreviations	11
List of Figures	12
List of Tables	14
Introduction	17
Chapter 1 Story of Artificial Intelligence: from mid-20th century to nowadays	20
1. Introduction	21
2. Origins of Artificial Intelligence	23
3. Evolution of Artificial Intelligence	24
3.1. Early age of modern Artificial Intelligence	24
3.1.1. Work of Alan Turing	24
3.1.2. Neural networks	25
3.1.3. Systems using other learning techniques	25
3.1.4. Reasoning systems	25
3.1.5. First applications of AI	26
3.1.5.1. Chess and checkers	26
3.1.5.2. Machine translation	26
3.2. Formal establishment of Artificial Intelligence	27
3.2.1. Dartmouth workshop	27
3.2.2. High expectations and challenges of Artificial Intelligence in 50s-60s	27
3.3. Rise of expert systems	29
3.4. Progress in machine learning	31
4. Enablers of Artificial Intelligence	32
5. Definition of Artificial Intelligence	33
5.1. First definition	33
5.2. Modern definitions	33
5.3. Standard definitions	36
6. Standardization efforts in Artificial Intelligence	37
6.1. Establishment and work of ISO/IEC JTC 1/SC 42 Artificial Intelligence	37
6.2. Other relevant standardization activities	40
6.2.1. ITU-T	40
6.2.2. IEEE	40
6.2.3. ETSI	41
6.2.4. CEN-CENELEC	42
7. Conclusion	42
References	43
Chapter 2 Artificial Intelligence: technology overview	46
1. Introduction	47

2.	AI paradigm: overview	47
3.	Searching	49
3.1.	Search problem definition	49
3.2.	Search algorithms	50
3.2.1.	Generic search algorithm	50
3.2.2.	Uninformed search strategies	51
3.2.3.	Heuristic search	54
4.	Machine learning	55
4.1.	Introduction to supervised, unsupervised and reinforcement learning	55
4.2.	Tasks solved by ML	57
4.3.	ML process and components	58
4.3.1.	Supervised and unsupervised machine learning	58
4.3.2.	Reinforcement learning	61
4.4.	ML algorithms and their business application	62
4.4.1.	Linear regression	62
4.4.2.	Logistic regression	63
4.4.3.	Decision tree	64
4.4.4.	Random forest	65
4.4.5.	Naïve Bayes classifier	66
4.4.6.	Support Vector Machine (SVM)	67
4.4.7.	K-Nearest Neighbors (kNN)	68
4.4.8.	Linear and quadratic discriminant analysis	69
4.4.9.	K-means	70
4.4.10.	Principal component analysis	71
4.4.11.	Hidden Markov model (HMM)	72
4.4.12.	Neural networks	73
4.4.13.	Q-learning	74
4.4.14.	AlphaZero	74
5.	Logic-based AI: knowledge representation and reasoning	75
5.1.	Overview of logical paradigms	75
5.2.	Knowledge representation and interpretation	78
5.2.1.	Propositional logic	78
5.2.2.	First-order logic	79
5.2.3.	Building a knowledge database	81
5.3.	Reasoning	82
6.	Examples of specialized AI systems	84
7.	Conclusion	84
	References	85
Chapter 3	Artificial Intelligence: use cases and applications	88
1.	Introduction	89
2.	Global market overview	89
3.	European activities in AI	92
4.	AI applications and use cases	93
4.1.	Horizontal AI applications	93
4.2.	AI application domains	94
4.3.	AI use cases	97
4.3.1.	Healthcare	97
4.3.2.	Banking and finance	101

4.3.3.	Manufacturing	105
4.3.4.	Transportation and Automotive sector	108
4.3.5.	Government and public sector	110
4.3.6.	Retail	114
5.	Key barriers to AI adoption	115
	References	117
Chapter 4	Towards trustworthy Artificial Intelligence	118
1.	Introduction	119
2.	Layers of trust	119
3.	Infrastructure layer	120
3.1.	Infrastructure optimization	120
3.1.1.	New computing paradigms	120
3.1.2.	Hardware optimization	121
3.2.	Infrastructure security	122
3.2.1.	Hardware faults	122
3.2.2.	Hardware security and verification	123
4.	Software and System layer	124
4.1.	Quality of AI system	124
4.1.1.	Product quality	125
4.1.2.	Quality in use	127
4.2.	AI-specific attributes for trustworthy and ethical AI systems	128
4.2.1.	Security	128
4.2.2.	Privacy and data protection	131
4.2.3.	Robustness	134
4.2.4.	Safety	135
4.2.5.	Data dependency	137
4.2.6.	Bias and fairness	138
4.2.7.	Transparency, accountability and explainability	140
4.2.8.	Autonomy and controllability	142
5.	Organizational layer	143
5.1.	AI ethical framework	143
5.2.	AI Governance	144
5.3.	Legal framework and certification schemes	145
6.	Conclusion	146
	Annex: Standards for Trustworthy AI	147
	References	150
	Conclusions and outlook	153

Abbreviations

AGI	Artificial General Intelligence
AI	Artificial Intelligence
AI HLEG	High-Level Expert Group on Artificial Intelligence
API	Application Programming Interface
CEN	European Committee for Standardization (Comité européen de normalisation)
CENELEC	European Committee for Electrotechnical Standardization (Comité européen de normalisation en électronique et en électrotechnique)
CL	Common Logic
CPU	Central Processing Unit
ETSI	European Telecommunications Standards Institute
EU	European Union
GDP	Gross domestic product
GOFAI	Good Old-Fashioned Artificial Intelligence
GPS	General Problem Solver (in Chapter 1)
GPS	Global Positioning System (in Chapter 2)
GPU	Graphics Processing Unit
HCI	Human-Computer Interface
HMM	Hidden Markov model
ICT	Information and Communication Technology
IEC	International Electrotechnical Commission
IEEE	Institute of Electrical and Electronics Engineers
IoT	Internet of Things
ISG	Industry Specification Group
ISO	International Organization for Standardization
IT	Information Technology
ITU-T	International Telecommunication Union - Telecommunication standardization sector
JTC 1	Joint Technical Committee 1 (of ISO and IEC)
kNN	K-nearest neighbors
ML	Machine Learning
NLP	Natural Language Processing
OECD	Organisation for Economic Co-operation and Development
OWL	Web Ontology Language
PaaS	Platform-as-a-Service
RAM	Random Access Memory
RPA	Robotic Process Automation
SC	Subcommittee
SG	Study Group
SQuaRE	Systems and software Quality Requirements and Evaluation
SVM	Support Vector Machine
TLO	Top-Level Ontology
TPU	Tensor Processing Unit
TR	Technical Report
TS	Technical Specification
W3C	World Wide Web Consortium
WG	Working Group
WIPO	World Intellectual Property Organization

List of Figures

Chapter 1

Figure 1:	Evolution of AI	21
Figure 2:	Tasks and capabilities of AGI	22
Figure 3:	ICT ecosystem supporting the AI systems	33
Figure 4:	Defining the AI as an intelligent agent	35
Figure 5:	AI definition described in “MIT Technology Review”	35

Chapter 2

Figure 1:	Depiction of different levels of world representation	47
Figure 2:	AI Knowledge Map	48
Figure 3:	Examples of a graph and a tree	49
Figure 4:	Simple search problem definition	50
Figure 5:	Illustration of a generic graph search	51
Figure 6:	Order in which the nodes get explored using breadth-first search	52
Figure 7:	Order in which the nodes get explored using depth-first search	52
Figure 8:	Example of iterative deepening search on a binary tree	53
Figure 9:	Illustration of heuristic function	54
Figure 10:	Supervised learning	55
Figure 11:	Unsupervised learning	56
Figure 12:	Agent-environment interaction	56
Figure 13:	ML Categorization	57
Figure 14:	Model creation process - main steps	59
Figure 15:	Typical loss functions	60
Figure 16:	Linear regression	62
Figure 17:	Logistic function and logistic regression classifier	63
Figure 18:	Decision tree	64
Figure 19:	Random Forest	65
Figure 20:	Bayes Rule and Naïve Bayes Classifier	66
Figure 21:	Support Vector Machine	67
Figure 22:	K-nearest neighbour	68
Figure 23:	Linear and Quadratic Discriminant Analysis	69
Figure 24:	k-means with k=3	70
Figure 25:	Principal Component Analysis	71
Figure 26:	Hidden Markov Model	72
Figure 27:	Different types of neural networks,	73
Figure 28:	Common Logic and its relation to other logics	77
Figure 29:	Syntax in Propositional Logic	78
Figure 30:	Truth table for the interpretation of sentences in propositional logic	78
Figure 31:	Example of a simple Knowledge base	79
Figure 32:	Syntax of first-order logic	80
Figure 33:	Knowledge representation in first-order logic	81
Figure 34:	Combining ML and reasoning over a knowledge base to provide explainable solutions	81
Figure 35:	Examples of logical equivalence	82

Chapter 3

Figure 1:	AI revenue worldwide	89
Figure 2:	Business opportunities generated by AI	90
Figure 3:	Gains from AI by 2030 expressed in GDP %	91
Figure 4:	Value gains from AI	91
Figure 5:	Adoption of AI by different industries and their estimated future investments in AI	96
Figure 6:	Distribution of number of AI patent families over application domains from 1980s until 2016	97
Figure 7:	Key barriers to AI adoption in Europe	116

Chapter 4

Figure 1:	ML system design combining Edge and Cloud computing	121
Figure 2:	Product quality characteristics according to ISO/IEC 25010:2011	125
Figure 3:	Quality in use characteristics according to ISO/IEC 25010:2011	127
Figure 4:	EU legislation relevant to AI	145

List of Tables

Chapter 1

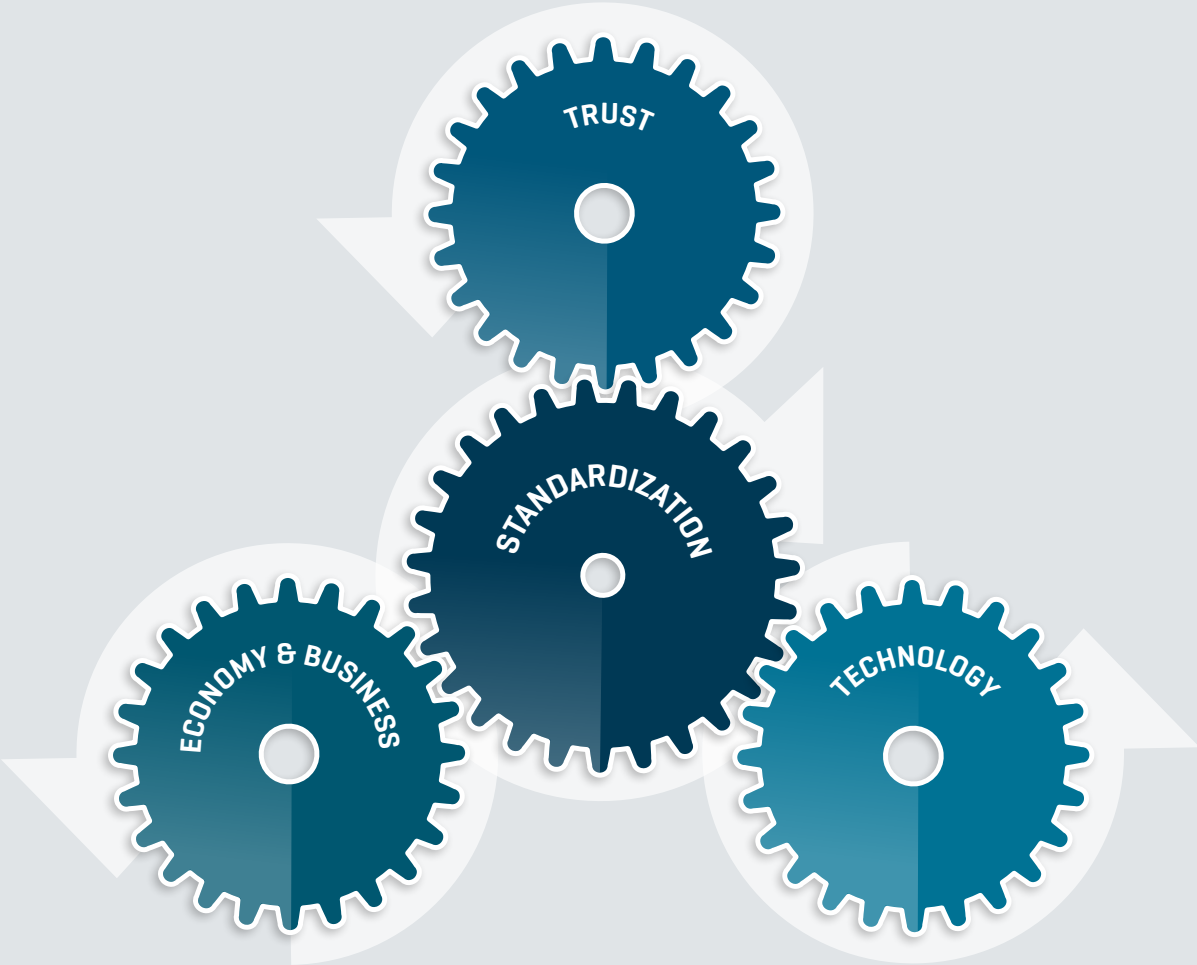
Table 1:	Projects of ISO/IEC JTC 1/SC 42 Artificial Intelligence	39
----------	---	----

Chapter 2

Table 1:	Mapping of ML algorithms to ML tasks	58
Table 2:	Confusion matrix	60
Table 3:	Statistical metrics for the evaluation of model performance	60
Table 4:	Linear regression	62
Table 5:	Logistic regression	63
Table 6:	Decision Tree	64
Table 7:	Random forest	65
Table 8:	Naïve Bayes Classifier	66
Table 9:	Support Vector Machine	67
Table 10:	K-nearest neighbour	68
Table 11:	Linear and quadratic discriminant analysis	69
Table 12:	k-means	70
Table 13:	Principal Component Analysis	71
Table 14:	Hidden Markov Model	72
Table 15:	Neural network	73
Table 16:	Reinforcement learning	74
Table 17:	AlphaZero	74

Chapter 4

Table 1:	Summary of defense measures against security attacks on AI	130
Table 2:	Description of privacy protection techniques in AI	133
Table 3:	Standards for Trustworthy AI	149



Introduction

Artificial Intelligence is progressively entering all areas of our lives. The technology has existed since 1956 but it is only now, with the development of other technologies, such as Cloud Computing and the Internet of Things (IoT), and the unprecedented volumes of data generated every minute in the world that it can realize its full potential. However, as with any technology, it does not come without risks and challenges. These challenges are technical, economic and societal; how to develop an AI system that would benefit rather than harm society?

In this context, the current white paper aims at providing the necessary background information to understand artificial intelligence and its challenges in an accessible way, to be able to take informed decisions. The document also introduces standardization activities related to artificial intelligence as one way of addressing both technical and economic challenges and building trust in the technology.

In this frame, the information provided in this document shall allow the readers to answer questions such as:

- What is the historical and technological context of AI?
- How do AI techniques work?
- What are the major challenges of achieving trustworthy AI, and how to address those challenges?
- What impact could AI have on the economy?
- How is AI applied in different sectors, and what are the national initiatives in these areas?
- What are the relevant AI technical standardization developments?

To achieve such outcomes, the rest of the white paper considers AI from three complementary perspectives - technology, economy and trust - while technical standardization perspective is addressed all along the chapters.

Introduction of AI concepts and technology:

- **Chapter 1** presents the evolution of artificial intelligence from its early days until today. To this end, it covers:
 - The developments that led to the establishment of AI as a branch of study in its own right,
 - The definition of AI,
 - Early research and application work on AI,
 - The modern technological context as an explanation of the current boom around AI, and
 - Existing standardization initiatives to support the development, deployment and acceptance of AI.
- **Chapter 2** provides an AI technical landscape. It introduces:
 - Three branches of AI techniques - searching, machine learning and reasoning - with a high-level explanation of how they work, and
 - The connections to the existing and under-development standards that support the development and validation of AI systems.

Economic and business perspective:

- **Chapter 3** aims first to provide insights on the potential economic impact of AI worldwide and in Europe, and second, to highlight a few major application domains of AI. For each application domain, the following information is provided:
 - A motivation of using AI in this domain,
 - Some top high-potential use cases,
 - Current applications that are fully implemented or being tested on the market,
 - The AI techniques mostly used for a given domain (in connection with Chapter 2),
 - The potential benefits of using AI,
 - The challenges to be addressed before fully reaping the rewards of using AI, and
 - Some standards - general purpose and domain-specific - that could help to address the challenges and accelerate the adoption of AI.

Building Trust in AI:

- **Chapter 4** paves a way towards trustworthy AI. It outlines the complex ecosystem of trust components that are present at different levels of AI implementation, and introduces the related challenges and mitigation measures. Standards are put forward as a helpful tool to achieve trust. They are therefore introduced all throughout the chapter to directly connect trustworthiness challenges with possible solutions. At the time of writing of this report, most of the standards for AI are still under development, leaving an excellent opportunity for the readers to join the standardization community and share their expertise, thus contributing to the European and international guidance towards trustworthy AI.

1

**Story of
Artificial
Intelligence:
from mid-20th
century to
nowadays**

1. Introduction

In early 2018, Gartner compared Artificial Intelligence (AI) to a hype itself¹¹. Indeed, multiple organizations look at the potential benefits of using AI and some of them already gather the fruits of AI-based products. It helps creating customized products and services or solving complex computational problems. It could be used in numerous application domains: healthcare, logistics, marketing, space discovery, etc. Hardly any domain could not benefit from AI.

AI has become popular in recent years, but the technology is not new and it has not always been a success. High expectations towards the technology in its early phase of development led to the disillusion and the cut of funding when the expected or promised results were not delivered. The new ways of approaching the problems brought new success stories and new deceptions. These deception periods are often called AI winters. Some authors considers only one winter period [1] separating symbolic, otherwise called “Good Old-Fashioned Artificial Intelligence” (GOF AI)¹², and connectionist approaches to AI. Others distinguish between two winters highlighting the three phases of AI development such as GOF AI, Expert Systems and Machine Learning (see Figure 1) [2]. It is important to note, however, that even during the winter, the work on AI never really stopped. Also, there is no clear separation between the phases of AI development. Even if machine learning became popular after 1990s, the description of the first neural networks was suggested by McCulloch and Pitts in as early as 1943 [3] [4] [5].

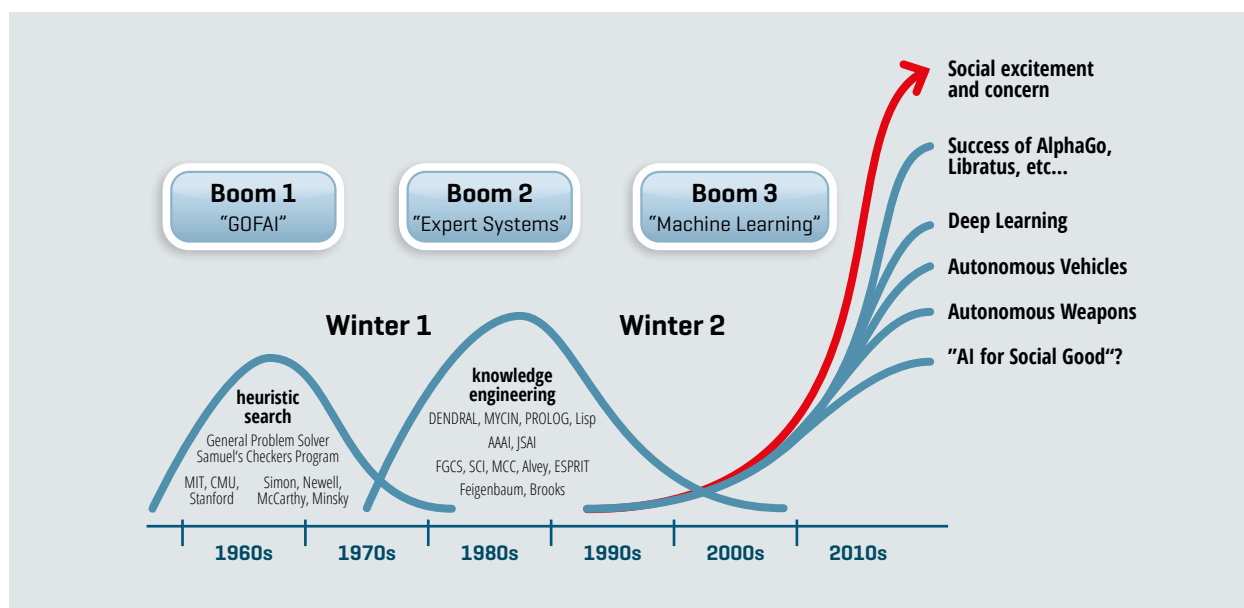


Figure 1: Evolution of AI [2]

An AI could be seen through the prism of the tasks that an intelligent system is supposed to be capable of. It could be a system that performs a wide range of tasks and is adaptable to new tasks and environments via generalization and learning transfer (without human intervention). Such a system is referred to as **Artificial General Intelligence (AGI)**. The learning capabilities and the tasks that an AGI should perform could be specified as in Figure 2. These tasks taken one by one could be studied separately and divided into the subtasks, forming a niche for the development of an efficient specialized AI system, such as for example image processing, speech recognition, knowledge representation or reasoning. An AI system specialized in one particular task is called **narrow AI** [6].

11 <https://www.gartner.com/doc/3883863/hype-cycle-artificial-intelligence->

12 https://en.wikipedia.org/wiki/Symbolic_artificial_intelligence

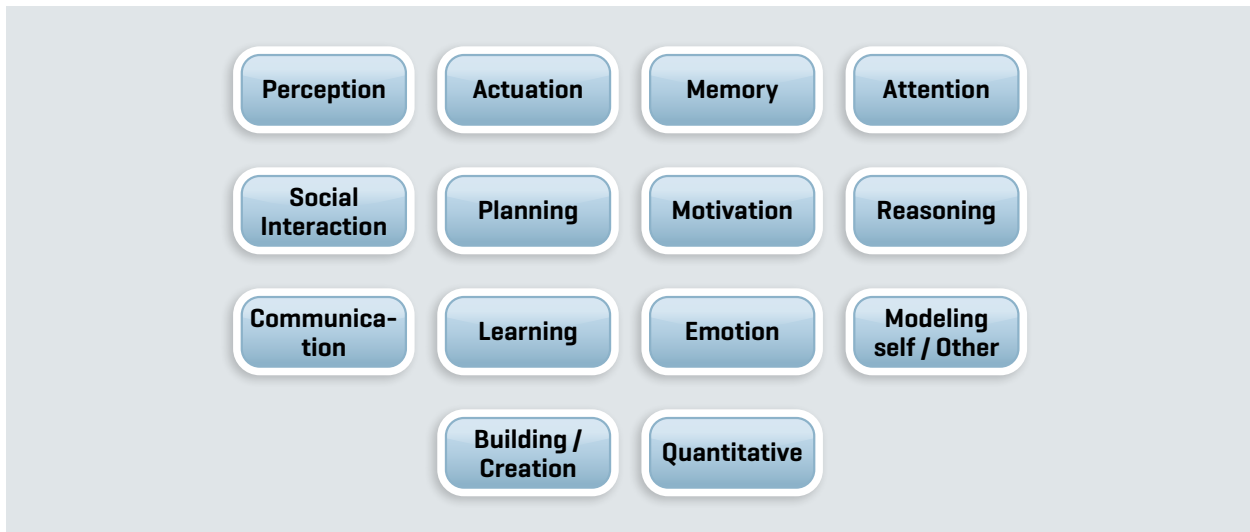


Figure 2: Tasks and capabilities of AGI (adapted from [6])

From the early stage, the research on AI has been driven by both work on specific tasks and on understanding and building general capabilities such as learning.

Current white paper provides an overview of the achievements in the field and presents the enabling factors that led to the boom of AI as we know it today. These developments set a background for the establishment of a standardization committee on AI whose objective is to provide international guidance for various stakeholders concerned by technology. For example, it will address the societal concerns about AI: bias, privacy issues, singularity, etc.

2. Origins of Artificial Intelligence

The term “*Artificial Intelligence*” was introduced in 1956 at the event that is currently referred to as Dartmouth Workshop. This first gathering to discuss the topic involved scientists from different backgrounds, such as computer science, mathematics, neurology, etc. [7]. Indeed, AI has its sources in multiple research fields and could be traced back to Ancient Greece [3] [4] [5].

The syllogistic logic of Aristotle, lingua characteristic of Leibniz, Boolean propositional calculus [4], first-order logic by Frege and Russell [3], and the invention of fuzzy logic by Lotfi Zadeh [5] contributed to the development of modern AI and particularly that of reasoning systems.

The incompleteness theorem of Gödel or NP-completeness theorem by Cook and Karp identify the limits of computation, while the notion of computability introduced by Church-Turing thesis give a general way to compute any computable function, which is crucial for running AI programs [3].

Theory of probability, statistical methods introduced by Bernoulli and Laplace, Bayes’ rule, etc. provide foundations for machine learning and dealing with uncertainties [3] [4].

Economic study of decision-making processes that maximize the pay-off gave rise to decision theory that is used in AI in the form of Markov decision processes. Another advancement in economics made by Simon, namely the one showing that humans tend to be satisfied with relatively easy achievable good decision rather than search for complex optimal solutions, could be compared with the usage of heuristics in AI [3].

Psychology and cognitive sciences, whose goal is to study mental processes and behavior, contributed to the definition of intelligence and understanding of human functioning to plan tasks and achieve goals on which rely the connectionist approach to AI [3] [4] [8]. One of the examples is the idea of reinforcement introduced by Skinner suggesting that a reward obtained for some behavior would make this behavior more likely to be reproduced in the future, the idea that inspired the reinforcement learning approach [4].

Cognitive sciences rely on the progress in neuroscience that tries to explain how the human brain functions. The idea to reproduce the functioning of neurons in human brain gave rise to the development of neural networks and deep learning [3] [4].

The development of computer science and modern computers made the modern AI possible by overcoming one of major problems of early AI: lack of computational power [3] [9].

Among other fields that inspires AI one could name biology and perception modeling, evolution theory and linguistics [4].

One can “*think of artificial intelligence as the scientific apotheosis of a venerable cultural tradition, the proper successor to golden girls and brazen heads, disreputable but visionary geniuses and crackpots, and fantastical laboratories during stormy November nights. Its heritage is singularly rich and varied, with legacies from myth and literature; philosophy and art; mathematics, science, and engineering; warfare, commerce, and even quackery*” [5].

When it comes to standardization, the story is similar. The dedicated technical committee for AI (ISO/IEC JTC 1/SC 42¹³) was created in 2017, but the work on relevant standards has started in other committees before that. Some examples of such work could be:

- Standardization related to software development, testing and quality by ISO/IEC JTC 1/SC 7 *Software and systems engineering*¹⁴,

¹³ <https://www.iso.org/committee/6794475.html>

¹⁴ <https://www.iso.org/committee/45086.html>

- Standardization related to big data by ISO/IEC JTC 1/WG 9 *Big Data* (now integrated in the committee on AI),
- Standardization of data representation and data quality by ISO/IEC JTC 1/SC 32 *Data management and interchange*¹⁵,
- Standardization of statistical methods for data analysis by ISO/TC 69 *Application of statistical methods*¹⁶,
- Study of the usage of compressed representation of neural networks for moving pictures by ISO/IEC JTC 1/SC 29 *Coding of audio, picture, multimedia and hypermedia information*¹⁷.

3. Evolution of Artificial Intelligence

Even if officially AI appeared in 1956, the work that is considered the “*gestation*” phase of modern AI started in the 1940s [2] [4]. The early work on AI was mainly done in the United States (US). Some European efforts also contributed to the development of the field. It also happened that the progress on the same topic was made in parallel by different researchers who were not aware of each other’s work due to the lack of communication [9]. This section presents non-exhaustive, yet influential, work on AI over the time.

3.1. Early age of modern Artificial Intelligence

3.1.1. Work of Alan Turing

Alan Turing is one of the key figures in modern computer science. His early paper “On computable numbers, with an application to the Entscheidungs problem” [10] defines the computing machine capable of computing any computable function, which is now called **Turing machine**. This work provided the basis for the theory of computation, underlying modern computer science in general and AI in particular.

Later, in 1950, Turing published a paper on “Computing machinery and intelligence” [11] that starts with a question “*Can machines think?*”. Such a question relates directly to the definition of AI. Instead of speculating about the definition of a thinking machine, Turing described an “*imitation game*” that is currently known as **Turing test**. He then replaces the original question with the following one: “*Are there imaginable digital computers which would do well in the imitation game?*”. The game presents a situation where there is one human interrogator and two responders, one human and one computer. The goal of interrogator is to determine who is who. The goal of computer is to trick the interrogator into thinking that it is a human, and the goal of the human respondent is to help the interrogator to make the correct identification. A thinking, or intelligent, machine would be the one that manages to convince the interrogator it is human.

The suggested test raised various objections – theological, mathematical, from continuity in the nervous system, etc. – to which Turing provided replies in his paper. As of today, researchers still argue about the plausibility of the test raising the two questions [1]:

- *Does imitating a human actually prove intelligence?*
- *Is intelligence possible without passing the Turing test?*

¹⁵ <https://www.iso.org/committee/45342.html>

¹⁶ <https://www.iso.org/committee/49742.html>

¹⁷ <https://www.iso.org/committee/45316.html>

However, the Turing test remains important for AI and a competition is organized annually to determine the most human-like computer program¹⁸. The competition will be closed after the victory of a program that judges would not be able to distinguish from a real human with the inputs necessitating deciphering and understanding text, visual, and auditory input.

The standardization community acknowledges the existence of this approach as one of the ways to define the systems that act like humans, but for the moment does not consider it for the proper definition of AI.

3.1.2. Neural networks

One of the early developments related to modern AI arises from the work of McCulloch and Pitts on **neural networks**. In 1943, the authors proposed a formal model of a neuron, having input, output and the main unit. They showed that networks of these formal “*neurons*”, where output of one neuron served as inputs for the connected neighboring neurons, could perform computational operations by switching on and off the relevant neurons in the network [3] [4].

Later, in 1949, Hebb suggested that the strength of connections between the neurons in the brain increased if it was frequently and successfully used. This so-called Hebb rule for updating the connection strength between neurons was actually observed in living animals [3] [4].

Based on these ideas, Clark and Farley described in 1955 an approach to pattern recognition based on the networks of formal neurons that became called neural networks. This was the first neural network to be run on the computer [4].

3.1.3. Systems using other learning techniques

In 1952, Oettinger described a computer program called Shopper that demonstrated the early successful implementation of **machine learning**. It was a simulation of shopping activity in a mall of eight shops. Upon the reception of first purchase instruction, the program would randomly visit the shops memorizing the items stored in them until it found the item sought. Next time, if the requested purchase item was the one that the program had seen before, it would go directly to the right shop [8].

In 1955, Selfridge described a technique for image **classification** based on significant **features** recognition. He suggested a model for extracting and counting those features in order to decide on the type of the image, for example count the number of corners to decide whether the image is a triangle or a square. This system had many limitations, for example, it could not recognize curvature or juxtaposition of singular points. Nevertheless, this work set the grounds for the future development of machine learning by recognizing that “*computer may improve its recognition by learning*” [4] [12].

Another interesting classification technique, now called **nearest neighbor**, was introduced in 1951 by Fix and Hodges. Given two sets of data points, the nearest neighbor method would attribute a new data point to the one having the closest to it data point [4].

3.1.4. Reasoning systems

The ideas behind the first **reasoning system**, Logic Theorist by Simon and Newell, were born in early 1950s, though the program itself was presented and published only in 1956. Since 1955, Simon and Newell began their collaboration with an objective to develop formal symbolic system that could simulate process of human thought¹⁹. Logic Theorist was a program for proving mathematical theorems similar to those in Russell and Whitehead's “Principia Mathematica”. By the end of 1955, Simon and Newell managed to implement some

¹⁸ https://en.wikipedia.org/wiki/Loebner_Prize

¹⁹ <https://history-computer.com/ModernComputer/Software/LogicTheorist.html>

heuristics for problem solving and simulate by hand the first proof. **Heuristic** is a practical method of finding a satisfactory solution to a problem, when reaching an optimal solution is impossible or impractical (for example, demanding high computational or time resources)²⁰. Later, using heuristic search became one of the key features in AI and gave rise to heuristic programming. With the help of Shaw, Simon and Newell later implemented the Logic Theorist on the computer [3] [4].

3.1.5. First applications of AI

Among the first practical applications of AI, come to mind the game playing programs, such as chess and checkers, and machine translation.

3.1.5.1. Chess and checkers

The first chess program in history could be attributed to Zeus, who is rather known for building the first general programmable modern computer. Zeus coded his chess program in 1945-1946 using the programming language Plankalkül that he designed specifically for the purpose. In a lack of implementation, this program was never run on a real machine [9].

This may be the reason why it is Shannon who is commonly acknowledged to develop the first chess program in 1950. In his ad-hoc paper, Shannon described two approaches to computer chess playing: Type A program and Type B program. The Type A considers in a given position all possible moves of the player and of the adversary in order to select the move that will optimize the chances to win using minimum-maximum search algorithm (that is minimizing the chances of the adversary and maximizing the own chances to win). The Type B considers the strategies used by the chess masters in order to reduce the number of variations to consider. The former is now referred to as "*brute force*" and the latter could be called "*strategic*". According to Shannon, the optimal strategy should take into account the capacities (computational speed) and the weakness (lack of analytical ability) of the computer. At the time, the Type B was considered more promising due to the limitations in computational power. However, the Deep Blue that defeated world's champion Kasparov in 1997 rather belonged to the Type A [1].

Another effort to design a chess playing program goes to Newell. In his 1955 paper, he suggested that the only way for a computer to play good chess is by learning. For this purpose, he first defined four problems that should be solved by computer while searching for the best next move. Then, he explored the concept of **likelihood** and the techniques, such as classification net (sort of decision tree), **reinforcement** and **bootstrapping** (sampling technique for better statistical assessment of a data set), that helped to provide a solution to these problems [13]. The design was not implemented but underlies later work by Newell, Shaw and Simon on NSS chess program.

Another game that attracted interest from the research community at the dawn of AI was the checkers. In 1951, Strachey wrote the first checkers program to run in Europe (UK). The first US checkers program was written by Samuel in 1952 [8]. The work that followed was quiet influential for machine learning as Samuel introduced the features allowing the program to **learn from experience**.

3.1.5.2. Machine translation

Machine translation is another early practical application of AI. The first Conference on Machine Translation took place in 1952 at MIT [9]. Promising demonstrations of the automatic translation of highly specialized texts from Russian to English, such as the Georgetown-IBM experiment in 1954, motivated the US government to support the researches in the field.

Modern range of applications is much vaster. A collection of use cases is provided, for instance, in ISO/IEC TR 24030 *Artificial Intelligence – Use cases* that will be used by standardization community to analyze the standardization needs and challenges of the companies using AI.

²⁰ <https://en.wikipedia.org/wiki/Heuristic>

3.2. Formal establishment of Artificial Intelligence

3.2.1. Dartmouth workshop

The development of interest in the topic resulted in the initiative by McCarthy, Minsky, Rochester and Shannon to propose a “2 month 10 men study of **artificial intelligence** during the summer of 1956 at Dartmouth” [7]. The proposal was approved and the event, now called Dartmouth workshop, took place and gave birth to the term “*Artificial Intelligence*”. The premise of the workshop was that “*every aspect of learning or any other feature of intelligence can in principle be so precisely described that a machine can be made to simulate it*” [7]. The following aspects of AI were considered during the gathering:

- Automatic computers,
- How can a computer be programmed to use a language,
- Neural networks,
- Theory of efficiency of a calculation,
- Self-improvement,
- Abstractions,
- Randomness and creativity.

The conference organizers thought that a significant progress could be done on one or more of the announced topics during one summer [7]. Aside from the four organizers, six other researchers – Samuel, Selfridge, Solomonoff, Newell, Simon and Bernstein – attended the conference. Each participant had its own ideas to present and discuss.

Simon and Newell came to Dartmouth with a working version of Logic Theorist (see Section 3.1.4 for more details). More generally, they were working on **symbol processing approach**. Minsky had started the work on **problem solving** using heuristics and mentioned its possible application to proving geometry theorems. McCarthy was interested in brain modeling and in designing an **artificial language to program a computer** to solve problems requiring conjecture and self-reference. Shannon was looking into the application of information theory concepts to computing machines and **brain models**. Rochester had background in neural networks. Solomonoff was interested in automating induction. Bernstein had been working on chess-playing program. Samuel had written a checkers program. Selfridge, as mentioned before (see Section 3.1.3), was working on **learning techniques** [4].

The event did not result in a breakthrough as the original proposal expected. Neither all the invited researchers could join for the full time, nor were they able to have constructive discussion, each one defending his own ideas [5]. However, the event set up an ecosystem and brought optimism for the work on AI for the forthcoming years.

3.2.2. High expectations and challenges of Artificial Intelligence in 50s-60s

Late 50s and the 60s were characterized by domination of a symbolic approach to the AI that consists in making the computers manipulate the symbols in a way similar to human perception. However, alternative approaches, such as neural networks or genetic algorithms, were also pursued.

Following the successful implementation of Logic Theorist, the advocates of symbolic approach Newell, Simon and Shaw continued working on the problem solving. In 1957, they ran the first version of their General Problem Solver, GPS. It was designed to **simulate human thinking**. It was dividing the problems given to it into sub-problems and applying a version of itself to solve a sub-problem, the process called **recursion**. While solving the problems, it was computing the differences between the problem to be solved and the existing knowledge and reducing this difference to solve the problem, a process Newell and Simon called “means-ends analysis” (similar

to backwards reasoning). GPS could be applied to a number of logical problems and puzzles, but it was highly dependent on the information that one could feed inside [3] [4].

To make it possible to run GPS on a computer the authors developed high-level Information Processing Language, **IPL**. This programming language was designed for a flexible data structures, such as lists. Later, in 1960, McCarthy combined IPL with lambda-calculus (a universal model of computation that could to simulate a Turing machine²¹) to obtain **LISP** (List Processor) programming language, which is used till today in some areas of AI [3].

Interested in **knowledge representation** and **reasoning**, McCarthy published in 1958 a paper entitled “Programs with Common Sense”. In this paper, he described the Advice Taker, a hypothetical program that could be seen as the first complete AI system. The Advice Taker would rely on the representation of the general knowledge of the world to solve problems. It would also add new knowledge to the database without being explicitly programmed to do so [3].

In 1959, Gelernter presented a paper where he described a program called the Geometry Theorem Prover. Similar to GPS, it was dividing the problems into sub-problems. Another interesting idea was to use the diagrams to close off unsuccessful search paths for solutions. The program was able to prove theorems that were difficult enough for many students in mathematics [3] [4].

Further development of chess playing programs was pursued. For example, Samuel included mechanisms for rote learning and generalization into his first program (described in Section 3.1.5). Thanks to these improvements, his program eventually won one game against a former Connecticut checkers champion in 1962 [8].

Driven by motivation to solve real-life problems, Minsky wrote a paper “Steps towards artificial intelligence” [14] where he described five areas that are essential to computer problem-solving using heuristics: **Search, Pattern recognition, Learning, Planning** and **Induction**. He also supervised students who were working on solving problems with rather limited scope but requiring intelligence to be solved. These limited problems came to be known as **microworlds**. The work on microworlds showed that AI could be used to solve real-life problems in limited application domains with heuristics, avoiding the challenges of representation of general knowledge of the world. Some examples of the work on microworlds are [3]:

- SAINT program by Slagle that was solving closed-form calculus integration problems typical of first-year college courses (1963),
- STUDENT program by Bobrow that was solving algebra problems written in natural language (1967),
- ANALOGY program by Evans that was solving geometric analogy problems that appear in IQ tests (1968).

Inspired by evolution theory, Friedberg did experiments where he was working on a population of random computer programs and attempting to evolve those that were successful in solving some specific tasks. He was doing that by taking the successful parts of the programs and introducing changes to the code of programs of next generation. The experiments of Friedberg gave birth to **genetic algorithms** of today [3].

The early work on neural networks was pursued as well. After the work of McCulloch and Pitts and Hebb, Rosenblatt continued the work on modeling human learning, cognition and memory in 50s-60s. The model he suggested was one of neural networks that he called **perceptrons** [4]. Perceptrons consisted of neural elements that had inputs, associated weights and one output. In the network, outputs of one element were inputs to others. Each input could have value of either 1 or 0. Neural element calculated the sum of inputs’ values multiplied by associated weights. If the sum was more than a threshold than the output would be 1, otherwise it would be 0. Rosenblatt defined different types of perceptrons, such as:

- series-coupled perceptrons, where the outputs of neural elements in one layer could go to the inputs of the elements in the next layer (currently known as **feed-forward networks**),

21 https://en.wikipedia.org/wiki/Lambda_calculus

- cross-coupled perceptrons, where the outputs of neural elements could become inputs to the elements in the same layer,
- back-coupled perceptrons, where the outputs of neural elements in one layer could become inputs to the elements in preceding layers,
- alpha-perceptron, for which a technique of adjusting weights that became known as **error-correction procedure** was used.

In 1969, Minsky and Papert wrote a book “Perceptrons” where they showed the limitations in representation capabilities of perceptrons. This work had negative impact on the future development of neural networks [3].

Independently of Rosenblatt, Widrow and his team started in 1959 the development of neural networks they called Adaline (for adaptive linear network). For training the Adalines, Widrow and Hoff suggested a Delta Learning rule, now known as **least means square adaptive method**, used to adjust the weights. Adaline could detect binary patterns. The next generation of their networks, called Madaline, was used to eliminate echo on the phone lines [4].

In 1958, Selfridge suggested a system called Pandemonium that could bridge the symbolic processing and the neural networks approach. The elements in Pandemonium could perform either nerve-cell type functions, as elements in neural networks, or higher-level cognitive functions, based on symbol processing. The structure of the system resembled the one of an organization, where at the bottom level were workers looking for specific features in the input data (for example, an image), and then shouting about their findings to higher elements in the hierarchy, middle-level managers. The louder the shouting, the better the chances of being heard. The middle-level managers were listening and shouting in their turn to the upper level managers about the most important findings from lower level, etc. At the upper level of hierarchy was a “decision demon” who was taking the final decision (for example, deciding what image was seen by workers: 9 or 0, R or A, etc.). The important notion behind the model described by Selfridge is the one of **parallel processing** as all the processes could run in parallel [4].

In 1966, the first chatbots were introduced by Weizenbaum and Colby. Weizenbaum built Eliza, program simulating human therapist. Colby built Parry, program simulating human paranoiac. Both programs could hardly be called intelligent since they worked only with a set of pre-registered sentences and a few programming tricks [8].

Most of these early developments remained either hypothetical or found very limited practical application faced with limited computational power and huge amount of time necessary to produce meaningful and useful results. Limitations of modern computers along with inflated and unsatisfied expectations resulted in criticism of AI, such as the Lighthill report [15] or the Dreyfus report [16]. This played an important role in the cut-off of the funding and the beginning of a period that may be referred to as first AI winter [2]. However, the research on AI did not stop completely. When it comes to reasoning systems, the attention was redirected from general-purpose knowledge representation and reasoning to specific application domains.

3.3. Rise of expert systems

Since the attempts to reproduce human intelligence or develop generalized knowledge systems were not very successful, the industry turned to more practical implementations, the **expert systems**, which could demonstrate the intelligent behavior when dealing with a specific task²².

²² <https://www.cs.swarthmore.edu/~eroberts/cs91/projects/ethics-of-ai/sec2.html>

Expert systems could be defined as “computer programs aiming to model human expertise in one or more specific knowledge areas” [1]. There are two basic components in an expert system: a knowledge database and an inference engine. Usually the expert systems also have an input/output interface to interact with users. The knowledge database contains the human experts’ knowledge in the area that is stored in the form of rules, for example of an “if X, then Y” structure. The inference engine processes the input information (for example, that X is true) and draws the deductions based on the rules (for example, Y) [1] [8].

The first known expert system is called Dendral, a common work of Feigenbaum, Buchanan and Lederberg from Stanford University (paper published in 1969). The system provided chemical analysis and could infer molecular structure based on the data from mass spectrometer. Without any domain knowledge, a system would need to consider enormous number of molecular combinations to solve the problem. However, upon consultation with chemists, the researchers found out about the existence of well-known patterns, the fact that helped to reduce the number of possible combinations significantly [3].

Another example of early expert systems is Mycin, by Feigenbaum, Buchanan, and Shortliffe. The system was providing medical diagnosis for blood infections and could prescribe medicines with dosage adequate to patient’s weight. Contrary to Dendral, there was no scientific theoretical framework to build knowledge base. The knowledge came from the doctors who in their turn got it from books and experience. This is why one of the important ideas of Mycin was the method of **calculating uncertainty**, which the authors called certainty factors [3] [8].

After these and some other successful applications, the expert systems became very popular in the industry where they were considered an important competitive advantage and were expected to provide magical solutions for most areas of human activities. In 1982, the Digital Equipment Corporation started using one of the first commercial expert systems, R1. The system was helping to configure orders for new computers and by 1986 was saving the company about \$40M per year. By the end of 80s more than a half of the companies from Fortune 500 list²³ were involved either in development or in maintenance of expert systems. Even now, expert systems remain important tool for decision support or even decision making (for non-expert users) in spite of various technological and business constraints [1] [3].

Multiple knowledge representation and reasoning languages were developed for expert systems. One is first-order logic based **Prolog** (1973). An alternative approach is based on Minsky’s idea of **frames** (1975), proposing a more structured approach, assembling facts about particular object and event types and arranging them into a large taxonomic hierarchy [3].

One of the main advantages of the expert systems is their explainability. However, there is a number of limitations and concerns for expert systems summed up in [1], such as:

- *Knowledge Acquisition and Analysis*: limited to the domain of expertise, human expertise not easily expressed in rules, etc.,
- *Handling Uncertain Situation*: in general the expert systems are not designed to handle uncertainty,
- *Validation and evaluation*: validation could be in comparison with human experts but yet not fully clear how verify the performance and reliability of the systems,
- *Maintenance Cost of Expert Systems*: maintaining the knowledge database could be expensive and technically challenging,
- *The systems make mistakes*: the user are in general more tolerant to human errors than to computer errors.

Current ongoing discussion about the possible standards related to input requirements, construction process (extraction, storage, fusion, etc.), performance metrics, application, etc. of knowledge database could help overcome some of these challenges by providing a set of good practices and recommendations.

²³ <http://fortune.com/fortune500/>

3.4. Progress in machine learning

With the limitations of the expert systems, machine learning regains ground starting from 80s [17]. In the 90s, with more and more data available, the machine learning shifts from knowledge-driven to data-driven approach and new algorithms to deal with big volumes of data begin appearing²⁴.

It has become possible to teach the computers to predict the right answer for a question based on a large number of examples with similar questions/answers. Thus, typical machine learning has three components: input data, learning process, and output data [18]. In order for the learning process to succeed, three main challenges should be addressed: data representation (to highlight the important features in the data), learning model evaluation (to evaluate model performance) and optimization (to find the best model) [19].

Typically, three different types of machine learning are distinguished [20]:

- **Supervised learning:** type of machine learning, where the algorithm uses labelled input/output data, to learn the pattern and be able to predict the values of outputs for new unseen input data.
- **Unsupervised learning:** type of machine learning, in which unlabeled data are used to train the algorithm, and the purpose is to explore the data and find some structure within.
- **Reinforcement learning:** type of learning where no raw data is given as input and instead the learning algorithm has to learn by means of reward to its actions.

Here are some findings that helped to shape modern machine learning.

After the work of Linnainmaa on **reverse mode of automatic differentiation** (later received the name of **back-propagation**) and its first application to neural networks by Werbos, Hinton and other researches further study back propagation in neural networks in 1985-1986. They tried different learning procedures and gained insights on the speed of learning and the scaling to bigger (having more inner layers) networks. This work characterized a revival of interest in neural networks [21] [22].

Also in 1986, Quinlan proposed a **Decision Tree** algorithm. A decision tree is a model that uses a tree-like graph to predict the output value of a target by learning simple decision rules inferred from the input characteristics/ input data features. Later in 2001, Breiman suggested using multiple decision trees where each of them is curated by a random subset of instances and each node is selected from a random subset of features. The algorithm got the name **Random Forest** [22] [23].

In 1989, Watkins developed **Q-learning**²⁵, a method to learn what action to take under which circumstances, which greatly advanced the development of reinforcement learning.

In 1995, Vapnik and Cortes proposed **Support Vector Machines** (Networks), SVM. SVM represents the input examples as a set of points of 2 types in N dimensional space and generates a (N - 1) dimensional hyperplane to separate those points into 2 groups. SVM then attempts to find a straight line that separates those points into 2 types and is situated as far as possible from all those points [22] [23].

An interesting technique, called **Adaboost**, was proposed in 1997 by Freund and Schapire. The idea behind Adaboost is to train simple and easy to train classifiers and after that use a weighted sum of their outputs to obtain the final output [22].

In 2006, Hinton, Osindero and Teh suggested a new way of training neural networks with multiple layers, where a fast greedy algorithm would allow initializing the parameters of a traditional neural network. The process got the name **deep learning** and revitalized the research around the topic [22] [24].

²⁴ <https://www.forbes.com/sites/bernardmarr/2016/02/19/a-short-history-of-machine-learning-every-manager-should-read/>

²⁵ <https://en.wikipedia.org/wiki/Q-learning>

Some highly mediated achievements that mark the progress of machine learning are the win of Deep Blue against the world's chess champion Kasparov in 1997, the victory of Watson at Jeopardy in 2011, the AlphaGo's beating the world's champion in Go in 2017, etc. [24].

Machine learning remains the mostly used branch of AI until today. For standardization community, this translates into a number of projects and discussions related to the explanation of an overall machine learning process, to the **robustness** of neural networks, to the assessment of **performance** of classification algorithms, etc.

It is possible to view the history and current interest in AI in terms of Gartner hype cycle. After the peak of inflated expectations in 50s-60s, the drop of the hype corresponding to AI winter and reaching plateau of productivity now [1]. A "One hundred year study on AI" [25] is a sign that AI is not disappearing any time soon. And the standardization efforts support its good and fair use.

4. Enablers of Artificial Intelligence

As the previous sections show, the AI is not a new technology. However, current interest and investments in the technology are much bigger than before [26].

Modern technological setting could be a reason for that. As discussed in Section 3.4, the recent boom of machine learning is data-driven. The huge volumes of data being generated nourish learning algorithms and allow better learning outcomes. More specifically, it is **big data** – characterized by big volumes, variety and velocity – that changes the ways the business is done today and fuels the AI. With AI, big data analysis could go beyond descriptive statistical analysis and generate even more value. Moreover, the development of tools for big data processing, such as for example Hadoop ecosystem, supports the application of AI techniques to big data [26] [27].

As presented in Section 3.4, the availability of data stimulated the researchers to find more accurate and fast algorithms. Thus, **new algorithms** were introduced, such as for example deep learning or random forest [26] [27].

In addition, the large annotated datasets are now available for AI researchers and developers that facilitates and accelerates the development of machine learning applications, especially the ones relying on the supervised learning algorithms [27]. Examples of such training sets are ImageNet, linguistic corpora (for example, British National Corpus), etc.

One of the major sources of data is **IoT**. IoT applications allow capturing environment-related data and using the outcomes of AI analysis to act back on the environment in a smart way. Already today, these applications are widely used in smart homes, logistics, etc. The positive outcomes for society help building trust and accelerate user acceptance of the technology. In the future, IoT coupled with AI could become key elements for time-critical applications, such as autonomous vehicles [27].

As mentioned in Section 3.2, early AI systems were often confronted with a lack of computational power. The development of **Cloud computing** dealt efficiently with this issue since it enables access to cost-efficient and scalable computing resources. Together, **Edge** and Cloud computing allow agile connectivity, real-time services, data optimization, application intelligence, as well as enhanced security and privacy protection. The hardware itself has become faster, especially with the introduction of **GPU** and **TPU** processors [26] [27].

Big data, IoT and computing resources could be considered as an ecosystem supporting the development and deployment of AI systems (see Figure 3). This is reflected at standardization level as well since the dedicated technical standardization committees working on these technologies are interconnected by means of enhanced

liaison relations. The developments and good practices in one domain are communicated to the committees working in other domains thus allowing for better interoperability within the whole ecosystem.

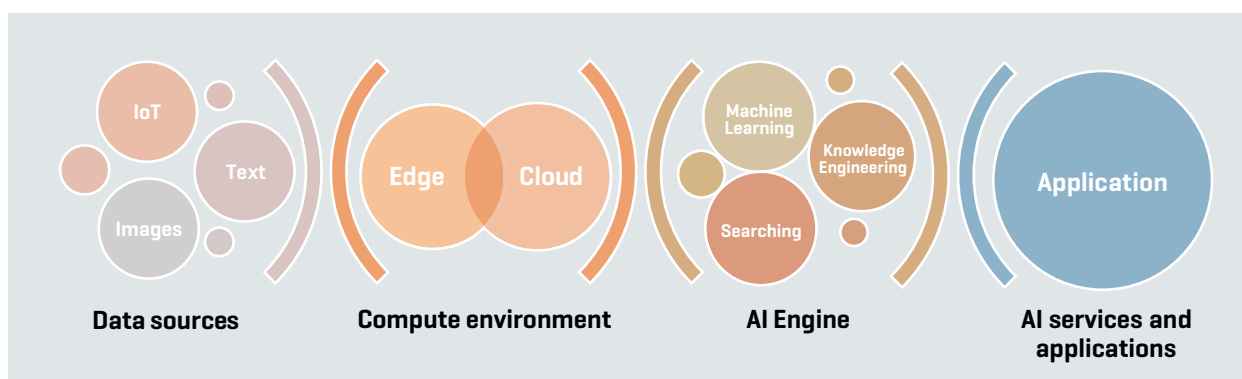


Figure 3: ICT ecosystem supporting the AI systems (inspired by Smart ICT Components in [37])

5. Definition of Artificial Intelligence

5.1. First definition

The term “*Artificial Intelligence*” was introduced in the proposal of McCarthy, Minsky, Rochester and Shannon for Dartmouth conference. The authors defined it as a problem of “*making a machine behave in ways that would be called intelligent if a human were so behaving*” [7]. This first definition introduces the two notions that were exploited in later definitions as well: intelligent and human-like behavior.

5.2. Modern definitions

The first definition refers to AI as a problem of creating intelligent machines. In a similar manner, it could also be defined as an activity of making intelligent machines/computer systems or as a branch of computer science dedicated to the purpose.

- Nillson defines AI as an “*activity devoted to making machines intelligent, and intelligence is that quality that enables an entity to function appropriately and with foresight in its environment*” [4].
- The English Oxford Living Dictionary²⁶ gives this definition of AI: “*The theory and development of computer systems able to perform tasks normally requiring human intelligence, such as visual perception, speech recognition, decision-making, and translation between languages*”.
- “One hundred year study on AI” refers to AI as “*a science and a set of computational technologies that are inspired by—but typically operate quite differently from—the ways people use their nervous systems and bodies to sense, learn, reason, and take action*” [25].

26 https://www.lexico.com/definition/artificial_intelligence

Another way of defining AI is to talk about the machines/systems that themselves present signs of intelligence.

- In the initial communication of European Commission, AI *“refers to systems that show intelligent behaviour: by analysing their environment they can perform various tasks with some degree of autonomy to achieve specific goals”* [28].
- Encyclopedia Britannica defines AI as *“the ability of a digital computer or computer-controlled robot to perform tasks commonly associated with intelligent beings”* [8].

Some sources combine these two ways of defining AI (for example, Encyclopedia Britannica [8]). This is the case of a definition proposed by the High-Level Expert Group on AI set up by the European Commission. This definition updates the initial one proposed by the European Commission so that it distinguishes between AI systems and a scientific discipline on the one hand, and elaborates on the capabilities of an AI system on the other hand:

“Artificial intelligence (AI) systems are software (and possibly also hardware) systems designed by humans that, given a complex goal, act in the physical or digital dimension by perceiving their environment through data acquisition, interpreting the collected structured or unstructured data, reasoning on the knowledge, or processing the information, derived from this data and deciding the best action(s) to take to achieve the given goal. AI systems can either use symbolic rules or learn a numeric model, and they can also adapt their behaviour by analysing how the environment is affected by their previous actions.

As a scientific discipline, AI includes several approaches and techniques, such as machine learning (of which deep learning and reinforcement learning are specific examples), machine reasoning (which includes planning, scheduling, knowledge representation and reasoning, search, and optimization), and robotics (which includes control, perception, sensors and actuators, as well as the integration of all other techniques into cyber-physical systems)” [29].

This definition was partly inspired by a definition of AI suggested by Norvig and Russel in their book *“Artificial Intelligence: a modern approach”* [3]. According to the authors, throughout the history, the development of the AI was motivated by discovering, analyzing, simulating and making the machines that would:

- Think humanly,
- Act humanly,
- Think rationally,
- Act rationally.

They retain the direction of acting rationally as the one requiring all the others and introduce a concept of a rational agent, given the agent is *“something that acts”* [3]. The approach to define an AI system through the introduction of a rational or intelligent agent, integrating the concepts of thinking and acting, is popular in AI literature [30]. An agent perceives the environment, reasons and acts (see Figure 4). The actions of an agent are driven by [30]:

- Its knowledge of the environment,
- History of interactions with the environment, such as received stimuli and past experiences,
- Its goals to achieve or preferences,
- Its abilities, that is the actions it is capable of carrying out.

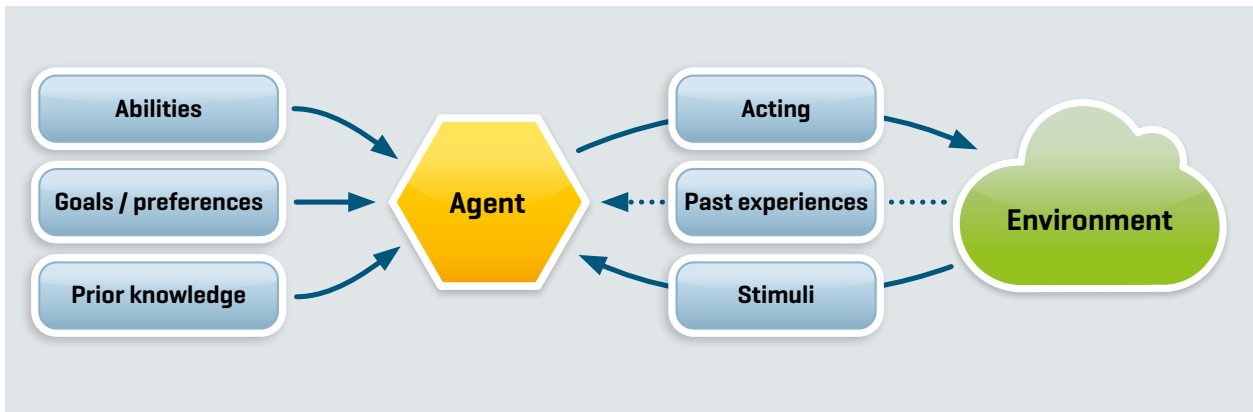


Figure 4: Defining the AI as an intelligent agent (adapted from [31])

Another practical approach to define AI was suggested in “MIT Technology Review”. It is based on a series of yes/no questions about what the system is doing and how it is doing that (see Figure 5) [31].

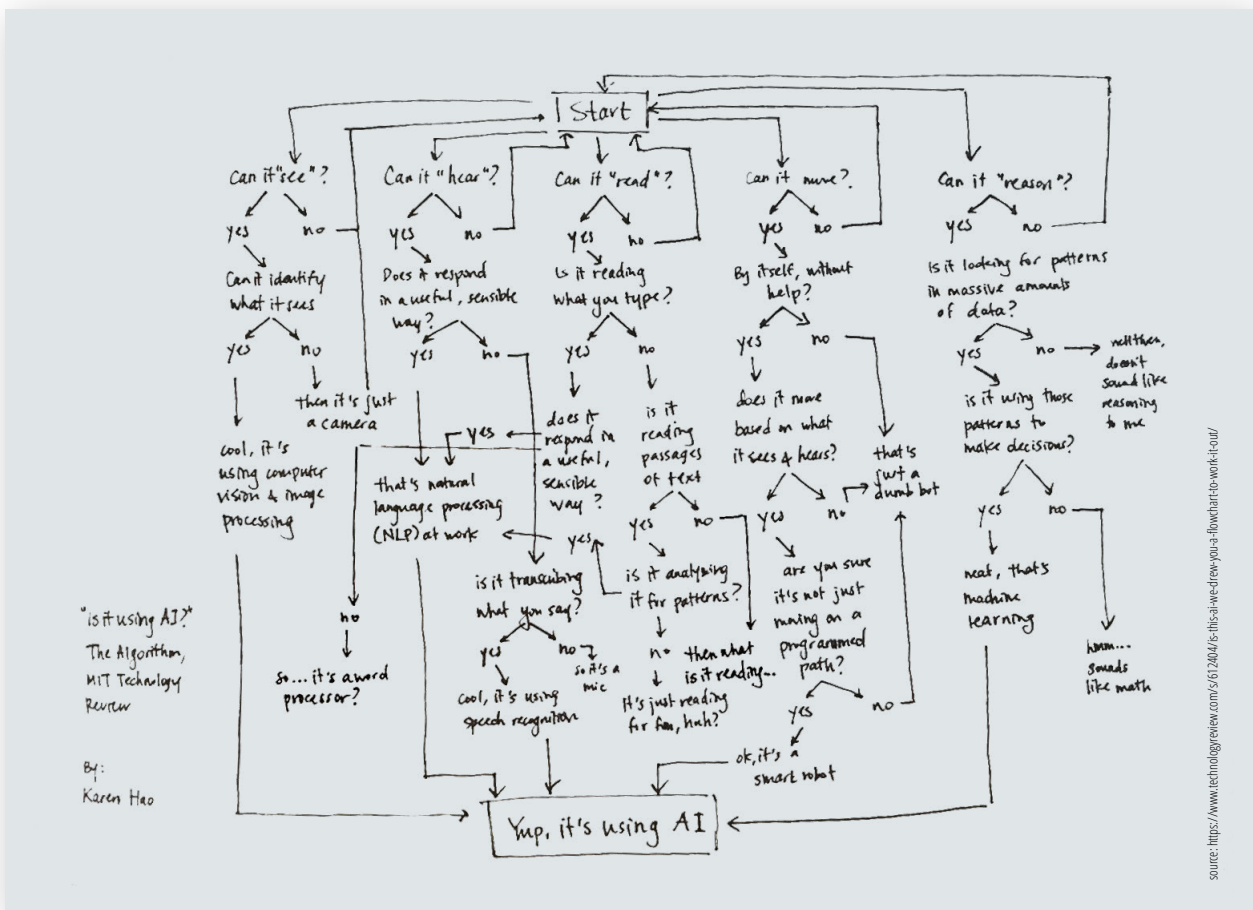


Figure 5: AI definition described in “MIT Technology Review” [31]

5.3. Standard definitions

Since one of the goals of standardization is to provide harmonized definitions for technologies in order to facilitate the exchanges between relevant stakeholders, the International Organization for Standardization, ISO, has been working on terminology related to AI and expert systems since 1989²⁷.

Published in 1993, the standard ISO/IEC 2382-1 on *Information technology vocabulary*, defined AI as a “*branch of computer science devoted to developing data processing systems that perform functions normally associated with human intelligence, such as reasoning, learning, and self-improvement.*”

Later, in 1995, two more definitions were suggested in ISO/IEC 2382-28:

- An update on the definition of ISO/IEC 2382-1 states that AI is an “*interdisciplinary field, usually regarded as a branch of computer science, dealing with models and systems for the performance of functions generally associated with human intelligence, such as reasoning and learning.*”
- A new definition that defines AI as a “*capability of a functional unit to perform functions that are generally associated with human intelligence such as reasoning and learning.*”

Currently all three definitions could be found in recent IT vocabulary standard ISO/IEC 2382 published in 2015. However, a dedicated technical committee is currently working to propose one single definition that could be used as a reference point.

²⁷ Standard terms and definitions are available through ISO Online Browsing Platform: <https://www.iso.org/obp/ui/#search>

6. Standardization efforts in Artificial Intelligence

6.1. Establishment and work of ISO/IEC JTC 1/SC 42 Artificial Intelligence

If the definitions of AI and related terms have been in ISO documents for a long time, it is not earlier than in 2017 that the dedicated technical committee, **ISO/IEC JTC 1/SC 42 Artificial Intelligence**²⁸, was established. This means the technology is:

- advanced and stable enough not to disappear from technical landscape in the near future,
- critical for economic development,
- would benefit from standard reference documents supporting various stakeholders.

Indeed, according to the preliminary report done by ISO/IEC JTC 1 *Information Technology*²⁹ group on emerging technologies and innovations (JETI), the AI is expected to impact multiple application domains. Different IT standardization committees have been studying AI techniques and their potential usage in respective domains and initiated a few standardization projects related to the application of AI. They also highlighted topics that necessitate standardization activities, such as interoperability, security, safety, ethics, etc. Thus, the report recommended to urgently initiate standardization activities on AI to provide guidance to relevant stakeholders on these and other relevant topics, such as terminology, reference architecture, etc. [32].

Accepting this recommendation, ISO/IEC JTC 1 established a new technical committee *SC 42 Artificial Intelligence* and suggested that this committee considers the following standardization areas:

- Foundational aspects,
- Computational methods,
- Trustworthiness,
- Societal concerns.

Currently, the work of *SC 42 Artificial Intelligence* is split among 6 groups and 22 ongoing projects, with 6 standards already published. Table 1 provides an overview of the activities of each group.

Group title	Project title	Status
Foundational standards	ISO/IEC 22989 <i>Artificial intelligence – Concepts and terminology</i>	Work in progress
	ISO/IEC 23053 <i>Framework for Artificial Intelligence (AI) Systems Using Machine Learning (ML)</i>	Work in progress
	ISO/IEC 42001 <i>Artificial intelligence – Management system</i>	Work in progress

²⁸ <https://www.iso.org/committee/6794475.html>

²⁹ <https://www.iso.org/isoiec-jtc-1.html>

Data	ISO/IEC 20546:2019 <i>Big data – Overview and vocabulary</i>	Published standard	
	ISO/IEC TR 20547-1:2020 <i>Big data reference architecture – Part 1: Framework and application process</i>	Published standard	
	ISO/IEC TR 20547-2:2018 <i>Big data reference architecture – Part 2: Use cases and derived requirements</i>	Published standard	
	ISO/IEC 20547-3:2020 <i>Big data reference architecture – Part 3: Reference architecture</i>	Published standard	
	ISO/IEC TR 20547-5:2018 <i>Big data reference architecture – Part 5: Standards roadmap</i>	Published standard	
	ISO/IEC 24668 <i>Artificial intelligence – Process management framework for Big data analytics</i>	Work in progress	
	ISO/IEC 5259-1 <i>Data quality for analytics and ML – Part 1: Overview, terminology, and examples</i>	Work in progress	
	ISO/IEC 5259-2 <i>Data quality for analytics and ML – Part 2: Data quality measures</i>	Work in progress	
	ISO/IEC 5259-3 <i>Data quality for analytics and ML – Part 3: Data Quality Management Requirements and Guidelines</i>	Work in progress	
	ISO/IEC 5259-4 <i>Data quality for analytics and ML – Part 4: Data quality process framework</i>	Work in progress	
	Trustworthiness	ISO/IEC TR 24028:2020 <i>Overview of trustworthiness in Artificial Intelligence</i>	Published standard
		ISO/IEC TR 24029-1 <i>Assessment of the robustness of neural networks – Part 1: Overview</i>	Being prepared for publication
ISO/IEC 24029-2 <i>Assessment of the robustness of neural networks – Part 2: Methodology for the use of formal methods</i>		Work in progress	
ISO/IEC TR 24027 <i>Bias in AI systems and AI aided decision making</i>		Work in progress	
ISO/IEC TR 24368 <i>Artificial intelligence – Overview of ethical and societal concerns</i>		Work in progress	

	ISO/IEC 23894 <i>Artificial Intelligence – Risk Management</i>	Work in progress
	ISO/IEC 25059 <i>Software engineering – Systems and software Quality Requirements and Evaluation (SQuaRE) – Quality Model for AI-based systems</i>	Work in progress
	ISO/IEC TR 5469 <i>Artificial intelligence – Functional safety and AI systems</i>	Work in progress
Use cases and applications	ISO/IEC TR 24030 <i>Artificial Intelligence – Use cases</i>	Being prepared for publication
	ISO/IEC 5338 <i>Artificial intelligence – AI system life cycle processes</i>	Work in progress
	ISO/IEC 5339 <i>Artificial Intelligence – Guidelines for AI applications</i>	Work in progress
Computational approaches and characteristics of artificial intelligence systems	ISO/IEC TR 24372 <i>Artificial intelligence (AI) – Overview of computational approaches for AI systems</i>	Work in progress
	ISO/IEC TS 4213 <i>Artificial intelligence – Assessment of machine learning classification performance</i>	Work in progress
	ISO/IEC 5392 <i>Artificial intelligence – Reference architecture of knowledge engineering</i>	Work in progress
Governance implications of AI	ISO/IEC 38507 <i>Governance of IT – Governance implications of the use of artificial intelligence by organizations</i>	Work in progress

Table 1: Projects of ISO/IEC JTC 1/SC 42 Artificial Intelligence

At the time of writing of the white paper, experts from 31 countries, among which Luxembourg, are actively involved in the activities of SC 42 *Artificial Intelligence*.

The technical committee also benefits from the expertise in other related domains through the liaison connections between committees. For example, experts from ISO/IEC JTC 1/SC 38 *Cloud Computing and Distributed Platforms*³⁰ and ISO/IEC JTC 1/SC 41 *Internet of Things and related technologies*³¹ contribute to better understanding of resources and infrastructure underlying many AI systems. Experts from ISO/IEC JTC 1/SC 27 *Information security, cybersecurity and privacy protection*³² provide insights on possible ways to address security and privacy challenges of Big Data and AI-based systems. Liaison experts from business-driven consortia in their turn bring up the business-related challenges to guide the standards development and support relevant stakeholders.

30 <https://www.iso.org/committee/601355.html>

31 https://www.iec.ch/dyn/www/f?p=103:7:0:::FSP_ORG_ID:20486

32 <https://www.iso.org/committee/45306.html>

6.2. Other relevant standardization activities

6.2.1. ITU-T

With the objective of supporting sustainable development, International Telecommunication Union, ITU, organizes various activities in the area of AI, thus helping stakeholders build a common understanding of technology. For example, since 2017 they organize AI for Good Global Summit. They also carry out standardization activities in the domain.

The most relevant ITU-T standardization activities related to AI are:

- *Study Group 16 Multimedia and its FG-AI4H³³ and FG-AI4AD³⁴*: Study Group 16 (SG16) has been working on standards related to multimedia coding, systems and application. In July 2018, the SG16 established a **Focus Group on Artificial Intelligence for Health** (FG-AI4H). It is motivated by the fact that AI has big potential to improve health care, however is rarely deployed in practice due to a number of business, legal, technical and other concerns. The objective of the group is to facilitate the global dialogue on the usage of AI for healthcare and develop standardized assessment framework of AI-based health solutions. Later, in October 2019, the SG16 established a **Focus Group on Artificial Intelligence for Autonomous and Assisted Driving** (FG AI4AD). Recognizing that AI can play an important role to reduce road deaths and injuries by encouraging safe, accessible and sustainable transport systems, the group seeks to increase public trust in autonomous and assisted driving, without which the widespread and socially acceptable deployment of AI on the roads is not possible. To do that, the group is working on the definition of a minimal performance threshold for the AI systems (such as AI as a Driver) that will meet or exceed the performance of a competent and careful human driver.
- *Study Group 13 Future Networks (& cloud) and FG-ML5G³⁵*: Study Group 13 (SG13) has been exploring the opportunities and addressing the challenges raised by the future networks, such as 5G, and by infrastructure and networking aspects of Cloud Computing and IoT. Under SG13, a **Focus Group on Machine Learning for Future Networks** including 5G (FG-ML5G) was established with the goal of studying relevant technological progress and helping the adoption of machine learning in future networks.
- *Study Group 20 IoT, smart cities and communities and FG-DPM³⁶*: Study Group 20 (SG20) has been working on the standards for IoT communication and services in the smart cities. Since IoT is a major source of big data, a **Focus Group on Data Processing and Management** (FG DPM) to support IoT and Smart Cities & Communities was established under direct responsibility of SG20. The FG DPM studied the questions of interoperability, data formats, data management, as well as the data quality, privacy and other aspects of enabling trust. As AI application rely on data, the work of this group could contribute to building trust in technology.

6.2.2. IEEE

The Institute of Electrical and Electronics Engineers, IEEE, mainly focuses on studying ethical aspects of technical standards related to AI. In March 2016, the IEEE Standards Association launched the **Global Initiative for Ethical Considerations in Artificial Intelligence and Autonomous Systems³⁷**, with the aim of helping people deal with the threats posed by AI and developing ethical design principles and standards that range from data privacy to

33 <https://www.itu.int/en/ITU-T/focusgroups/ai4h/Pages/default.aspx>

34 <https://www.itu.int/en/ITU-T/focusgroups/ai4ad/Pages/default.aspx>

35 <https://www.itu.int/en/ITU-T/focusgroups/ml5g/Pages/default.aspx>

36 <https://www.itu.int/en/ITU-T/focusgroups/dpm/Pages/default.aspx>

37 <https://standards.ieee.org/content/ieee-standards/en/industry-connections/ec/autonomous-systems.html>

fail-safe engineering. Under this umbrella, the IEEE has initiated a P7000 series³⁸ of 14 standard covering ethical considerations, bias, privacy challenges, etc. Out of these 14 standards, IEEE P7010™ *Recommended Practice for Assessing the Impact of Autonomous and Intelligent Systems on Human Well-being*, was published in 2020.

Among other relevant IEEE initiatives there is a group working on **Intelligent Process Automation (IPA)**³⁹. It has published in 2017 and is revising an IEEE P2755 *Guide for Terms and Concepts in Intelligent Process Automation*, which intends to provide a set of definitions established by and for the community involved with Software Based Intelligent Process Automation (SBIPA).

6.2.3. ETSI

The European Telecommunications Standards Institute, ETSI, considers AI through standardization of various topics as indicated in their White Paper No. #34 “Artificial Intelligence and future directions for ETSI” [33]:

- **5G:** in 5G specifications AI is referenced in the context of Core Network capabilities (5G NG Core) and Radio Access Network (5G RAN). In both cases, AI can increase 5G network automation and effective management and orchestration. The 5G standard specifies the capabilities of services using AI and the way other services can access the results.
- **Network optimization and end-to-end service assurance:** AI is considered as key enabler for the creation of largely autonomous networks capable of self-configuration, self-monitoring, self-healing and self-optimization (it can be a tool in network virtualization, it can drive adaptive behavioral changes in the network, it can support automated service assurance, etc.). In this context, AI is referred to in the specifications for Network Function Virtualization (NFV), Zero-touch Network and Service Management (ZSM), Experiential Networked Intelligence (ENI).
- **IoT, data acquisition and management, governance and provenance:** in the context of machine-to-machine, the opportunities for improving AI/ML performance through its introduction into the IoT systems are considered. A reference architecture introducing AI into IoT system is described, along with a proof of concept that shows how the AI/ML-related functional requirements can be addressed and how AI/ML can be used as-a-service in IoT solutions. Moreover, the issues of data provenance, data quality metadata, semantic interoperability, etc. are considered.
- **Security and Privacy:** with the objective to minimize harm from AI and maximize its benefits, three aspects of AI and security are considered. First, how to secure an AI-based system from attack. Second, how to protect a system/service/etc. from an attack that exploits AI capabilities. Third, how to enhance conventional security measures using AI. To address the topic of security, a dedicated **Industry Specification Group Securing AI**⁴⁰ was put in place by ETSI.
- **Testing:** having identified a need for a Test and Certification Framework for AI Models for autonomic networks, an AI Model Life Cycle Management Process was proposed, covering development, training, testing, certification, and deployment stages and the support expected from three main associated stakeholders. Moreover, a new specification “AI in Test Systems and Testing AI Models” was initiated to illustrate the usage of AI in test systems and to propose a guide for testing AI Models.
- **Health and Societal applications of AI:** recognizing the potential of AI in healthcare, the standardization efforts aim to address the safety, security and privacy concerns related to the usage of AI in this application domain by providing clear descriptions of AI systems along with the guidance on how to monitor and control them. Another topic of interest is the usage of AI in the response to emergencies, including the mission critical aspects, accuracy and display of geo-location, etc.

38 <https://standards.ieee.org/initiatives/artificial-intelligence-systems/standards.html#p7000>

39 <https://standards.ieee.org/standard/2755-2017.html>

40 <https://www.etsi.org/committee/sai>

6.2.4. CEN-CENELEC

At the beginning of 2019, two European organizations for standardization, European Committee for Standardization (CEN) and European Committee for Electrotechnical Standardization (CENELEC), put in place a **Focus Group on Artificial Intelligence** with the objective to prepare a European standardization roadmap for the technology. By October 2020, the group provided CEN-CENELEC their report where they considered the format of future European standardization for AI and the topics to be addressed. Namely, it was recommended to establish a Joint Technical Committee on AI to work on the development of the European standards giving the priority to the following topics: accountability, quality, data for AI, security and privacy, ethics, engineering of AI systems, safety of AI systems.

7. Conclusion

AI is not a new technology. It has been maturing for more than 60 years and is advanced enough to bring the real added value to economy and society. It draws inspiration from multiple domains and in its turn impacts most of the areas of our lives.

As any technology, it comes with its own challenges. People fear bias and unemployment; questions of liability preoccupy stakeholders' minds in case something would go wrong; the possible ethical choices behind the actions of an AI system are carefully studied since they are based on regional and cultural preferences; privacy, security and safety of the end users must be kept in mind by developers and legislators [25] [34] [35].

The list could continue. This white paper sets the scene for the presentation of modern AI applications and algorithms. Next chapters discuss their limitations and challenges, demonstrate how these challenges could be addressed and what role technical standardization plays in it.

References

- [1] C. Smith, B. McGuire, T. Huang and G. Yang, "The History of Artificial Intelligence," 2006. [Online]. Available: <https://courses.cs.washington.edu/courses/csep590/06au/projects/history-ai.pdf>. [Accessed 11 2020].
- [2] C. Garvey, "Broken promises & empty threats: the evolution of AI in the USA, 1956-1996," 03 2018. [Online]. Available: <http://www.technologystories.org/ai-evolution/>. [Accessed 11 2020].
- [3] S. Russel and P. Norwig, *Artificial Intelligence: a modern approach (3rd edition)*, Upper Saddle River, New Jersey: Pearson Education, Inc., 2010.
- [4] N. J. Nilsson, *The Quest for Artificial Intelligence: A History of Ideas and Achievements*, New York: Cambridge University Press, 2009.
- [5] P. McCorduck, *Machines who think : a personal inquiry into the history and prospects of artificial intelligence (2nd edition)*, Natick, Massachusetts: A K Peters, Ltd., 2004.
- [6] S. S. Adams, I. Arel, J. Bach, R. Coop, R. Furlan, J. S. Hall, B. Goertzel, A. Samsonovich, M. Scheutz, M. Schlesinger, S. C. Shapiro and J. F. Sowa, "Mapping the landscape of human-level artificial general intelligence," *AI Magazine*, vol. 33, no. 1, pp. 25-42, 2012.
- [7] J. McCarthy, M. L. Minsky, N. Rochester and C. E. Shannon, "A proposal for the Dartmouth summer research project on artificial intelligence," 1955. [Online]. Available: <http://www-formal.stanford.edu/jmc/history/dartmouth/dartmouth.html>. [Accessed 11 2020].
- [8] J. Copeland, "Artificial Intelligence," *ENCYCLOPÆDIA BRITANNICA*, [Online]. Available: <https://www.britannica.com/technology/artificial-intelligence>. [Accessed 11 2020].
- [9] W. Bibel, "Artificial Intelligence in a historical perspective," *AI Communication*, vol. 27, no. 1, pp. 87-102, 2014.
- [10] A. Turing, "On computable numbers, with an application to the entscheidungsproblem," *Proceedings of the London Mathematical Society*, vol. 42, pp. 230-265, 1936.
- [11] A. Turing, "Computing machinery and intelligence," *Mind*, vol. 49, pp. 433-460, 1950.
- [12] O. Selfridge, "Pattern recognition and modern computers," in *AFIPS'55 - Western Joint Computer Conference*, Los Angeles, 1955.
- [13] A. Newell, "The chess machine: an example of dealing with a complex task by adaptation," in *AFIPS'55 - Western Joint Computer Conference*, Los Angeles, 1955.
- [14] M. Minsky, "Steps toward Artificial Intelligence," *Proceedings of the IRE*, vol. 49, no. 1, pp. 8-30, 1961.
- [15] J. Lighthill, "Artificial Intelligence: A General Survey," *Science Research Council*, 1973.
- [16] H. Dreyfus, *What Computers Can't Do: A Critique of Artificial Reason*, New York: Harper & Row, 1972.
- [17] P. Brey, "Hubert Dreyfus: Humans Versus Computers," in *American Philosophy of Technology: The Empirical Turn*, Indiana University Press, 2001, pp. 37-63.
- [18] Gartner, "Preparing and architecting for Machine Learning," 2017. [Online]. Available: <https://www.gartner.com/en/documents/3573617/preparing-and-architecting-for-machine-learning>. [Accessed 11 2020].
- [19] J. Brownlee, "Basic Concepts in Machine Learning," 2015. [Online]. Available: <https://machinelearningmastery.com/basic-concepts-in-machine-learning/>. [Accessed 11 2020].
- [20] Towards Data Science, "Machine Learning For Beginners," 2018. [Online]. Available: <https://towardsdatascience.com/machine-learning-for-beginners-d247a9420dab>. [Accessed 11 2020].
- [21] D. Plaut, S. Nowlan and G. Hinton, "Technical report CMU-CS-86-126: Experiments on learning by back-propagation," Department of Computer Science, Carnegie-Mellon University, 1986.
- [22] E. Golge, "Brief History of Machine Learning," [Online]. Available: <http://www.erogol.com/brief-history-machine-learning/>. [Accessed 11 2020].
- [23] S. Sakr, R. Elshawi, A. Amjad , W. Qureshi, C. Brawner, S. Keteyian, M. Blaha and M. Al-Mallah, "Comparison of machine learning techniques to predict all-cause mortality using fitness data: the Henry ford exercise testing (FIT) project," *BMC medical informatics and decision making*, vol. 17, no. 1, 2017.
- [24] "A history of machine learning," [Online]. Available: <https://cloud.withgoogle.com/build/data-analytics/explore-history-machine-learning/>. [Accessed 11 2020].

- [25] Report of the 2015-2016 Study Panel, "One hundred year study on artificial intelligence: artificial intelligence and life in 2030," 09 2016. [Online]. Available: <https://ai100.stanford.edu/2016-report>. [Accessed 11 2020].
- [26] B. Hodjat, "The AI Resurgence: Why Now?," 03 2015. [Online]. Available: <https://www.wired.com/insights/2015/03/ai-resurgence-now/>. [Accessed 11 2020].
- [27] IEC, "Artificial intelligence across industries," 2018. [Online]. Available: <https://basecamp.iec.ch/download/iec-white-paper-artificial-intelligence-across-industries-en/>. [Accessed 11 2020].
- [28] European Commission, "Factsheet: Artificial Intelligence for Europe," 04 2018. [Online]. Available: <https://ec.europa.eu/digital-single-market/en/news/factsheet-artificial-intelligence-europe>. [Accessed 11 2020].
- [29] High-Level Expert Group on AI, "A definition of AI: Main capabilities and disciplines," European Commission, 2019.
- [30] D. L. Poole and A. K. Mackworth, *Artificial Intelligence: Foundations of Computational Agents*, Cambridge University Press, 2017.
- [31] K. Hao, "Intelligent machines: Is this AI?," 11 2018. [Online]. Available: <https://www.technologyreview.com/s/612404/is-this-ai-we-drew-you-a-flowchart-to-work-it-out/>. [Accessed 11 2020].
- [32] ISO/IEC JTC1, "ISO/IEC JTC 1 JETI N13587, Report on Artificial Intelligence (AI) and Autonomous Systems (AS)," 2017.
- [33] ETSI, "Artificial Intelligence and future directions for ETSI," 2020.
- [34] A. Campolo, M. Sanfilippo, M. Whittaker and K. Crawford, "AI Now 2017 Report," AI Now Institute, New York, 2017.
- [35] European Parliament, "Understanding artificial intelligence," European Parliamentary Research Service, 2018.
- [36] ISO/IEC JTC 1/SC 42 Artificial Intelligence, "ISO/IEC TR 24030, Information technology -- Artificial Intelligence (AI) -- Use cases," [Online]. Available: <https://www.iso.org/standard/77610.html>. [Accessed 11 2020].
- [37] ILNAS, "Standards Analysis Smart ICT V2.0," 2018.

NOTE: While any hyperlinks included in this chapter were valid at the time of consultation, ILNAS cannot guarantee their long-term validity.

2

Artificial Intelligence: technology overview

1. Introduction

Applications of Artificial Intelligence (AI) are numerous and diverse. AI could be used to predict the weather, identify spam messages, answer questions, play games, and so on. To have a successful solution, each of these tasks relies on a different technique. Some would require a knowledge base to search for the answer, others would achieve the result by learning from experience. This chapter aims at providing a high-level overview of three major types of AI techniques. For each type, a general description will be provided along with the introduction of tasks being solved and the most popular algorithms to solve them.

2. AI paradigm: overview

AI helps to solve business problems, using different types of techniques such as [1]:

- searching for the optimal solution,
- learning,
- reasoning.

Searching techniques are used to determine a sequence of actions that starts from an initial state and leads to a final desired state in some state space. **Learning** techniques rely on mathematical and statistical functions to get insights from data, generally large volumes of it. **Reasoning** techniques work on a domain model where domain knowledge representation and inference rules are logic-based [1] [2].

In [3], the authors describe these problem-solving techniques with respect to the complexity of the underlying **world representation**, such as shown in Figure 1. Searching techniques use atomic representation, where the states of the world have no internal structure. Learning and some reasoning (using propositional logic) techniques are based on factored representation where each state of the world is split into a fixed set of variables - or attributes - that can have specific values. Finally, complex reasoning is based on a structured representation where objects and their relationships are explicitly described.

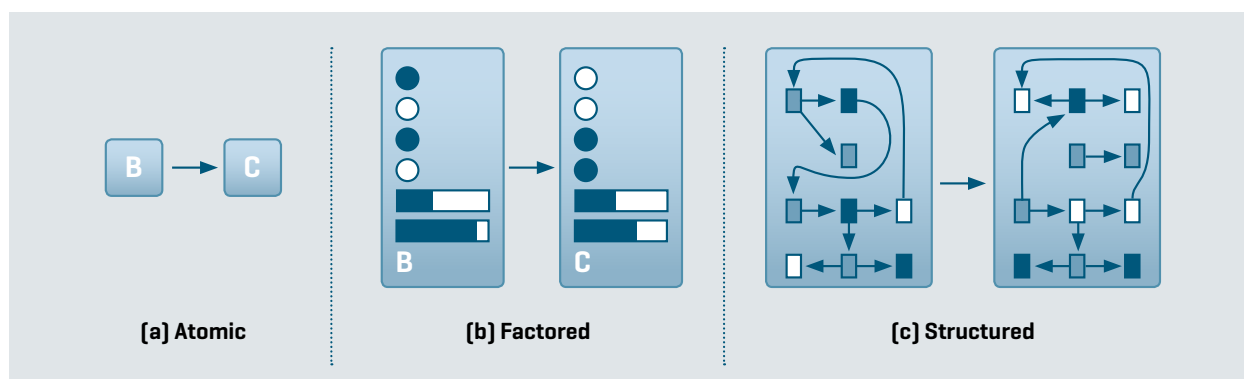


Figure 1: Depiction of different levels of world representation [3]

An alternative classification of AI techniques can be found in the dedicated entry of the Stanford Encyclopedia of Philosophy, where the authors distinguish between intelligent agent-based AI, logic-based AI and non-logicist AI [4]. Intelligent agent-based AI may be considered as a mix of state space exploration and learning through experience. Logic-based AI consists in developing limited-objective reasoning systems that rely on large bodies

of declarative information. Finally, non-logicist AI consists in learning, using either probabilistic symbolic (for example, Bayesian networks) or non-symbolic (such as neural networks) approaches (see Chapter 1 for the introduction of symbolic and non-symbolic approaches) [4].

A more complex classification of a variety of AI techniques is presented in the AI Knowledge Map [5]. This classification is based on two criteria: which AI paradigm is used and to which AI problem domain it is applied. AI paradigms are the main global approaches to solving AI problems; these are the symbolic, statistical, and sub-symbolic approaches. AI problem domains are categorizations of problems that are traditionally solved by AI, namely perception, reasoning, knowledge, planning and communication. Based on these two criteria, different classes of AI techniques can be described, as shown in Figure 2 [5].

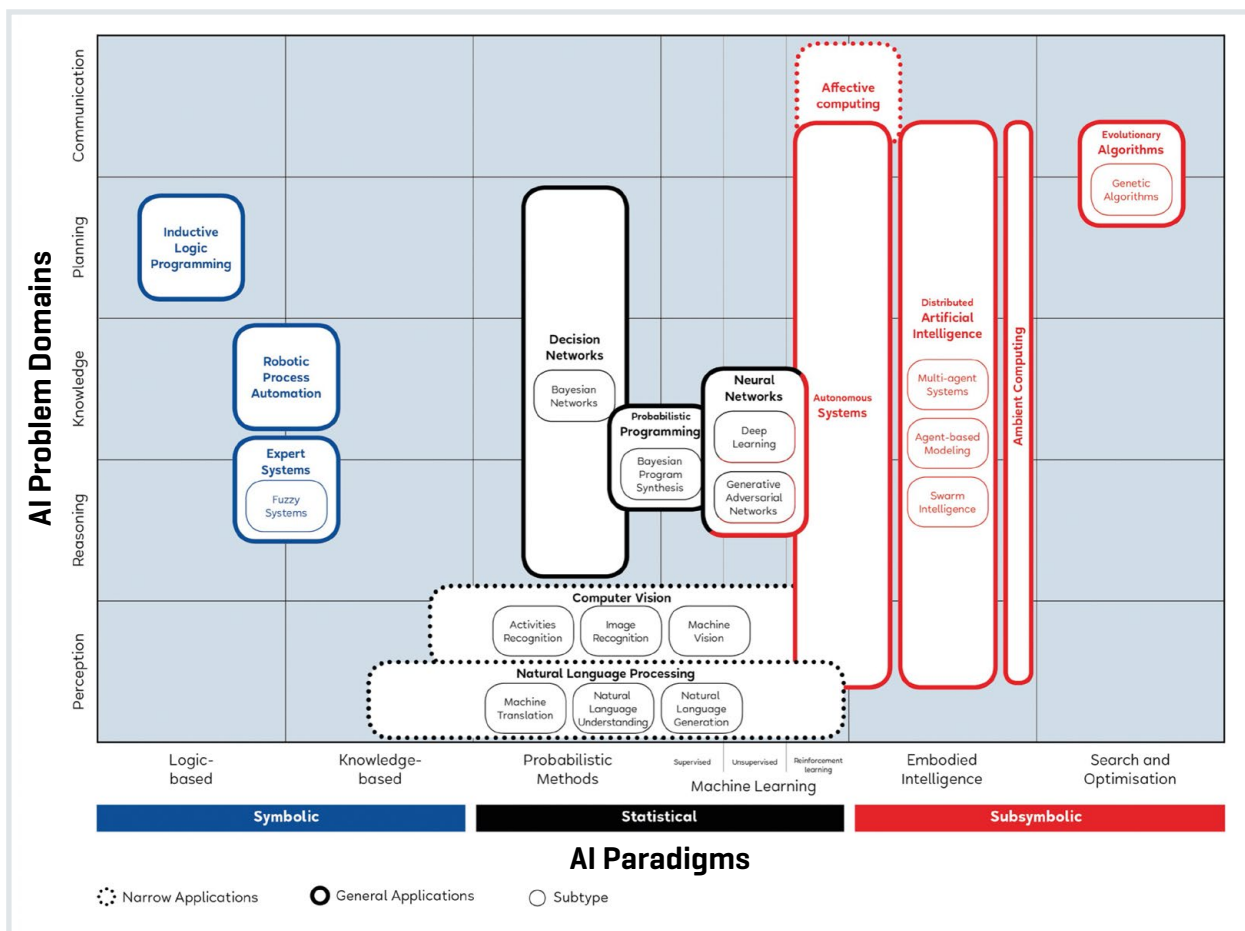


Figure 2: AI Knowledge Map [5]

Following the standard approach of ISO/IEC JTC 1/SC 42/SG 1 on *Computational approaches and characteristics of artificial intelligence systems* [1], the first classification is adopted for the purpose of this white paper. This chapter provides further details on searching, machine learning (ML), and reasoning techniques. First, basic search optimization algorithms are described, then different types of ML and some popular algorithms are presented, and finally different types of knowledge representation and basic logic models are introduced. A technical report ISO/IEC TR 24372 *Information technology – Artificial intelligence (AI) – Overview of computational approaches for AI systems*⁴¹, enhancing this presentation, is currently being developed by ISO/IEC JTC 1/SC 42/WG 5 on *Computational approaches and computational characteristics of artificial intelligence systems*⁴².

41 <https://www.iso.org/standard/78508.html>

42 ISO/IEC JTC 1/SC 42/WG 5 was established following the Study Report and the recommendations made by ISO/IEC JTC 1/SC 42/SG 1 on *Computational approaches of AI systems*. Hence, the similarity in the title of the group and the continuity in the program of work.

3. Searching

This section describes **searching techniques to solve problems**. Simply speaking, such problems consist of an initial state, a set of goals and a set of possible actions. The objective is to search for a series of feasible actions that lead the agent from the initial state to its goal(s). This problem definition is similar to the planning problem. However, searching is based on an atomic representation of the world, where the agent's actions have no effect on the world, in contrast to planning, which requires a more complex world representation in order to be able to reflect the effects of the agent's actions [3] [6].

When talking about searching, the problem of finding a sequence of actions that leads to achieving the goal is generally abstracted as a **search problem in a directed graph** [6]. A directed graph consists of:

- a set N of nodes (or vertices),
- a set A of arcs (or edges), where an arc is an ordered pair of nodes.

Thus, a **state** is a representation of a world configuration, and a **node**⁴³ is a part of a graph corresponding to a state but also including information about parent nodes, the actions that led to this node, etc. [7].

For the illustration of search algorithms, a special type of a directed graph - called a tree - will be used in this section. A tree contains no loops, and every node has exactly one incoming arc with the exception of a root node that has no incoming arcs [6]. Examples of a graph and a tree are shown in Figure 3.

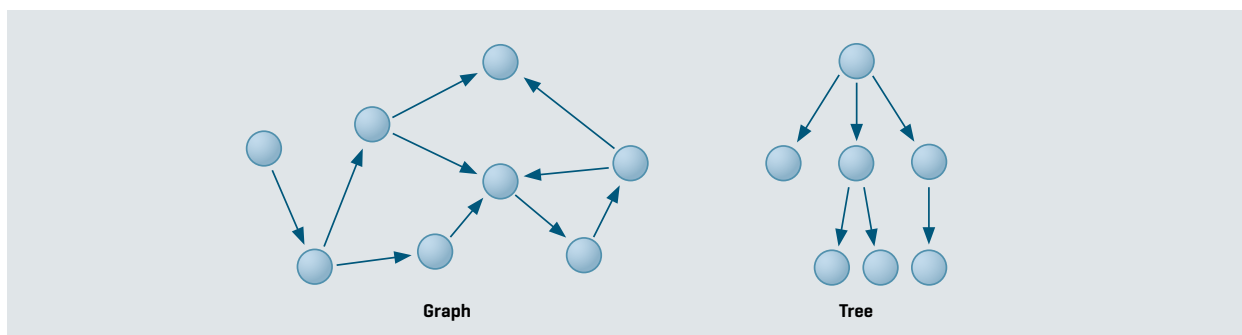


Figure 3: Examples of a graph (on the left) and a tree (on the right) [2]

This section formalizes the search problem and presents different search algorithms.

3.1. Search problem definition

Before starting the search for solutions, a goal should be identified and a problem should be well defined [3].

A problem is formally defined with the following components [2] [3] [6] (as illustrated in Figure 4):

- a set of states $S = \{s_1, s_2, \dots, s_n\}$,
- the initial state S_{start}
- for each state, a set of actions $Actions(s) = \{a_1, a_2, \dots, a_n\}$ available in that state,

⁴³ The node corresponds to a state in the search space. In this document for simplicity the term "state" is typically used.

- an **action function** that returns a new state (or Successor) given a state and an action (or a transition model describing the results of the actions),
- a **goal function** that returns **true** if the state **s** is a **goal** state and **false** otherwise,
- a path cost function **Cost (s,a)** that assigns a numeric cost to each path.

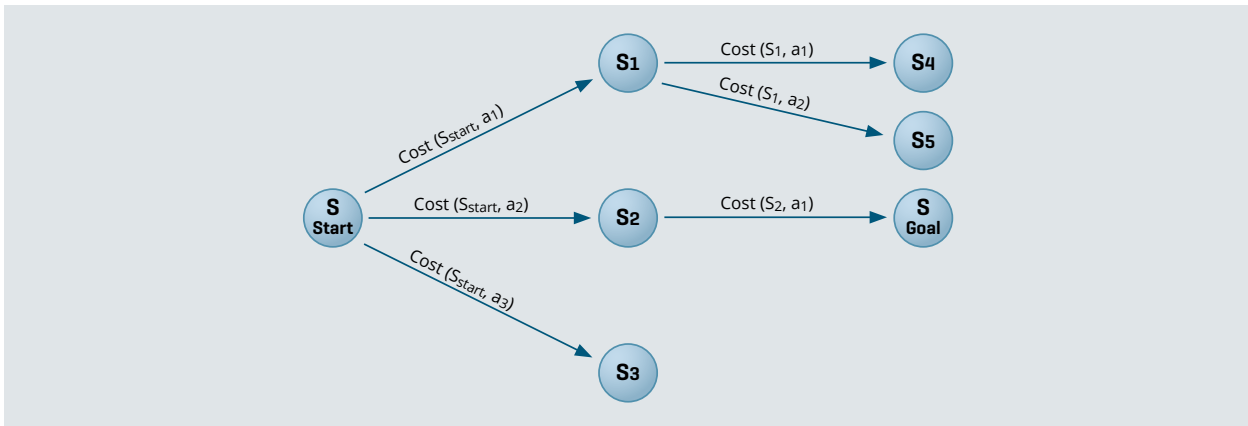


Figure 4: Simple search problem definition (adapted from [2])

The cost function reflects the desired performance measure. For example, in the case of a GPS navigator it could be the distance travelled (shortest path), the time required to achieve the destination (fastest path), etc.

A path that leads from the initial state to a goal state is a solution. The quality of the solution is measured by the total path cost and an optimal solution has the lowest path cost among all solutions.

3.2. Search algorithms

3.2.1. Generic search algorithm

The inputs to the search function are the problem, containing the initial state, a set of states with the associated actions (a search tree), and a goal function.

The output of the search function is the proposed path to the state for which the goal function returns true or an empty set if there are no solution paths.

In order to keep track of the exploration of a state space, a concept of a **frontier** is introduced. A frontier in a search algorithm is a set of paths from the initial state that have been explored so far and that could form segments of paths from initial state to the goal state.

The **search algorithm** can be summarized as follows [3] [6]:

- Initialize the Frontier with the initial state $\{(S_{start})\}$
- Go into the loop while the Frontier is not empty:
 - Select and remove a path (S_{start}, \dots, S_k) from the Frontier
 - If the S_k is the goal, return the path (S_{start}, \dots, S_k)
 - Else update the Frontier: check all possible actions from a current state, and add the new possible states to the Frontier
- Return the empty set if the Frontier becomes empty and no solution was found

Figure 5 [6] illustrates the generic search algorithm.

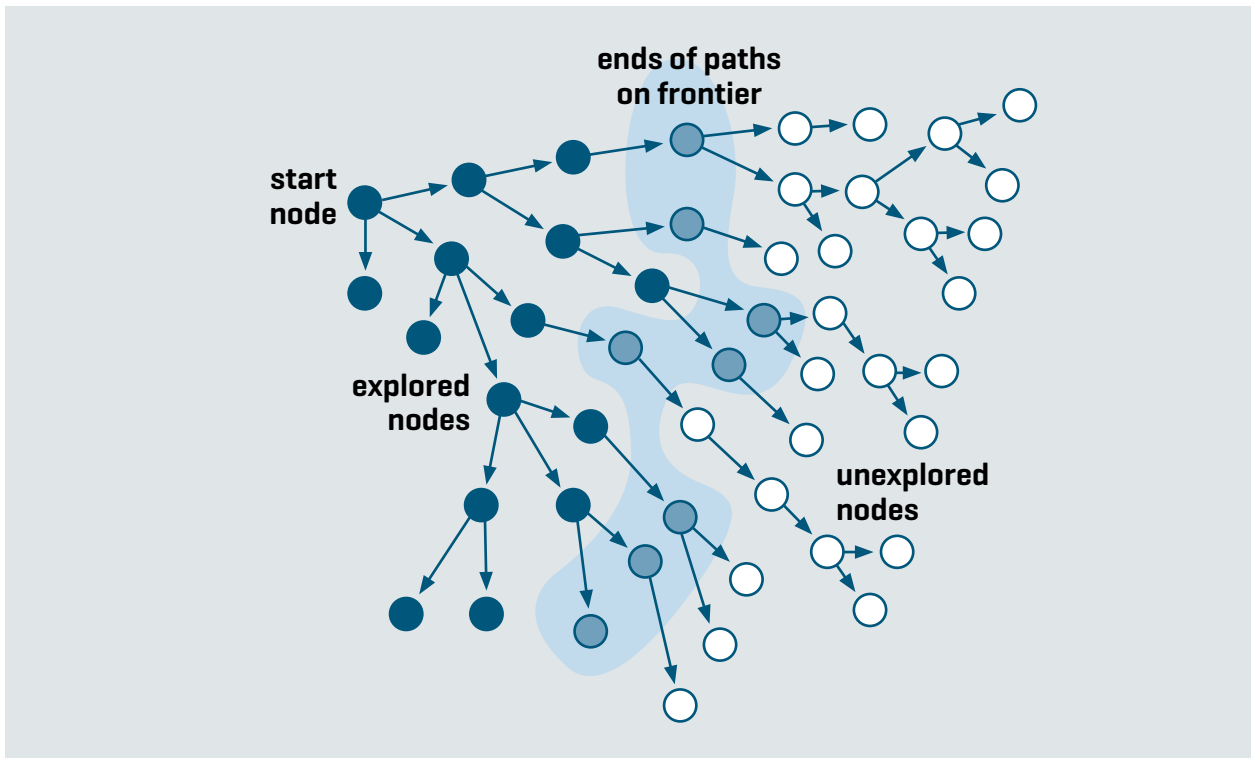


Figure 5: Illustration of a generic graph search [6]

Since a graph can be cyclic (contain loops), a graph search function also keeps track of the explored states in order to not explore them multiple times or reverse course, dealing thus with potential redundancy [3].

There exist different search algorithms and the difference between them arises from the strategy to explore the states and update the frontier. The algorithms can be evaluated with respect to their performance: does it always find a solution, how many states it had to explore before finding a solution (memory required by the algorithm), how fast the solution is found, etc. Indeed, some algorithms will allow to find a solution faster in a small state space, but will be unable to find a solution in an infinite state space, while other algorithms guarantee to find a solution but will not be necessary optimal in terms of time consumed [3]. In the following sections some basic search algorithms are introduced. These could be applied to different domains (for example, health or logistics) by tailoring the generic principle to the specificity of domain.

3.2.2. Uninformed search strategies

Uninformed search strategies do not take into account the knowledge about the environment and the location of the goal. They advance in a graph without knowing where they are going until they find the goal.

Breadth-first search (BFS) is a search strategy where the graph is parsed level-by-level. Thus, in BFS the shallowest nodes are explored first. The search starts with an initial node, then all its successors are expanded, then their successors, and so on, until the goal is found or all the nodes have been explored. Figure 6 illustrates in what order the nodes get explored using BFS strategy [2] [3].

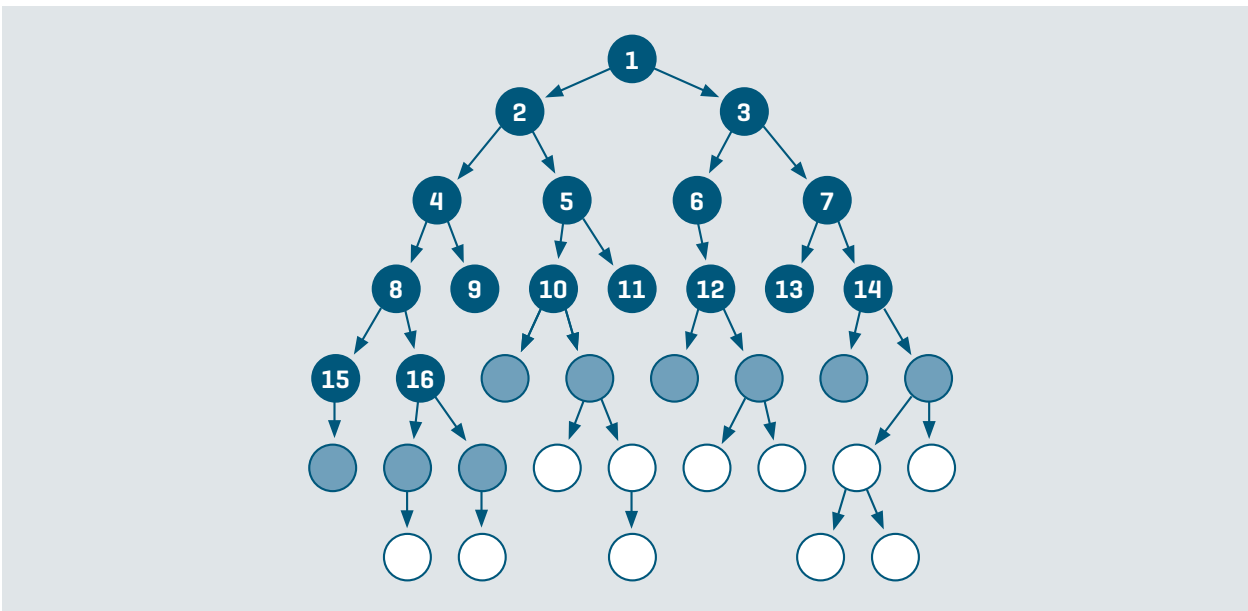


Figure 6: Order in which the nodes get explored using breadth-first search (the light blue nodes are the Frontier) [6]

Depth-first search (DFS) is a search strategy where the deepest path is expanded first. The graph is parsed by following each path as deeply as possible. The search proceeds immediately to the deepest level of the search tree, where the nodes have no successors. As those nodes are expanded, they are dropped from the frontier, so then the search “backs up” to the next deepest node that still has unexplored successors. Figure 6 illustrates in what order the nodes are explored using DFS strategy [2] [3].

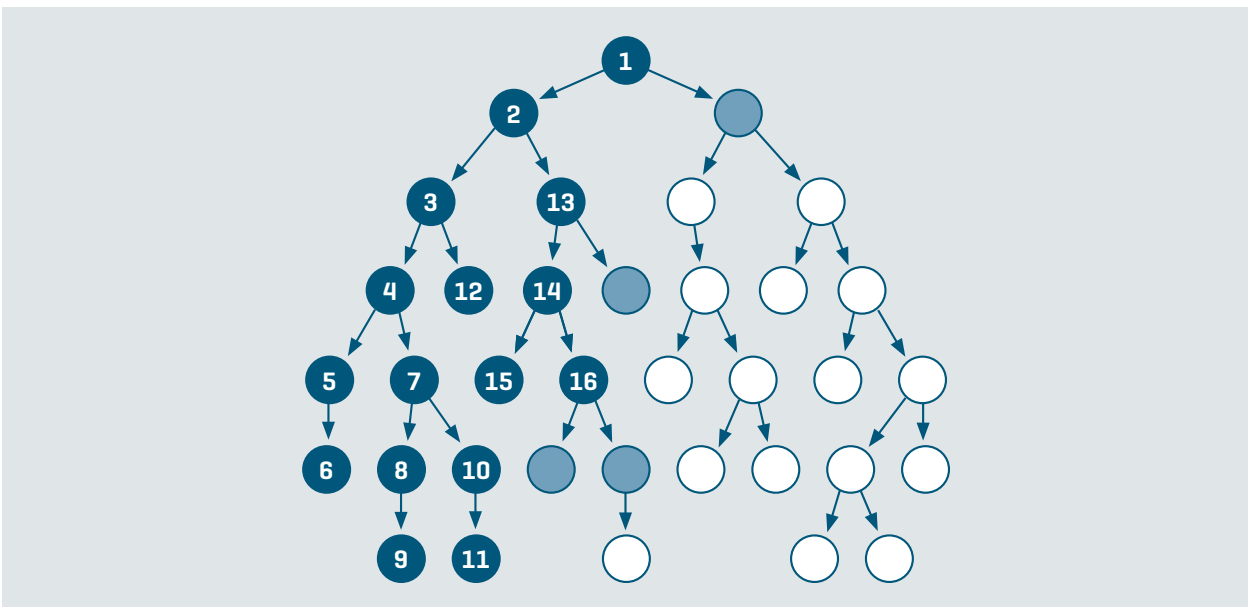


Figure 7: Order in which the nodes get explored using depth-first search (the light blue nodes are the Frontier) [6]

Iterative deepening search is a modification of the depth-first search algorithm. It iteratively increases the depth-limit to apply the depth-first search strategy but at each iteration never explores the paths with more arcs than the defined depth bound. Iterative deepening starts with a depth-search to depth 1 by building paths of length 1. If the goal was not found, it does the depth-search of depth 2, then of depth 3, etc. until a solution is found (if there is any). Thus, it combines the advantages of depth-first and breadth-first search strategies. Figure 8 shows how the nodes are parsed using an iterative-deepening search strategy [2] [3] [6].

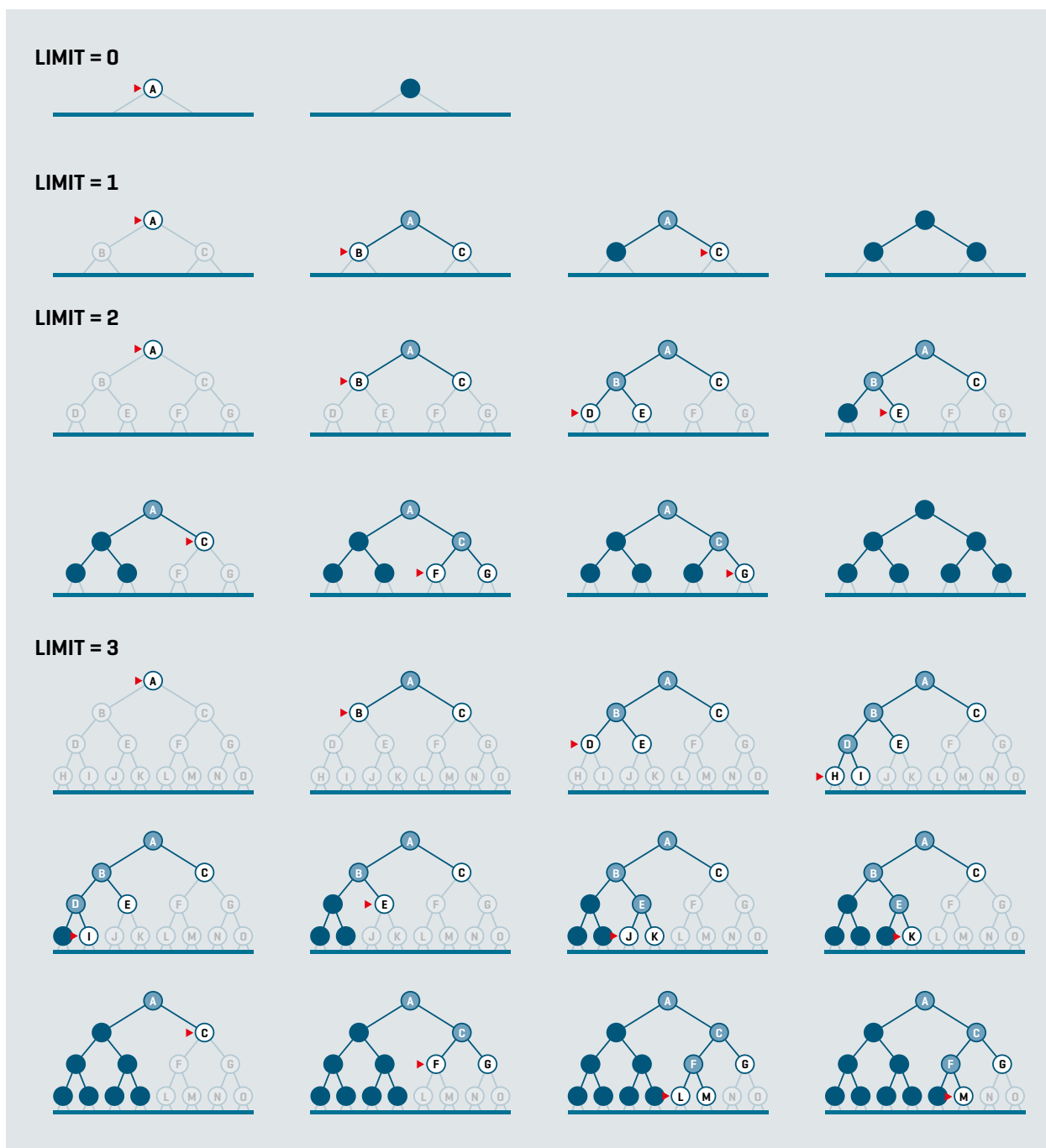


Figure 8: Example of iterative deepening search on a binary tree [3]

Uniform cost search (UCS) (or **Lower-cost-first search**) is a search strategy that aims at finding a solution with the lowest cost. It expands the nodes with the lowest path cost and is optimal for general step costs [3] [6].

3.2.3. Heuristic search

Searching for a solution without any prior knowledge of the environment could be time - and resource -consuming. In order to optimize the search and find solutions more efficiently, problem-specific knowledge (called heuristics) can be used. While exploring the graph using heuristics, it is possible for the search algorithm to know which non-goal states are more promising. Search strategies using such problem-specific knowledge are called informed (or heuristic) search strategies [3].

The **heuristics** are formalized using a heuristic function $h(s)$ that provides the estimated cost of the cheapest path from the node with state s to a goal state (an estimation of future cost, as shown on Figure 9).

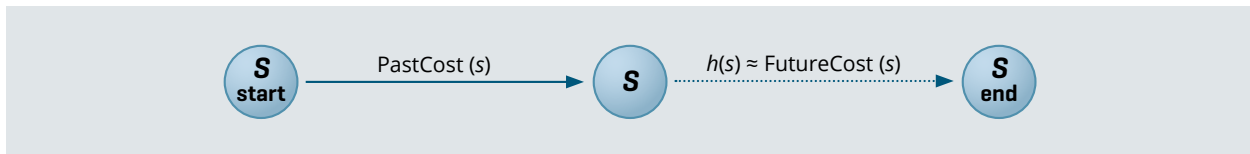


Figure 9: Illustration of heuristic function [2]

The heuristic has to satisfy the admissibility condition. A heuristic is admissible if the heuristic function $h(s)$ is always less than or equal to the actual cost of lowest-cost path from the current node to the goal. Thus, the admissible heuristic never overestimates the cost to reach the goal [2] [3].

Greedy best-first search is a search strategy that only takes into account the future cost to reach the goal. It selects a path on the frontier with the lowest heuristic value, meaning it explores first the nodes with minimal $h(s)$ [3].

A* search is a strategy that takes into account the total path cost $f(p)$ from the start node to the goal. The total path cost is composed of the cost $Cost(s)$ of the path found to reach the state and the heuristic function $h(s)$ estimating the future cost from the state to the goal (for all the paths on the frontier): $f(p) = Cost(s) + h(s)$. It expands nodes with minimal total path cost value [3] [6].

4. Machine learning

4.1. Introduction to supervised, unsupervised and reinforcement learning

In Chapter 1, three main components of machine learning were introduced: input data, learning process, and output data [8]. With respect to what input data a ML process receives and how it handles this data, three types of learning can be distinguished: supervised, unsupervised and reinforcement learning [9] [10] [11] [12].

- **Supervised learning:** input data is labeled to learn the relationship of given inputs to given outputs.

The learning is done on training data that contains input-output pairs. Once the relationship between inputs and outputs from the training data is learnt, it can be used to predict the outputs on new inputs (see Figure 10).

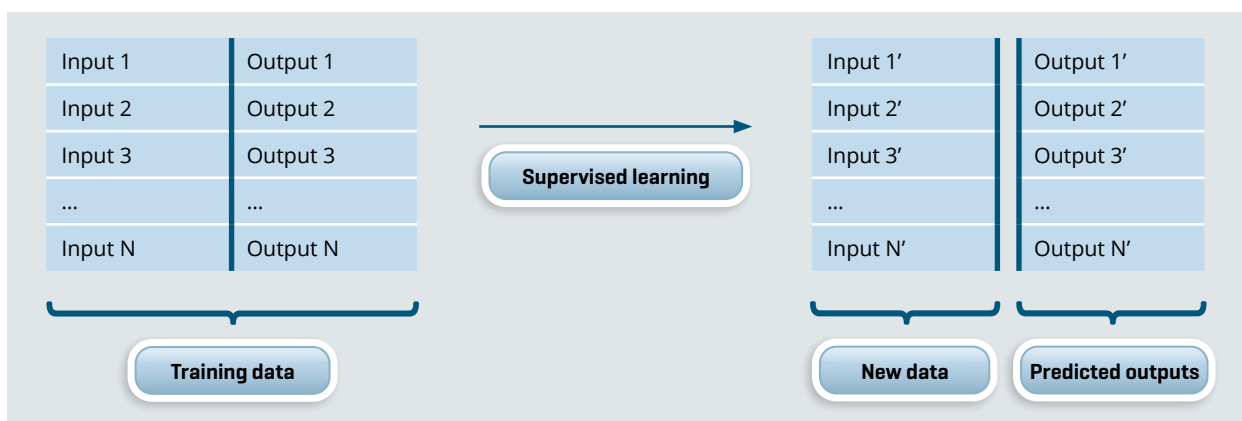


Figure 10: Supervised learning

- **Unsupervised learning:** input data is not labeled and its structure is discovered through the learning process.

The learning is meant to find out the structure in the input training data and identify groups of data points that exhibit similarities (see Figure 11).

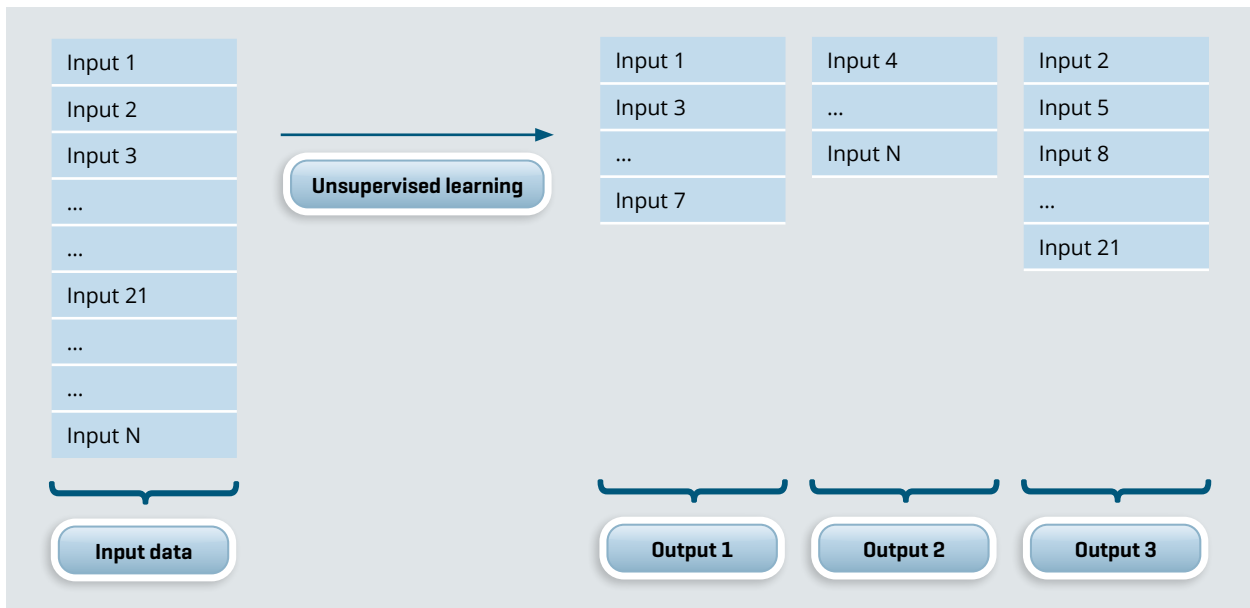


Figure 11: Unsupervised learning

- **Reinforcement learning:** does not receive input data as such but learns through the interaction with the environment where it performs tasks by trying to maximize rewards it receives for its actions [13].

Reinforcement learning can be described using the concept of agent that takes actions in the environment and receives feedback for its actions (see Figure 12). The environment in which the agent operates could be known or not. Each concrete and immediate situation in which agent finds itself is called a state. In order to reach its objective, the agent employs a strategy, or policy, to determine its next action based on its current state. The agent could receive positive or negative feedback for its actions, called a reward. The optimal strategy is the one that maximizes the reward [13].

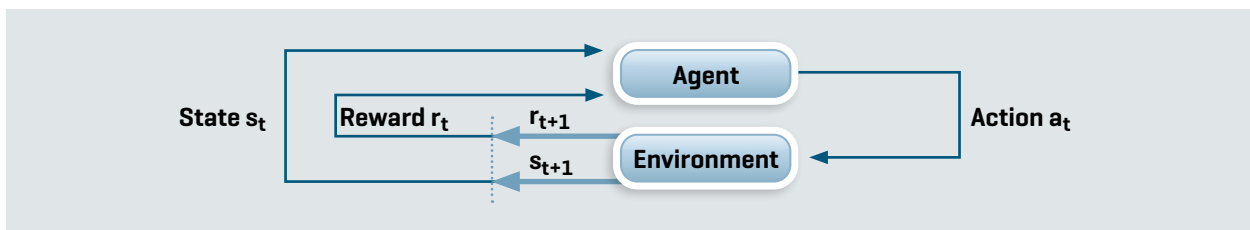


Figure 12: Agent-environment interaction [23]

Deep learning, which is a popular ML approach, can be used with any input data and thus could fall under any category of learning: supervised, unsupervised or reinforcement [13] [14]. This approach is based on neural networks that are introduced in Section 4.4.12. The section presents the variations of neural networks and provides examples of its usage.

4.2. Tasks solved by ML

Another way of classifying ML is with respect to the learning tasks and expected output (Figure 13 illustrates different categories of ML).

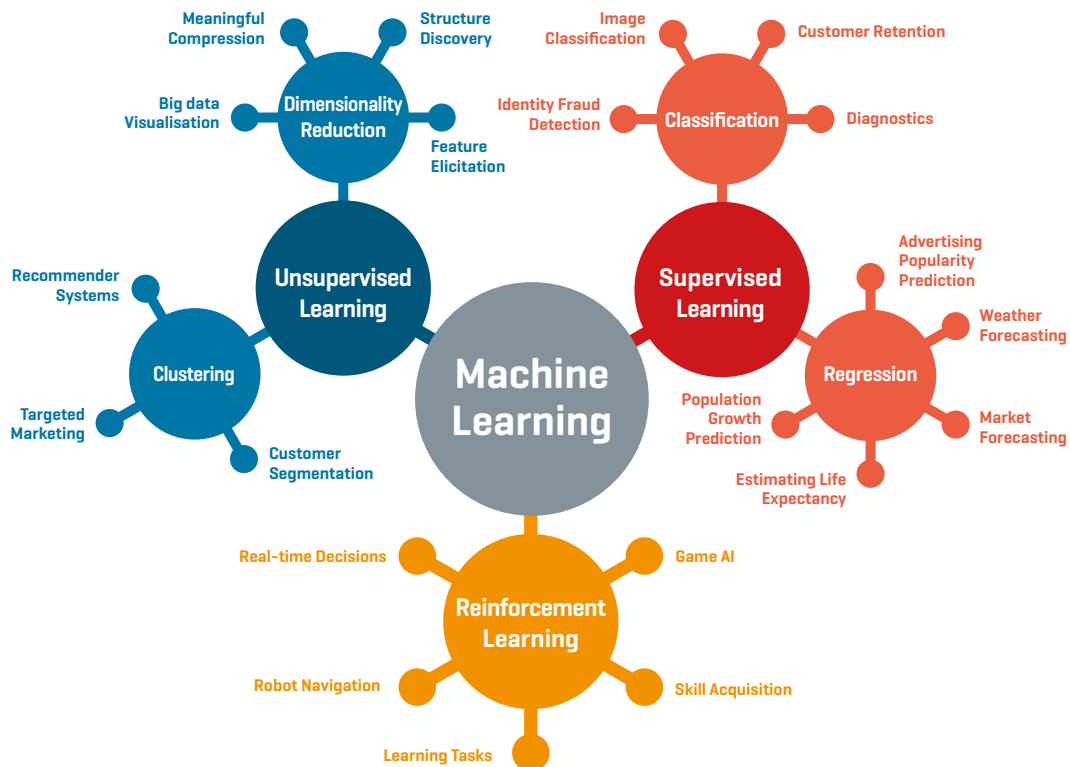


Figure 13: ML Categorization (adapted from [11])

The non-exhaustive list of tasks includes [9] [15] [16] [17] [18]:

- **Classification:** classification of the new inputs into the defined set of groups based on the training data,
- **Regression:** prediction of a continuous numeric value based on a number of input features,
- **Clustering:** identifying the inputs that bear similarities and grouping them into clusters,
- **Anomaly detection:** identification of unusual data records given the pattern(s) identified in the training set,
- **Summarization:** compact representation of data,
- **Ranking:** identifying a position on a scale based on some defined/learned feature(s)/criteria,
- **Recommendation:** a recommendation of an item(s) that would satisfy some condition(s) deduced from the training data,
- **Data generation:** generate novel data that would fit the pattern based on training data,
- **Optimization:** finding an optimal solution for a given problem in a given setting (game playing, space navigation, etc.).

For each of these tasks there may be a few ML algorithms. For example, one could distinguish between crispy classification, where the output is the label of a class to which the input belongs, and the probabilistic classification, where the output is a probability of an input to belong to each class [19]. Table 1 provides a mapping of some common algorithms to the ML tasks they are used to solve.

ML task	ML algorithms
Classification	Logistic regression, Support Vector Machine (SVM), decision tree, random forest, k-nearest neighbors (kNN), Convolutional Neural Networks (CNN), naïve Bayes classifier, linear/quadratic discriminant analysis
Regression	Linear regression, Neural Networks, SVM, kNN
Clustering	K-means, Hidden Markov Models (HMM)
Anomaly detection	kNN, SVM, Neural Networks
Summarization	Principal Component Analysis (PCA)
Ranking	RankSVM, Neural Networks
Recommendation	Collaborative filtering, association rule, kNN, matrix factorization
Data generation	Generative adversarial networks, HMM
Optimization	Reinforcement learning (Q-learning, AlphaZero)

Table 1: Mapping of ML algorithms to ML tasks [16] [17] [18] [20]

4.3. ML process and components

4.3.1. Supervised and unsupervised machine learning

ML is used to solve business problems. In this respect, the **ML process** is business-driven and thus relies on a business understanding and definition of the problem in the beginning and a business validation at the end of the process. As a result of a ML process, a **ML model** is created, that can be defined as a computational structure that generates an inference or prediction based on input data. In the phase of industrialization, when ML is used in production, the need to update the ML model (for example, because it does not perform well on new input data) or archive/terminate the model would also be motivated by business needs [8] [16]. This section focuses on the steps of creation of ML model that come after the business problem definition and before the business validation and deployment.

As of yet, there is no standard definition of ML process. The related challenges are highlighted in ISO/IEC TR 24028 *Information technology – Artificial Intelligence (AI) – Overview of trustworthiness in Artificial Intelligence*⁴⁴. More efforts to describe the standard ML process are being made by ISO/IEC JTC 1/SC 42/WG 1 on *Foundational standards*, namely through the development of future standard ISO/IEC 23053 *Framework for Artificial Intelligence (AI) Systems Using Machine Learning (ML)*⁴⁵. Generally, the following steps are considered for **ML model creation**, as shown on Figure 14 [8] [16] [21]:

44 <https://www.iso.org/standard/77608.html>

45 <https://www.iso.org/standard/74438.html>

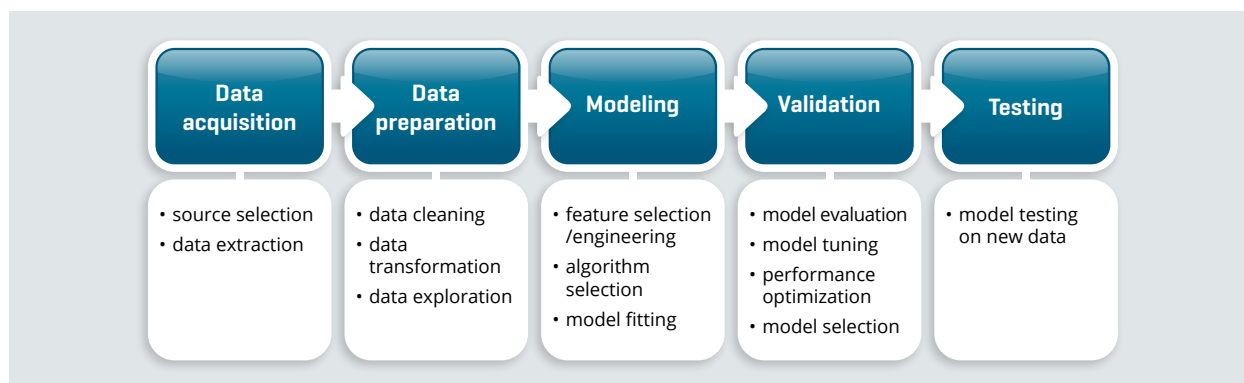


Figure 14: Model creation process - main steps

Identifying the data sources and acquisition of relevant data to solve the problem is the first step of ML model creation. To be able to build a quality ML model, it is necessary to guarantee the good quality of the input data (check for inconsistencies, formatting, remove duplication, etc.) which is done at the data preparation step.

For supervised machine learning it is common to split the input data into three subsets: a **training set** (used for modeling), a **validation set** (used for evaluation and validation) and a **test set** (used for testing). As a next step, the raw data is transformed into features, the input variables that will be used by the model to make predictions. Then, the learning algorithm is selected. Applying the algorithm to the training data is called model training or model fitting.

Some models need additional **hyper-parameters** (for example, the depth of a decision tree, the number of layers in a neural network, etc.) that are selected after training the model, during model tuning.

While tuning the model, the model variants using different hyper-parameters are evaluated on the validation dataset that is different from the one on which the model was trained. The selection of the model is based on its performance compared to other model variants.

After that, the testing data set is used to guarantee that the final selected model would perform well on unseen data.

In order to evaluate a **model's performance** (how well the model fits the data), different methods exist for regression and classification tasks. For regression tasks, the loss function is defined. The loss function calculates the difference between the predicted value and the real data value. Typical loss functions for supervised learning are shown in Figure 15. To assess the performance of the model, a cost function is defined as a sum over the values of a loss function for the entire dataset: the lower the cost function, the more fitting the model [22]. Examples of cost functions are the Mean Squared Prediction Error (MSPE), R-squared, etc.

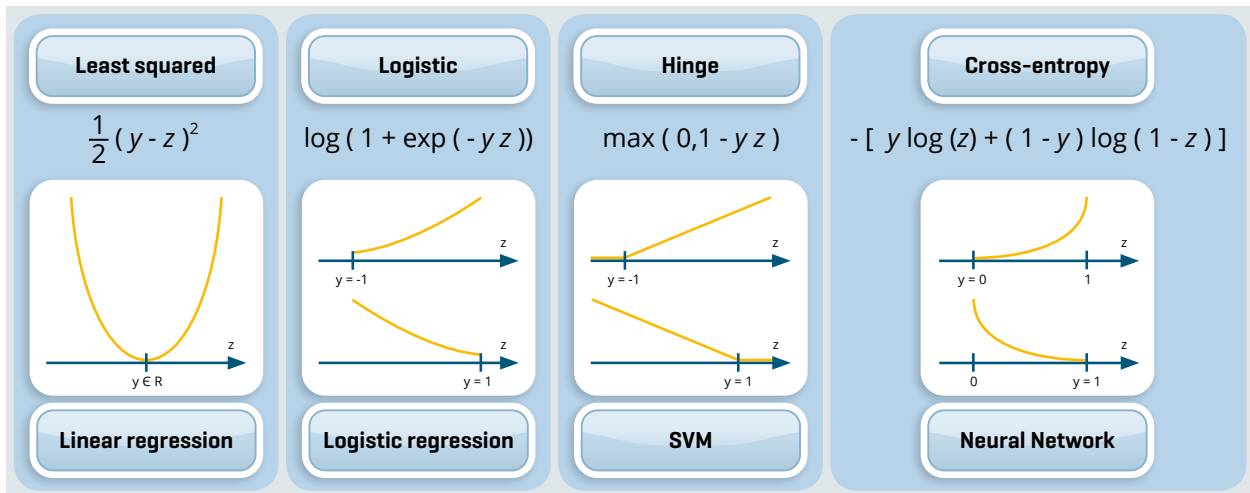


Figure 15: Typical loss functions [22]

In the case of classification problems, statistical metrics are used to evaluate model performance. Common statistical metrics, such as accuracy, precision, F1-score, etc. (some of which are listed in Table 3), are based on a Confusion Matrix that takes into account the predicted class labels and the actual class labels (explained in Table 2) [16].

From a standardization point of view, special attention is paid to the performance of neural networks. A dedicated technical report ISO/IEC TR 24029-1 *Artificial Intelligence (AI) – Assessment of the robustness of neural networks – Part 1: Overview*⁴⁶, where the robustness is defined as the ability of an AI system to maintain its level of performance under any condition, introduces various statistical, formal and empirical methods to assess the robustness of neural networks. Its second part ISO/IEC 24029-2 *Assessment of the robustness of neural networks – Part 2: Methodology for the use of formal methods*⁴⁷ considers specifically on the role of formal methods in the robustness of the neural networks. Moreover, there is a standard under development ISO/IEC 4213 *Artificial Intelligence – Assessment of machine learning classification performance*⁴⁸ that will specify the model-agnostic methodology for comparing the classification performance of ML models.

		Predicted label	
		Positive	Negative
Actual label	Positive	True Positive (TP)	False negative (FN)
	Negative	False positive (FP)	True negative (TN)

Table 2: Confusion matrix

Metric	Definition	Formula
Accuracy	Percentage of predictions that are correct	$(TP+TN)/(TP+TN+FP+FN)$
Precision	Percentage of positive predictions that are correct	$TP/(TP+FP)$
Recall (sensitivity)	Percentage of positive labels that were predicted as positive	$TP/(TP+FN)$
Specificity	Percentage of negative labels that were predicted as negative	$TN/(TN+FP)$
F1-Score	Trade-off between precision and recall	$2 * \text{precision} * \text{recall} / (\text{precision} + \text{recall})$

Table 3: Statistical metrics for the evaluation of model performance

46 <https://www.iso.org/standard/77609.html>

47 <https://www.iso.org/standard/79804.html>

48 <https://www.iso.org/standard/79799.html>

In the case of unsupervised learning, there is no division of data into training, validation and testing sets. The model creation is done by iterating over the data points in the input data, and then the same data is used to make inferences [21].

4.3.2. Reinforcement learning

The reinforcement learning process is different from supervised or unsupervised learning. It can be described using the concept of an **agent** that learns and make decisions, and an **environment** that represents everything outside the agent [23].

The agent has a task, which it tries to accomplish by interacting with the environment. The interactions of the agent with the environment constitute a sequence of steps. At each step, the agent perceives the current **state** \mathbf{s} of the environment and makes a decision about what **action** \mathbf{a} it should take next. After taking that action, the agent receives a **reward** \mathbf{r} from the environment, which could be any real number, and finds itself in a new state \mathbf{s}' [23].

The agent may or may not have a model of its environment. If the agent has (or has learned) a model of the environment, it can predict the state transitions and rewards to be received. However, providing a true and unbiased model to the agent can be a challenging task. Thus, it is common to distinguish between model-free (such as Q-learning) and model-based (such as AlphaZero) reinforcement learning algorithms [24].

While trying to accomplish its task, the goal of the agent is to maximize the cumulative reward, called **return** (or utility). The return is a sum of all the rewards received by the agent. While the task is not accomplished, the agent tries to maximize its expected return. The expected return is defined as the immediate reward the agent receives for its action and the future rewards it expects to receive. In order to favor immediate rewards over future rewards, a **discount factor** γ can be introduced. Ranging from 0 to 1, the discount factor is multiplied by future rewards making future rewards worth less than immediate rewards [13] [23].

In order to select the action (to take) based on its current state, the agent follows a rule, called the **policy** π . Reinforcement learning methods vary with respect to how the policy is updated as a result of the agent's experience [23] [24].

In order to evaluate how good the policy is, value functions are introduced. Value functions allow measuring the expected return under a given policy. There are two types of value functions: **the state-value** $v_{\pi}(\mathbf{s})$ and the **action-value** $q_{\pi}(\mathbf{s}, \mathbf{a})$. The state-value function $v_{\pi}(\mathbf{s})$ calculates for a state \mathbf{s} an expected return from that state given that the agent uses the policy π . The action-value policy function $q_{\pi}(\mathbf{s}, \mathbf{a})$ is similar to the state-value function. However, since policy defines a particular way of acting, the action-value function $q_{\pi}(\mathbf{s}, \mathbf{a})$ calculates the return of taking action \mathbf{a} in a state \mathbf{s} under the policy π . The functions that assign the largest expected return achievable by any policy are called optimal state or action-value functions. The optimal action-value function is denoted q^* , and this is where the name **Q-learning** comes from [23].

4.4. ML algorithms and their business application

Having described the ML process, this section introduces some popular ML algorithms. For each algorithm, a short explanation and a few business problems illustrating its application are provided.

4.4.1. Linear regression

Type of learning	Supervised
Tasks being solved	Regression (continuous value prediction)
Explanation	<p>Linear regression is a highly interpretable, standard method for modeling the past relationship between independent input variables (features) and dependent output variables (which can have an infinite number of values). As a result, the model is used to predict values of the output variables for new input data.</p> <p>Figure 16 bellow illustrates a simple relationship between one input variable and one output variable, for example the evolution over time (x-axis) of the average price of a house (y-axis). The red points are the original data points and the blue line is the learned model, according to which the estimation of house price could be obtained for the future.</p> <div data-bbox="475 958 1385 1447" data-label="Figure"> </div> <p>Figure 16: Linear regression⁴⁹</p> <p>NOTE: There is an alternative to linear regression, called non-linear regression, which would fit the data points not with a straight line but with a curve (such as a quadratic, sinusoidal, or other non-linear function).</p>
Sample business use case	<ul style="list-style-type: none"> • Understand product-sales drivers such as competition prices, distribution, advertisement, etc. • Optimize price points and estimate product-price elasticities

Table 4: Linear regression [12]

49 Source: <https://www.geeksforgeeks.org/linear-regression-using-tensorflow/>

4.4.2. Logistic regression

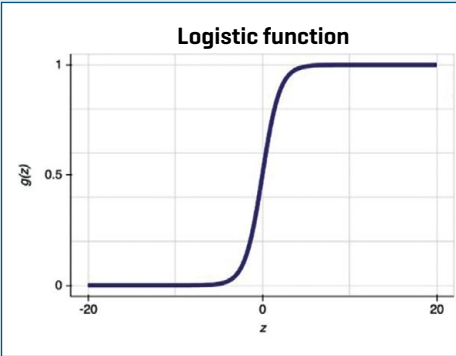
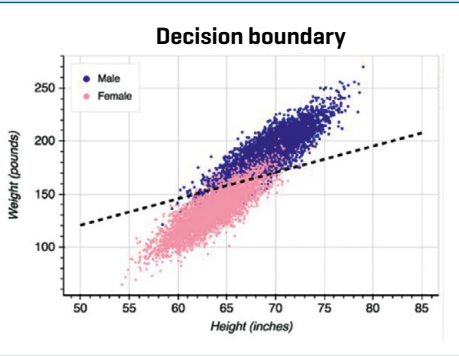
Type of learning	Supervised
Tasks being solved	Classification
Explanation	<p>Logistic regression is an extension of linear regression used for classification tasks. In this case, the output variable values are not continuous (like in linear regression) but discrete. The classification could be binary (with only two classes) or multi-class (more than two classes). Logistic regression uses likelihood to establish a boundary and a logistic function to give a probability of data points to belong to one class or the other.</p> <p>Figure 17 shows an example of binary classification: predicting a person’s gender (male/female) based on two input variables, weight (y-axis) and height (x-axis). On the right side there are input data points and an extracted decision boundary and on the left side there is a logistic function (the boundary is at 0,5).</p> <div style="display: flex; justify-content: space-around;">   </div> <p style="text-align: center;"><i>Figure 17: Logistic function (on the left) and logistic regression classifier (on the right)⁵⁰</i></p>
Sample business use case	<ul style="list-style-type: none"> ● Classify customers based on how likely they are to repay a loan ● Predict if a skin lesion is benign or malignant based on its characteristics (size, shape, color, etc.)

Table 5: Logistic regression [12]

50 Source: <https://towardsdatascience.com/understanding-logistic-regression-step-by-step-704a78be7e0a>

4.4.3. Decision tree

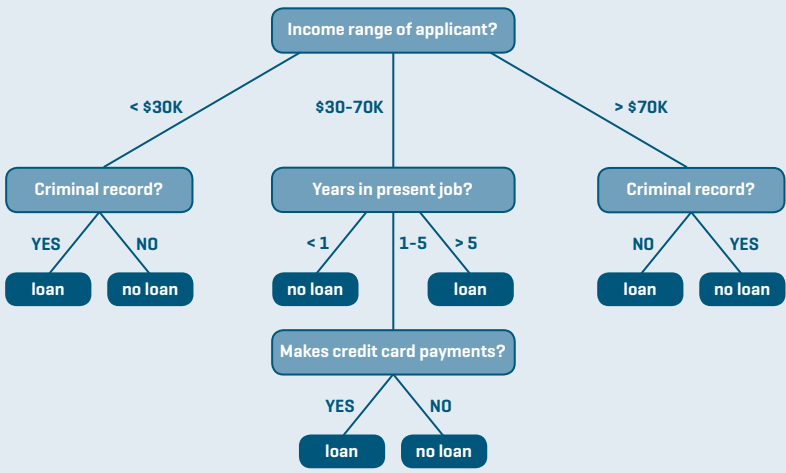
Type of learning	Supervised
Tasks being solved	Classification, Regression
Explanation	<p>A decision tree algorithm is a highly interpretable classification or regression algorithm. It splits the input data into subgroups based on the values of the input features (for example, if the feature is “color”, each color from the dataset could become a basis for the split). Each split forms a decision node in the tree, and the sequence of splits gives rise to a branch of the tree. At the end of each branch, there are leaves that are the labels of the class (in case of classification) or some numerical values (in case of regression) corresponding to the input data points.</p> <p>Figure 18 presents a decision tree for a money loan (with two possible decisions: loan or no loan) based on the input features such as income range, existence of a criminal record, number of years working for the current employer, and credit card payments.</p>  <p style="text-align: center;"><i>Figure 18: Decision tree⁵¹</i></p>
Sample business use case	<ul style="list-style-type: none"> ● Provide a decision framework for hiring new employees ● Understand product attributes that make a product most likely to be purchased

Table 6: Decision Tree [12]

51 Source: <https://towardsdatascience.com/what-is-a-decision-tree-22975f00f3e1>

4.4.4. Random forest

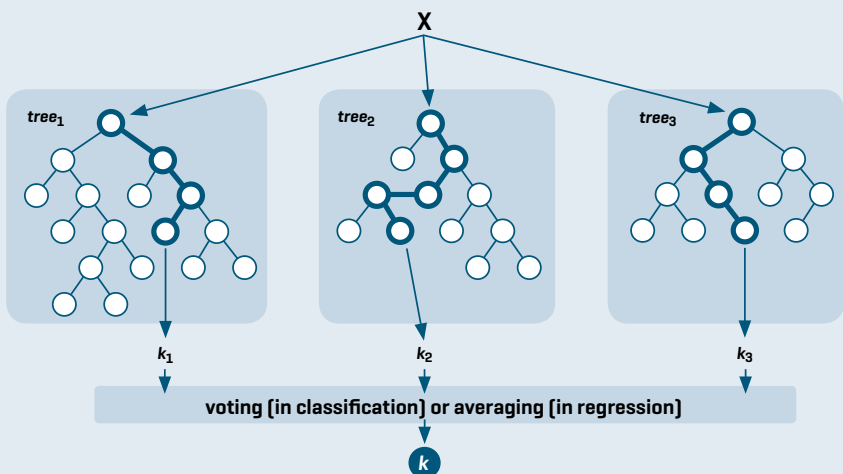
Type of learning	Supervised
Tasks being solved	Classification, Regression
Explanation	<p>Random forest is classification or regression algorithm that improves the accuracy of a simple decision tree by generating multiple decision trees and taking a majority vote of them to predict the output, which could be a class label (for example, black, white, or red) for classification problems or an averaged numerical value (such as age) for a regression problem.</p> <p>Figure 19 provides a schematic explanation of how a random forest algorithm works, presenting a forest of three trees.</p>  <p style="text-align: center;"><i>Figure 19: Random Forest⁵²</i></p>
Sample business use case	<ul style="list-style-type: none"> ● Predict call volume in call centers for staffing decisions ● Predict power usage in an electrical-distribution grid

Table 7: Random forest [12]

52 Adapted from: https://www.researchgate.net/figure/Architecture-of-the-random-forest-model_fig1_301638643

4.4.5. Naïve Bayes classifier

Type of learning	Supervised
Tasks being solved	Classification
Explanation	<p>Naïve Bayes is a probabilistic technique for constructing classifiers. It applies Bayes' theorem, which allows calculating the probability of an event based on the knowledge of factors that may affect it. For example, if an email contains the word "money" (an affecting factor) then the probability of it being spam (the event) is high. An advantage of naïve Bayes is that it only requires a small amount of training data to estimate the necessary probabilities and that the classifier can be trained incrementally.</p> <p>Figure 20 presents the Bayesian rule (on the left), where $P(C X)$ is the probability of the event C given the affecting factor X. On the right, the classification of a new data point using Naïve Bayes Classifier is explained. The example has input data points of two classes, green and yellow. Based on the number of data points of each color and the total number of points in the dataset it is possible to estimate the probability of any data point to belong to one class or another. To classify the new data point, its vicinity is considered (within black circle). Given the number of data points of each class in the vicinity and the general probability of having a data point of this respective class, the probability of the new data point to belong to each class is calculated. The highest probability is retained to classify the new data point.</p> <div style="border: 1px solid black; padding: 10px; margin: 10px 0;"> <div style="display: flex; justify-content: space-between;"> <div style="width: 45%;"> <p style="text-align: center;">Likelihood Class Prior Probability</p> $P(c x) = \frac{P(x c)P(c)}{P(x)}$ <p style="text-align: center;">Posterior Probability Predictor Prior Probability</p> $P(c X) = P(x_1 c) \times P(x_2 c) \times \dots \times P(x_n c) \times P(c)$ </div> <div style="width: 50%;"> <p style="text-align: right;"> $P(\text{yellow}) = \frac{7}{17}$ $P(\text{green}) = \frac{10}{17}$ $P(? \text{green}) = \frac{3}{10}$ $P(? \text{yellow}) = \frac{1}{7}$ </p> <p style="text-align: right;"> <small>prior probabilities number of samples in a given class divided by the total number of samples</small> <small>we consider just the vicinity of the new sample we want to classify</small> </p> <p style="text-align: right;"> <small>posterior probability</small> $P'(? \text{ is green}) = P(\text{green}) * P(? \text{green}) = \frac{10}{17} * \frac{3}{10} = \frac{30}{170}$ <small>posterior probability</small> $P'(? \text{ is yellow}) = P(\text{yellow}) * P(? \text{yellow}) = \frac{7}{17} * \frac{1}{7} = \frac{7}{119}$ </p> </div> </div> </div> <p style="text-align: center;"><i>Figure 20: Bayes Rule (on the left)⁵³ and Naïve Bayes Classifier (on the right)⁵⁴</i></p>
Sample business use case	<ul style="list-style-type: none"> Analyze sentiment to assess product perception in the market Create classifiers to filter spam emails

Table 8: Naïve Bayes Classifier [12] [25]

53 Source: http://uc-r.github.io/naive_bayes

54 Source: <https://www.ijrte.org/wp-content/uploads/papers/v8i3/C4258098319.pdf>

4.4.6. Support Vector Machine [SVM]

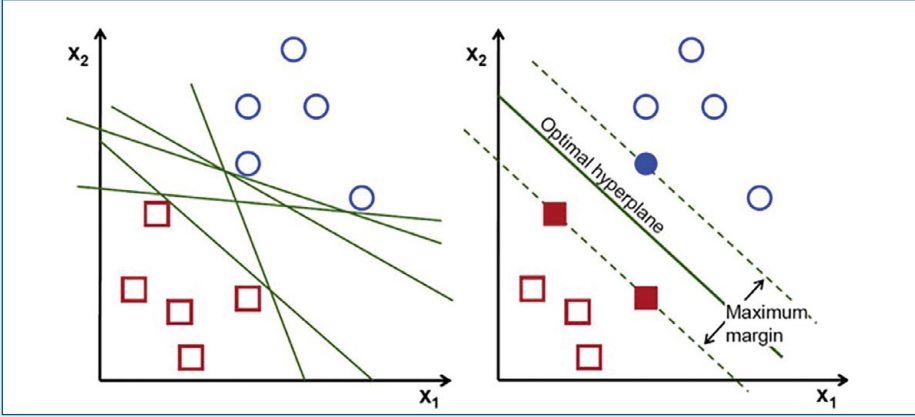
Type of learning	Supervised
Tasks being solved	Classification, Regression, Anomaly detection
Explanation	<p>Support Vector Machine (SVM) is typically used for classification but can be transformed to perform regression. It is typically used with a linear model but can be generalized to solve non-linear problems. SVM is used to draw an optimal division line (or plane, depending on the number of input features) between classes so that it has the maximum possible distance to the nearest data points. The lines passing through these nearest data points are called support vectors. With the evolution of the input data set and introduction of new data points the support vectors as well as the optimal class separation line could change.</p> <p>Figure 21 shows the classification using SVM. The image on the left presents various possible separation lines (in green) between two classes of data points. The image on the right presents an optimal separation line (in green) and the support vectors (dashed lines).</p>  <p style="text-align: center;"><i>Figure 21: SVM⁵⁵</i></p>
Sample business use case	<p>Predict how many patients a hospital will need to serve in a time period</p> <p>Predict how likely someone is to click on an online ad</p>

Table 9: SVM [12]

55 Source: <https://gdcoder.com/support-vector-machine-vs-logistic-regression/>

4.4.7. K-Nearest Neighbors [kNN]

Type of learning	Supervised
Tasks being solved	Classification, Anomaly detection, Recommendation, Regression
Explanation	<p>K-Nearest Neighbors (kNN) is a non-parametric and instance-based learning algorithm that can be used for both classification and regression problems. The output value for the new data point is computed by averaging the values of k data points closest to it from initial/training dataset, called nearest neighbors. kNN can be useful in case of nonlinear data as well as when dealing with datasets that do not follow mathematical theoretical assumptions, since it does not rely on underlying data distribution to provide the output.</p> <p>Figure 22 illustrates the different steps of kNN. The first image from the left shows the initial data set with data points belonging to two classes (class A and class B) as well as a new data point that needs to be classified. The middle image shows the next step, where the distances to the data points of the initial data set are calculated in order to be able to identify the nearest neighbors. Finally, the last image shows different possibilities for the choice of the number of nearest neighbors (k=3 and k=7) highlighting the impact such a choice has on the classification of the new data point (class B in case of k=3 and class A in case of k=7).</p>
	<p>Figure 22: kNN⁵⁶</p>
Sample business use case	<ul style="list-style-type: none"> ● Predict the credit rating of customers ● Predict whether the loan is safe or risky ● Classifying potential voters in two classes “will vote” or “will not vote”

Table 10: kNN [26]

56 Adapted from source: <https://www.datacamp.com/community/tutorials/k-nearest-neighbor-classification-scikit-learn>

4.4.8. Linear and quadratic discriminant analysis

Type of learning	Supervised
Tasks being solved	Classification
Explanation	<p>Discriminant analysis is a classification technique based on the probabilistic distribution of input variables. There are two types of discriminant analysis, linear and quadratic. In case of linear discriminant analysis (LDA) the decision boundary between classes is linear, and in case of quadratic discriminant analysis (QDA) it is non-linear. When a new data point is presented to discriminant analysis, its class label is determined by the decision boundaries.</p> <p>Figure 23 illustrates the discriminant analysis in the case of three classes. The left-side image shows the probability distribution of the data points and the decision boundaries between classes. The right-side image provides a visualization of decision boundaries in two dimensions based on linear discriminant analysis (LDA) and quadratic discriminant analysis (QDA).</p>
Sample business use case	<ul style="list-style-type: none"> ● Predict client churn ● Predict a sales lead’s likelihood of closing

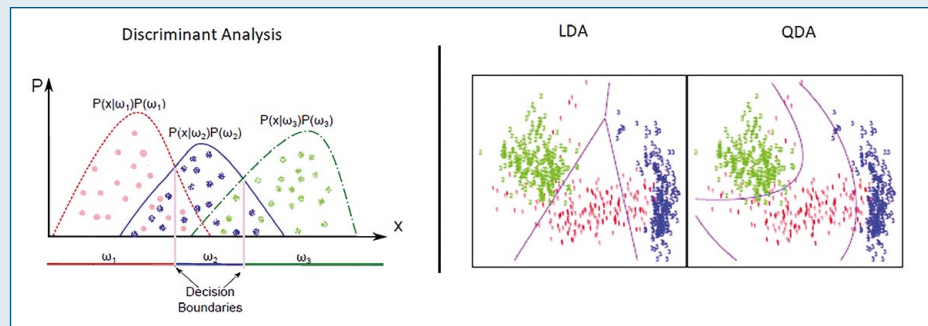


Figure 23: Linear and Quadratic Discriminant Analysis⁵⁷

Table 11: Linear and quadratic discriminant analysis [12] [27]

57 Adapted from source: https://www.researchgate.net/publication/308015273_Linear_vs_quadratic_discriminant_analysis_classifier_a_tutorial

4.4.9. K-means

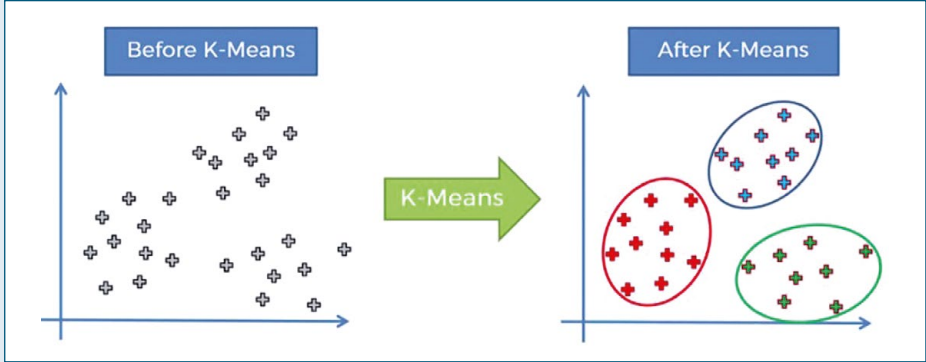
Type of learning	Unsupervised
Tasks being solved	Clustering
Explanation	<p>K-means puts data into a number (k) of groups such that each group contains data with similar characteristics. The characteristics are determined by the model during the training process. However, the number of groups k has to be defined in advance manually.</p> <p>Figure 24 presents the dataset (on the left side) and the outcomes of k-means clustering (on the right side), where $k=3$.</p>  <p style="text-align: center;"><i>Figure 24: k-means with $k=3$⁵⁸</i></p>
Sample business use case	<ul style="list-style-type: none"> Segment customers into groups by distinct characteristics (for example, age group) – for instance, to better assign marketing campaigns or prevent churn

Table 12: k-means [12] [25]

58 Source: <https://towardsdatascience.com/k-means-clustering-identifying-f-r-i-e-n-d-s-in-the-world-of-strangers-695537505d>

4.4.10. Principal component analysis

Type of learning	Unsupervised
Tasks being solved	Summarization (pattern detection and dimensionality reduction)
Explanation	<p>Principal component analysis (PCA) is a statistical method of identifying patterns in data and expressing the data in such a way as to highlight their similarities and differences. Once the patterns are detected, it is possible to compress the data, by reducing the number of dimensions, without much loss of information.</p> <p>Figure 25 shows the transformation of data using PCA. The image on the left presents original data with two input features, expressed along the x-and-y-axes. The image on the right depicts the data along the one “combined” feature.</p> <div style="text-align: center;"> </div> <p style="text-align: center;"><i>Figure 25: PCA [28]</i></p>
Sample business use case	<ul style="list-style-type: none"> ● Image compression and recognition based on patterns detection ● Stock portfolio construction and management

Table 13: Principal Component Analysis [28]

4.4.11. Hidden Markov model [HMM]

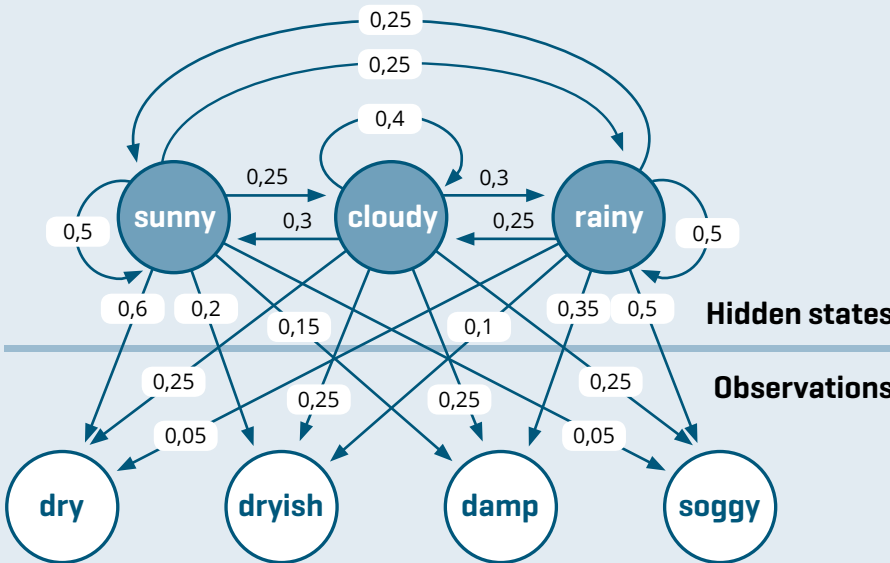
Type of learning	Supervised, unsupervised, reinforcement
Tasks being solved	Clustering, Data generation
Explanation	<p>Hidden Markov models (HMM) is a statistical method to capture hidden information (for example, parts of speech) from observable sequential symbols (for example, the sequences of words). In a HMM, a system is modelled through a set of states and transition probabilities between states, called a Markov process. Some of the parameters (states and/or transition probabilities) are unknown, and the challenge is to determine the hidden parameters from the observable ones. A good HMM accurately models the real world source of the observed real data and has the ability to simulate the source.</p> <p>Figure 26 presents the HMM for prediction of weather (unknown information) based on the level of humidity (observable information).</p>  <p>The diagram illustrates a Hidden Markov Model (HMM) for weather prediction. It consists of two layers: Hidden states and Observations.</p> <ul style="list-style-type: none"> Hidden states: sunny, cloudy, rainy. Observations: dry, dryish, damp, soggy. <p>Transition probabilities (between hidden states):</p> <ul style="list-style-type: none"> sunny to sunny: 0,5 sunny to cloudy: 0,25 sunny to rainy: 0,25 cloudy to sunny: 0,3 cloudy to cloudy: 0,4 cloudy to rainy: 0,3 rainy to sunny: 0,25 rainy to cloudy: 0,25 rainy to rainy: 0,5 <p>Emission probabilities (from hidden states to observations):</p> <ul style="list-style-type: none"> sunny to dry: 0,6 sunny to dryish: 0,2 sunny to damp: 0,15 sunny to soggy: 0,05 cloudy to dry: 0,25 cloudy to dryish: 0,25 cloudy to damp: 0,25 cloudy to soggy: 0,25 rainy to dry: 0,35 rainy to dryish: 0,1 rainy to damp: 0,25 rainy to soggy: 0,05
Sample business use case	<ul style="list-style-type: none"> ● Speech recognition and generation ● Predict exons and introns in genomic DNA

Figure 26: Hidden Markov Model⁵⁹

Table 14: Hidden Markov Model [29]

59 Source: <http://article.sciencepublishinggroup.com/html/10.11648.j.acm.s.2017060401.12.html>

4.4.12. Neural networks

Type of learning	Supervised, unsupervised, reinforcement
Tasks being solved	All
Explanation	<p>A neural network is a model that can be used to classify data or find the relationship between variables in regression problems. Such a network is built of artificial neurons that could be seen as software-based calculators. A simple neural network is built of three layers of artificial neurons: an input layer, a hidden layer, where calculations take place, and an output layer. Networks that are more complex have more hidden layers. Deep learning is a class of ML algorithms based on neural networks using multiple hidden layers.</p> <p>There are three common types of neural networks:</p> <ul style="list-style-type: none"> ● Feed Forward Neural Networks (FFNs) are one of the most common type of neural networks. In this architecture, information moves in only one direction, forward, from the input layer, through the “hidden” layers, to the output layer. The neurons in each hidden layer learn something simple and pass this information to the neurons of the next layer. FFNs are the most general-purpose type of neural networks. ● Convolutional Neural Network (CNN) is a multilayered neural network with a special architecture designed to extract increasingly complex features by first learning small, patterned features and then aggregating and re-identifying them. CNNs are well suited for image classification. ● Recurrent neural network (RNN) is a multilayered neural network that can store information in context nodes, allowing it to learn data sequences and output a number or another sequence. <p>Figure 27 illustrates these three types of neural networks. Note that many other types of neural networks exist.</p>
Sample business use case	<ul style="list-style-type: none"> ● Predict the probability that a patient joins a healthcare program ● Diagnose health diseases from medical scans ● Predict whether registered users will be willing or not to pay a particular price for a product ● Provide language translation

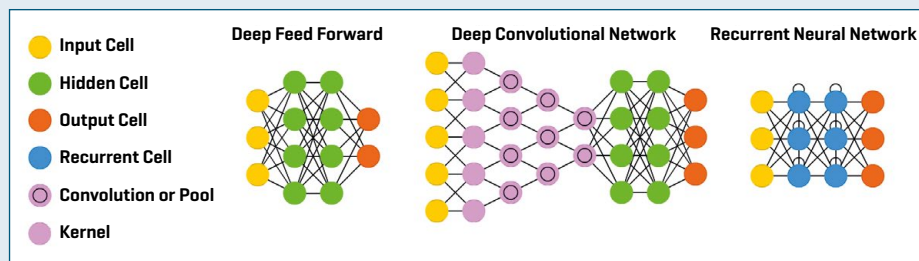


Figure 27: Different types of neural networks^{60, 61}

Table 15: Neural network [12] [30]

60 Source: <https://becominghuman.ai/cheat-sheets-for-ai-neural-networks-machine-learning-deep-learning-big-data-science-pdf-f22dc900d2d7>

61 Source: <https://towardsdatascience.com/the-mostly-complete-chart-of-neural-networks-explained-3fb6f2367464>

4.4.13. Q-learning

Type of learning	Reinforcement
Tasks being solved	Optimization
Explanation	<p>Q-learning is a model-free off-policy reinforcement learning algorithm. The algorithm seeks to maximize the total reward of the agent, which is estimated by an action-value function $q_{\pi}(s, a)$ (this is where the Q comes from). The algorithm is called model-free because it does not have access to the model of the environment. It is considered off-policy because the optimal action-value function q^* is learnt by taking the actions “randomly”, without them being prescribed by the policy. This is why Q-learning is not always efficient, but it has an advantage of being simple to implement.</p>
Sample business use case	<ul style="list-style-type: none"> ● Traffic light control ● Web System configuration ● Personalized recommendations

Table 16: Reinforcement learning [24] [31] [32]

4.4.14. AlphaZero

Type of learning	Reinforcement
Tasks being solved	Optimization
Explanation	<p>AlphaZero is a model-based reinforcement learning algorithm. It was originally designed to learn the games of Go, Chess and Shogi, teaching itself from scratch how to play those games. Using only the rules of a game as a model it iteratively learns and updates explicit representations of the policy. It uses searching techniques to select the most promising candidate moves out of the ones suggested by the policy, the latter being updated afterwards.</p> <p>The learning process is encoded using a deep neural network. Starting from a random move, the system adjusts the parameters of the neural network by learning from its wins and losses, thus increasing the probability of choosing the most advantageous moves in the future.</p> <p>Having shown the ability to master complex games, the algorithm can be successfully exploited to solve other problems, for example in engineering.</p>
Sample business use case	<ul style="list-style-type: none"> ● Master the games of Chess, Shogi and Go ● Assembling a car given the detached parts ● Simulating physical experiments/interactions

Table 17: AlphaZero [24] [33] [34]

5. Logic-based AI: knowledge representation and reasoning

Using **logic** in AI has a long history and was motivated and advocated for by one of the “fathers” of AI, John McCarthy, as early as 1959. One motivation for using logic was that a logical formalization could help to understand the reasoning process itself. The long-term objective was to **formalize common sense reasoning** in order to solve everyday problems, such as diagnosis, spatial reasoning, reasoning about the attitudes of other agents, etc. [35].

Applications of logic-based AI currently include diagnostics, be it in the medical, manufacturing or some other field, monitoring systems, process control, information extraction, etc. It can play a role of advising, interpreting inputs, predicting and explaining results, justifying a conclusion, suggesting an alternative solution, etc. [36]. Logic-based AI can be successful where ML fails, for example, in mathematical reasoning [37]. Moreover, as compared to ML, it has an advantage of being highly interpretable [38].

As discussed in Chapter 1, the early successful implementations of logic-based AI were the **expert systems**, composed of knowledge database and an inference engine. More recent applications are the knowledge databases for graph-based representation formats, or **knowledge graphs** [39]. Generally speaking, three uses of logic in AI can be distinguished: as a tool of analysis, as a basis for knowledge representation and as a programming language [32]. Since the last one is more on the implementation side, the focus of this section is on the theoretical grounds for the first two.

The section starts with an introduction of different logical paradigms. Then, various knowledge representation schemes are presented. Finally, the reasoning (inference) engine is discussed showing the advantages and limitations of logic-based AI.

5.1. Overview of logical paradigms

One definition of logic in Collins dictionary states that logic is “*any particular formal system in which are defined axioms and rules of inference*”⁶². Based on the **axioms**, more complex statements can be constructed. In order for the statements to be correct, they have to respect the **syntactic rules** of the logic. And in order for the statements to convey a message their meaning should be understood, that is the **semantics** must also be defined in the logic [3]. There is a famous example in the area of natural language processing, which shows the importance of semantics by providing a syntactically correct English sentence that is completely meaningless: “*Colorless green ideas sleep furiously*”⁶³.

The semantics allows defining which sentences are true and which are not in a given model of the world [3]. For example, in the world of Irrational numbers $\sqrt{2}$ is valid, while it is not in the world of Integers or even Rational numbers. The logical systems vary with their syntactic rules, but also with their semantics, which should be well adjusted to capture the complexity of the world they purport to describe.

One very simple logic is called **propositional logic**. It deals with sentences describing facts about the world [3]. But it cannot describe the individual entities in the world or their relationships. This can be done in description or in first-order logic. The former is considered as a subset of the latter but at the same time allows for the expression of some different relations. **Description logic** is widely used in designing the ontologies and the Semantic Web, upon which the search engines operate. **First-order logic** builds on propositional logic. It adds terms to represent objects, and has universal and existential quantifiers to construct assertions about all or some

of the possible values of the quantified variables [3].

Propositional, description and first-order logics can be called classical logics. Given such a variety of logics, the standard ISO/IEC 24707:2018 *Information technology — Common Logic (CL) — A framework for a family of logic-based languages*⁶⁴ was developed by the international technical committee ISO/IEC JTC 1/SC 32 *Data management and interchange*⁶⁵. As specified in the scope of this standard, it “*defines an abstract syntax and an associated model-theoretic semantics for a specific extension of first-order logic. The intent is that the content of any system using first-order logic can be represented in this document. The purpose is to facilitate interchange of first-order logic-based information between systems*”.

Among other well-known standards, there is **Web Ontology Language (OWL)** used for knowledge representation and ontologies authoring⁶⁶. OWL is developed by the W3C consortium⁶⁷. How the Common Logic (CL) connects to other logics is shown in Figure 28.

It is interesting to note that the international technical committee ISO/IEC JTC 1/SC 32 has also been making an effort to define terms and relations of some widely-used upper-level ontologies and specify them in both CL and OWL to support the implementation. The result of this effort is a multipart standard entitled ISO/IEC 21838 *Information technology — Ontologies — Top-Level Ontologies (TLO)*, of which:

- Part 1: Requirements - defines the criteria for a top-level ontology⁶⁸;
- Part 2: Basic Formal Ontology (BFO) – describes BFO, a TLO that is domain-neutral and thus provides terms representing only highly general categories which pertain to all domains of reality⁶⁹;
- Part 3: Descriptive ontology for linguistic and cognitive engineering (DOLCE) – describes DOLCE, a TLO that captures ontological categories found in natural language and human common sense⁷⁰;
- Part 4: TUpper – describes TUpper, a TLO that considers an upper ontology to be a modular ontology composed of generic ontologies that cover concepts including those related to time, process, and space⁷¹.

64 <https://www.iso.org/standard/66249.html>

65 <https://www.iso.org/committee/45342.html>

66 https://en.wikipedia.org/wiki/Description_logic

67 <https://www.w3.org/TR/owl-ref/>

68 <https://www.iso.org/standard/71954.html>

69 <https://www.iso.org/standard/74572.html>

70 <https://www.iso.org/standard/78927.html>

71 <https://www.iso.org/standard/78928.html>

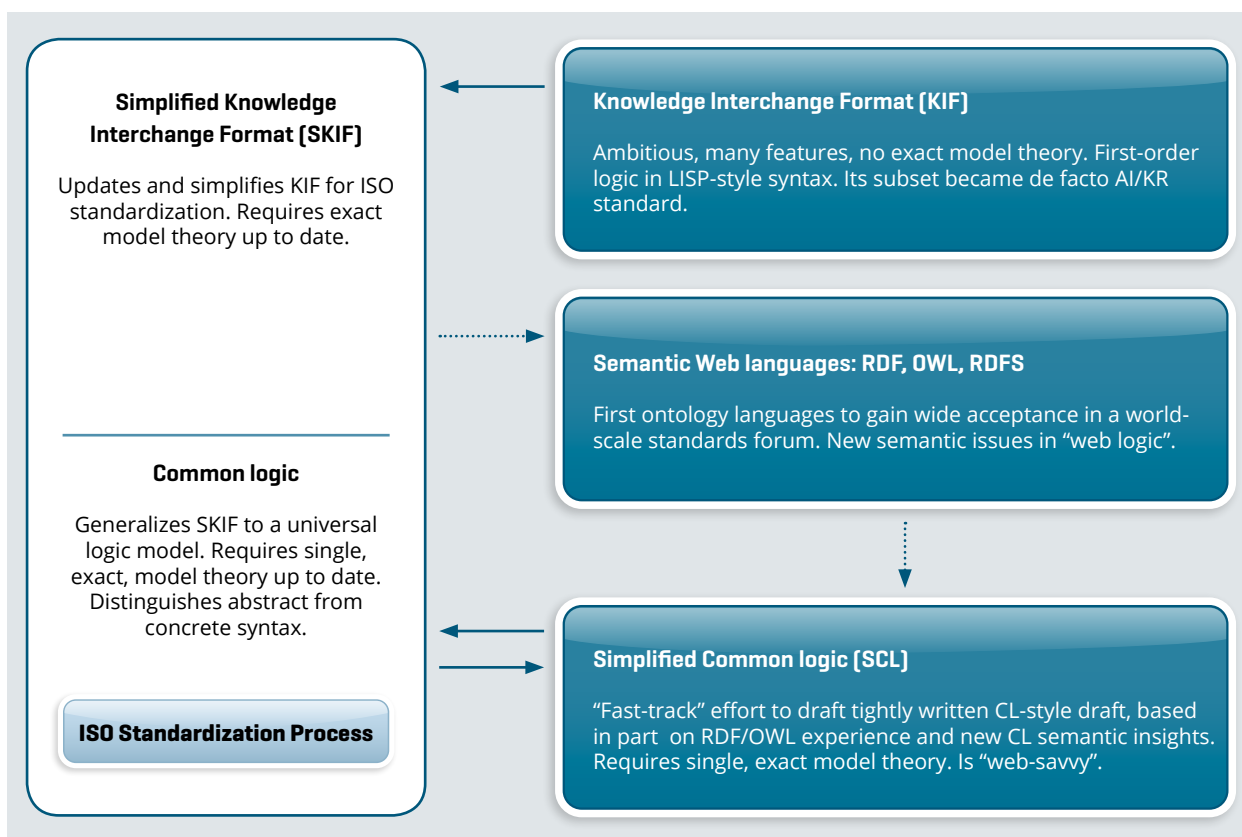


Figure 28: Common Logic and its relation to other logics (adapted from <https://www.w3.org/2004/12/rules-ws/slides/pathayes.pdf>)

One of the limitations of classical logics is that for each input query they can only output **true** or **false**. However, it is not always straightforward to attribute such an output. **Fuzzy logic** was invented to be able to output more values (any real value from 0 to 1) and to describe partial truths⁷². If one considers the temperature and the scale cold-warm-hot, it is practical to use fuzzy logic to describe the “warm” concept, which reflects a certain degree between cold and hot.

Moreover, classical logics are monotonic, meaning that the number of outputs could only increase as information is added to the knowledge base [3]. **Monotonic logics** have difficulties handling exceptions, reasoning about uncertain situations, not including all the knowledge in the reasoning even if the probability of some event is close to zero, excluding some situations base on newly received inputs, etc. In order to deal with these limitations, **non-monotonic logics** were introduced [35].

For the purposes of this white paper, the following two sections will focus on propositional (Section 5.2.1) and first-order logic (Section 5.2.2), providing some intuition behind non-monotonic logic (Section 5.3).

⁷² https://en.wikipedia.org/wiki/Fuzzy_logic

5.2. Knowledge representation and interpretation

5.2.1. Propositional logic

As explained in the previous section, **propositional logic** provides a means of expressing **facts**. It builds upon atomic sentences. Each atomic sentence consists of a single proposition symbol. There are two propositional symbols with fixed meanings (true and false). Other propositional symbols can be interpreted as true or false depending on the world's model. More complex sentences are built from atomic sentences using syntactic rules. These rules are based on five logical connectives [3]:

- Negation \neg meaning "not",
- Conjunction \wedge meaning "and",
- Disjunction \vee meaning "or",
- Implication \implies meaning "implies",
- Biconditional \iff meaning "if and only if".

The syntax of propositional logic is captured by Figure 29.

```

Sentence  $\rightarrow$  AtomicSentence | ComplexSentence
AtomicSentence  $\rightarrow$  True | False | P | Q | R | ...
ComplexSentence  $\rightarrow$  ( Sentence ) | [ Sentence ]
                |  $\neg$  Sentence
                | Sentence  $\wedge$  Sentence
                | Sentence  $\vee$  Sentence
                | Sentence  $\implies$  Sentence
                | Sentence  $\iff$  Sentence
OPERATOR PRECEDENCE:  $\neg$ ,  $\wedge$ ,  $\vee$ ,  $\implies$ ,  $\iff$ 
```

Figure 29: Syntax in Propositional Logic [3]

Each sentence has a meaning. In propositional logic, the meaning of any sentence could be either **true** or **false**. The interpretation of atomic propositional symbols depends on the model of the world. The interpretation of complex sentences is defined by semantic rules allowing to compute the value of any sentence given a model. These rules can be expressed in what is called truth tables that provide the truth value of a complex sentence for each possible combination of values of its constituents. A **truth table** for the interpretation of complex sentences based on five syntactic rules is presented in Figure 30 [3].

P	Q	$\neg P$	$P \wedge Q$	$P \vee Q$	$P \implies Q$	$P \iff Q$
false	false	true	false	false	true	true
false	true	true	false	true	true	false
true	false	false	false	true	false	false
true	true	false	true	true	true	true

Figure 30: Truth table for the interpretation of sentences in propositional logic [3]

The knowledge base in propositional logic is built of sentences for which the interpretation is true in all possible models of the world [2]. One example of knowledge representation using propositional logic could be a list of possible health problems (fever, runny nose, headache, cough, etc.) and the corresponding illnesses (as shown in Figure 31).

IF fever is yes	IF fever is no
AND cough is yes	AND cough is yes
AND runny nose is yes	AND runny nose is yes
AND headache is yes	THEN diagnosis is cold
THEN diagnosis is flu	

Figure 31: Example of a simple Knowledge base⁷³

5.2.2. First-order logic

Propositional logic only allows for a schematic representation of world, as introduced in Section 2. The logic that allows more **complex knowledge representation** is **first-order logic**. It is expressive enough to describe the objects, the relations among the objects (*is adjacent to*, *red*, etc.) and functions on the objects (specific relations with only one possible output for an input, such as *is father of*, *head of*) [3]. It is considered the foundation of many other representation languages.

In first-order logic the **objects** are mapped to terms. A term could be a specific object in the world, referred to by a constant symbol. However, it is not always possible to have a distinct symbol for every object. To deal with this issue, functions on the objects come into picture. Functions allow defining additional objects related to the set of constant symbols. For example, Alice and Bob are specific objects in the world referred to by constants A and B respectively. Then, body parts of Alice and Bob could be referred to using Functions *Head_of(A)*, *RightHand_of(B)*, etc. Thus, the outputs of the functions are also terms but they refer to the unnamed objects of the world. Moreover, first-order logic allows referring to collections of objects instead of enumerating objects one by one. Such a reference would be an expression of a common property of the objects. This is done using two quantifiers, universal (\forall) and existential (\exists). Thus, in first-order logic it is possible to talk about properties that are common “for all” or “for some” objects respectively. With quantifiers, the objects are referred to by a variable that corresponds to any object that satisfies the condition. Finally, the **relations** between objects are mapped to predicate symbols (or simply predicates). Predicates may have varying numbers of arguments depending on the number of objects that the relation links together [3].

Atomic sentences to express facts in first-order logic are composed of a predicate symbol optionally followed by a parenthesized list of terms, its arguments. Complex sentences are built using syntactic rules with five logical connectives, as in propositional logic, with the addition of a rule for using quantifiers [3]. A summary of syntactic rules in first-order logic is presented in Figure 32.

⁷³ <http://www.primaryobjects.com/2018/07/23/logical-based-artificial-intelligence-and-expert-systems/>

```

Sentence → AtomicSentence | ComplexSentence
AtomicSentence → Predicate | Predicate ( Term, ... ) | Term = Term
ComplexSentence → ( Sentence ) | [ Sentence ]
                | ¬ Sentence
                | Sentence ∧ Sentence
                | Sentence ∨ Sentence
                | Sentence ⇒ Sentence
                | Sentence ⇔ Sentence
                | Quantifier Variable, ... Sentence

Term → Function ( Term, ... )
      | Constant
      | Variable

Quantifier → ∀ | ∃
Constant → A | X1 | John | ...
Variable → a | x | s | ...
Predicate → True | False | After | Loves | Raining | ...
Function → Mother | Left Leg | ...

OPERATOR PRECEDENCE : ¬ , = , ∧ , ∨ , ⇒ , ⇔

```

Figure 32: Syntax of first-order logic [3]

As in propositional logic, a sentence in first-order logic can be either **true** or **false**. Since the interpretation of the sentence depends on the **model of the world**, and the world has objects, functions on them and relations between them, the model has to provide the mapping from objects, functions and relations to constants, functions and predicate symbols respectively. The models can vary in how many objects they contain and in the way the constant symbols map to objects. With an increasing number of objects, the number of possible models grows exponentially. It is thus not possible to enumerate all possible interpretations to compute the meaning of a sentence [3].

Figure 33 provides an example of knowledge representation possible with first-order logic. It describes a world with two companies of which one is a contractor for another. It has constants such as "Alice Reddy", "Bob Jones", etc. referring to persons, or "Widgets Inc", "Consult Inc" referring to the specific companies in the described world. It has predicates with one argument such as "Company", "Employee", etc. to express common properties of objects. And it has predicates with two arguments such as "Works for", "Has contract with", etc. to express the relations between objects.

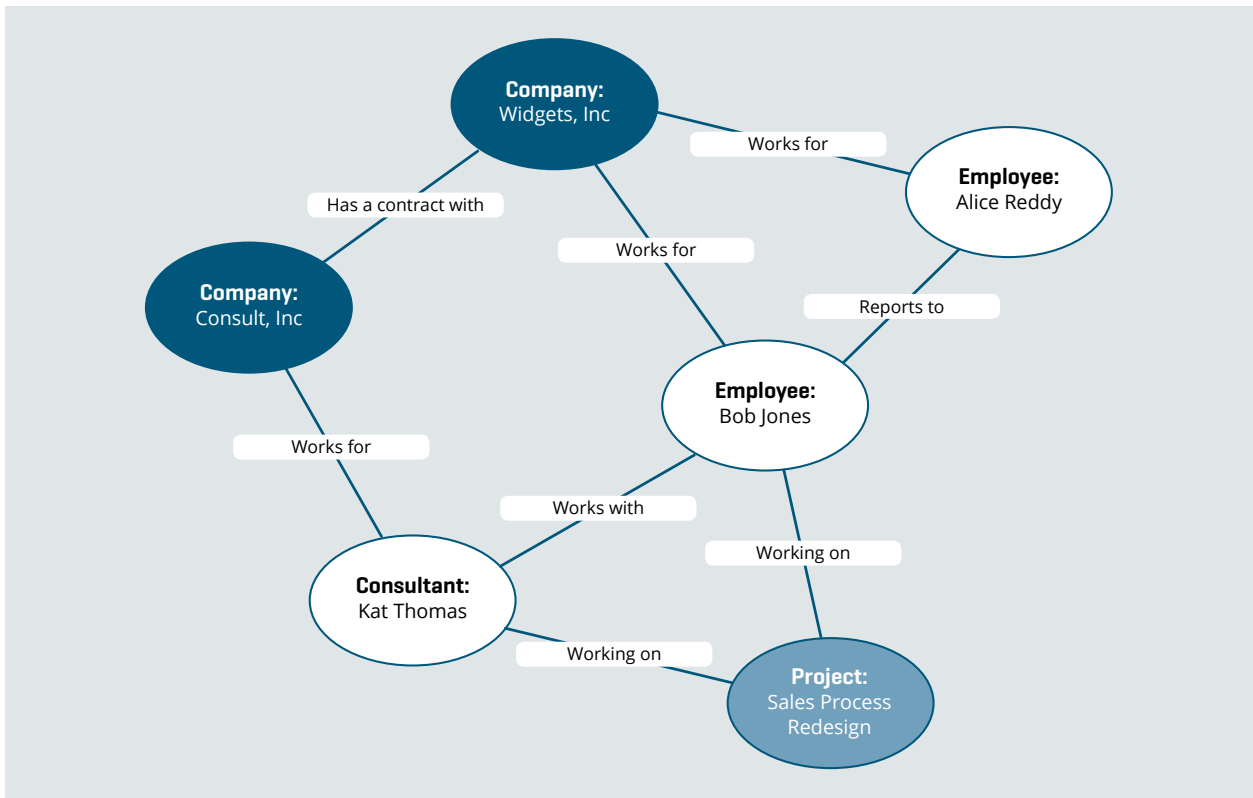


Figure 33: Knowledge representation in first-order logic⁷⁴

5.2.3. Building a knowledge database

Traditionally, a **knowledge database** is built by human experts who share their expertise for a given domain. In recent years, less specialized and more complex knowledge databases have become popular. To build such knowledge bases, ML techniques can be used. As a result, the AI system relying on such knowledge database takes advantage of both ML (for building a knowledge base) and logic-based AI (for the explainability of its outcomes), as highlighted in Figure 34.

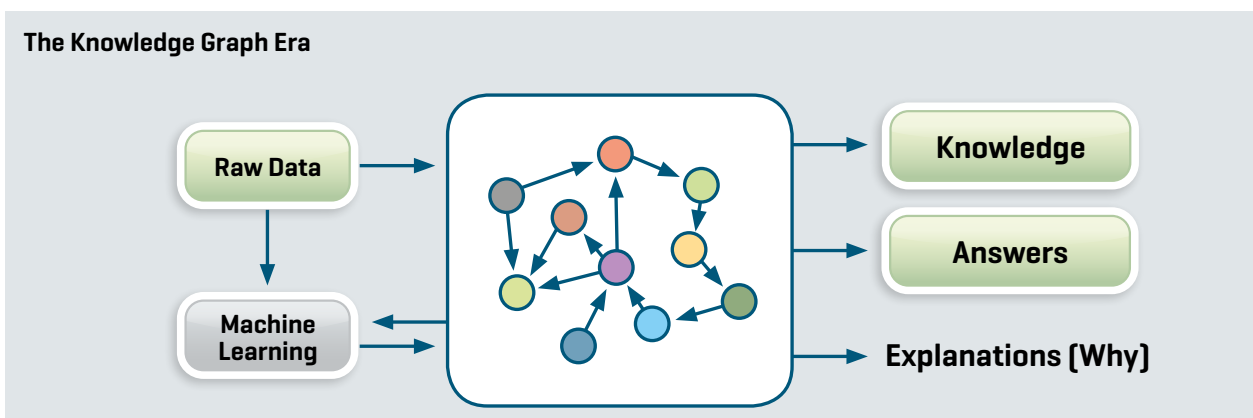


Figure 34: Combining ML and reasoning over a knowledge base to provide explainable solutions⁷⁵

74 Source: <https://towardsdatascience.com/ontology-and-data-science-45e916288cc5>

75 Source: <https://www.slideshare.net/dmccreary/ai-knowledge-representation-and-graph-databases-key-trends-in-data-science>

5.3. Reasoning

Once the knowledge base is built, it can be used in **logical reasoning** to solve problems. One basic concept of logical reasoning is the **logical entailment** between sentences. Sentence α entails sentence β , or sentence β follows from sentence α , if and only if in every model in which α is true β is also true. Entailment is the basis of logical inference, which could be a deduction of new information from the known facts [3].

In propositional logic, since it is possible to enumerate all possible models, the simplest inference algorithm is model checking. If the task is to find out if the sentence is true given the knowledge base, it simply considers the models where the knowledge base is true and checks if a sentence is true in all those models. In first-order logic, it is impossible to enumerate all the models. For this reason, and also because enumeration of models in propositional logic could be lengthy, the entailment by **theorem proving** is used. Theorem proving in this context stands for the application of formal inference rules to the sentences from the knowledge base, without consulting the models, in order to verify if the sentence in question is true [3].

One well-known inference rule is called Modus Ponens, which states that whenever sentences of the form $\alpha \Rightarrow \beta$ and α are given, then the sentence β can be inferred [3]. It is written as:

$$\frac{\alpha \Rightarrow \beta, \alpha}{\beta}$$

Another known inference rule is And-Elimination, which states that, if a sentence $\alpha \wedge \beta$ is given then any of its parts can be inferred [3]. It is written as:

$$\frac{\alpha \wedge \beta}{\alpha} \quad \text{or} \quad \frac{\alpha \wedge \beta}{\beta}$$

One useful concept in theorem proving is the concept of logical equivalence. Two sentences are equivalent if they are true in the same set of models. Replacing a sentence from the knowledge base by its equivalent could facilitate the inference by providing a more practical form of a sentence to handle within a particular inference request. There exist a few rules stating the equivalence of sentences in first-order logic, some examples are shown in Figure 35 [3].

$(\alpha \Rightarrow \beta) \equiv (\neg \beta \Rightarrow \neg \alpha)$	contraposition
$(\alpha \Rightarrow \beta) \equiv (\neg \alpha \vee \beta)$	implication elimination
$(\alpha \Leftrightarrow \beta) \equiv ((\alpha \Rightarrow \beta) \wedge (\beta \Rightarrow \alpha))$	biconditional elimination
$\neg(\alpha \wedge \beta) \equiv (\neg \alpha \vee \neg \beta)$	De Morgan
$\neg(\alpha \vee \beta) \equiv (\neg \alpha \wedge \neg \beta)$	De Morgan

Figure 35: Examples of logical equivalence [3]

In first-order logic, the inference rules are modified and some new rules are introduced in order to handle the

objects and the quantifiers. For example, there are rules of Universal Instantiation and Existential Instantiation (shown in a formula below). Given a sentence α , these rules serve to substitute a variable (v in Formula 1) by a constant symbol (g and k in Formula 1) related to a concrete object in the world and are applied to variables used with the universal and existential quantifiers respectively. In the case of the universal quantifier, the generalized substitution is more complex since it provides a list of all possible substitutes for a given variable in a given world.

$$\frac{\forall v \alpha}{\text{SUBST}(\{v/g\}, \alpha)} \qquad \frac{\exists v \alpha}{\text{SUBST}(\{v/k\}, \alpha)}$$

Formula 1: Universal and Existential Substitution in first-order logic [3]

Thus, in first-order logic the generalized Modus Ponens inference rule with n premises p_1, \dots, p_n , states that if for these premises there exists a substitution θ with which the implication of sentence q is true, then it can be inferred from these premises.

$$\frac{p_1', p_2', \dots, p_n', (p_1 \wedge p_2 \wedge \dots \wedge p_n \Rightarrow q)}{\text{SUBST}(\theta, q)}$$

Formula 2: Modus Ponens in first-order logic [3]

Using these different rules, two types of reasoning could be distinguished: forward chaining and backward chaining. **Forward-chaining reasoning** is used to answer the question “What can happen next?”. It starts with known facts and deduces the **possible outcomes**. **Backward-chaining reasoning** is used to provide the **explanations**. If the outcome is known, it would search for possible facts (premises) that could have led to this outcome [3] [40].

However, as already mentioned in Section 5.1, there are limits for reasoning with first-order logic. Three major needs that motivated the development of alternative forms of reasoning, namely of what is called **non-monotonic logic**, are belief revision, closed-world reasoning, and planning. **Belief revision** is needed to update the knowledge base, and in some cases the support for a belief can consist in the absence of contradictory beliefs. Reasoning with a closed-world assumption leads to consider only known facts in the database; any possible **unknown information** is not taken into consideration. **Planning** requires the ability to reason about the outcomes of a series of actions. In a complex world the usual assumption is to consider that most of the things will remain unchanged after the performance of actions. However, it is difficult to formalize what is subject to change and what is not [35].

Thus, one of the general inference rules in non-monotonic logic requires not only the presence of a set of proved conclusions, but the absence of certain other conclusions: in the presence of $\{A_1, \dots, A_n\}$ and in the absence of $\{B_1, \dots, B_n\}$, conclude C .

6. Examples of specialized AI systems

Sections 3 to 5 presented the basic techniques used in AI and showed that different algorithms work better on different problems. Thus, being a complex mechanism, an AI system usually exploits different algorithms and techniques.

Some typical AI systems are the ones specialized in computer vision, natural language processing, recommendations, etc.

Computer vision systems are intended to imitate the human visual system and perform tasks such as acquiring, processing, analyzing, and understanding digital images. Computer vision can be used for image classification, object detection, target tracking, etc. For example, in the healthcare domain, computer vision could be applied to identify disease features on medical images. In the field of security, it could be used to identify suspects on the surveillance footage. In the case of autonomous vehicles, it would detect traffic signs, buildings, other vehicles and pedestrians, etc. Modern computer vision systems are usually built using various types of neural networks, including deep learning [1] [41].

Natural language processing (NLP) systems are developed to imitate the human capacity to use languages. NLP is a vast domain, covering natural language understanding (NLU) and natural language generation (NLG) that could be used for both written (text) and spoken (speech) form. NLP is used in a variety of tasks: text understanding, text summarization, information extraction, machine translation, speech recognition and synthesis, etc. Intelligent personal assistants exploit NLP, for example, to understand the requests of the user and speak out a response. Spam detection could have an NLP component analyzing the content of email. The autocomplete feature in messaging applications is another example of NLP usage. In NLP, one finds the use of various algorithms and techniques: SVM as example for classification (such as spam detection), HMM for speech or text generation, neural networks for machine translation, logic-based methods for text summarization, etc. [1] [41].

Recommender systems are vastly used in e-commerce. They recommend to a user the articles that should match his or her preferences. Typically, the recommender system would be an association of ranking (listing of the articles ordered by potential interest to the user) and recommendation (suggestion of one or more of the first articles from the list) tasks. A number of algorithms, such as collaborative filtering, Naïve Bayes, kNN, etc. are used in recommender systems [1] [41].

7. Conclusion

This chapter introduced some of the main pillars of common AI techniques, such as searching, machine learning and logic-based. There is no one technique that fits all the cases, and it is crucial for businesses to understand their problem and to find an appropriate technique. In Chapter 3 different use cases will be considered to show the real-world application of some techniques.

On the other hand, the complexity of some techniques could be an obstacle to user acceptance and business adoption. While this chapter aimed at shedding some light on the functioning of AI and touch upon some development aspects that would make an AI system trustworthy (such as evaluation of ML algorithms in Section 4.3), it does not explain in detail the challenges related to the building of a trustworthy AI. Chapter 4 will provide more information with respect to these challenges and how they could be addressed.

References

- [1] ISO/IEC JTC1/SC 42/SG1 on Computational approaches, "Study Report on Computational Approaches and AI Systems," 2019.
- [2] A. Amidi and S. Amidi, "CS 221 — Artificial Intelligence," 05 2019. [Online]. Available: <https://stanford.edu/~shervine/teaching/cs-221/>. [Accessed 11 2020].
- [3] S. Russel and P. Norwig, *Artificial Intelligence: a modern approach (3rd edition)*, Upper Saddle River, New Jersey: Pearson Education, Inc., 2010.
- [4] S. Bringsjord and N. S. Govindarajulu, "Artificial Intelligence," *The Stanford Encyclopedia of Philosophy*, Fall 2018. [Online]. Available: <https://plato.stanford.edu/entries/artificial-intelligence/#ApprAI>. [Accessed 11 2020].
- [5] F. Corea, "AI knowledge map: how to classify AI technologies," *Forbes / Cognitive World*, 08 2018. [Online]. Available: <https://www.forbes.com/sites/cognitiveworld/2018/08/22/ai-knowledge-map-how-to-classify-ai-technologies/#72d6fb857773>. [Accessed 11 2020].
- [6] D. L. Poole and A. K. Mackworth, *Artificial Intelligence: Foundations of Computational Agents*, Cambridge University Press, 2017.
- [7] Paul G. Allen School of Computer Science and Engineering, "Lecture notes on Uninformed Search for CSE573: Artificial Intelligence I," 2012. [Online]. Available: <https://courses.cs.washington.edu/courses/cse573/12sp/lectures/02-search.pdf>. [Accessed 11 2020].
- [8] Gartner, "Preparing and architecting for Machine Learning," 2017. [Online]. Available: <https://www.gartner.com/en/documents/3573617/preparing-and-architecting-for-machine-learning>. [Accessed 11 2020].
- [9] J. Brownlee, "Basic Concepts in Machine Learning," 2015. [Online]. Available: <https://machinelearningmastery.com/basic-concepts-in-machine-learning/>. [Accessed 11 2020].
- [10] Towards Data Science, "Machine Learning For Beginners," 2018. [Online]. Available: <https://towardsdatascience.com/machine-learning-for-beginners-d247a9420dab>. [Accessed 11 2020].
- [11] J. Chugh, "Types of Machine Learning and Top 10 Algorithms Everyone Should Know," 12 2018. [Online]. Available: <https://blogs.oracle.com/ai/types-of-machine-learning-and-top-10-algorithms-everyone-should-know>. [Accessed 11 2020].
- [12] McKinsey & Company, "An executive's guide to AI," 2018. [Online]. Available: <https://www.mckinsey.com/business-functions/mckinsey-analytics/our-insights/an-executives-guide-to-ai>. [Accessed 11 2020].
- [13] skymind, "A.I. Wiki: A Beginner's Guide to Deep Reinforcement Learning," [Online]. Available: <https://wiki.pathmind.com/deep-reinforcement-learning>. [Accessed 11 2020].
- [14] skymind, "A.I. Wiki: A Beginner's Guide to Neural Networks and Deep Learning," [Online]. Available: <https://wiki.pathmind.com/neural-network>. [Accessed 11 2020].
- [15] ISO/IEC JTC 1/SC 42/SG 2 on Trustworthiness, "Study group report on robustness," 2018.
- [16] CSSF, "Artificial Intelligence : opportunities, risks and recommendations for the financial sector," 2018.
- [17] A. Ticlavilca and A. Torres, "Data Driven Models and Machine Learning (ML) Approach in Water Resources Systems," [Online]. Available: https://usu.instructure.com/files/62445095/download?download_frd=1. [Accessed 11 2020].
- [18] M. Chui, R. Chung, N. Henke, S. Malhotra, J. Manyika, M. Miremadi and P. Nel, "Mapping AI techniques to problem types," 10 2018. [Online]. Available: <https://www.mckinsey.com/industries/financial-services/our-insights/mapping-ai-techniques-to-problem-types>. [Accessed 11 2020].
- [19] I. Moise, E. Pournaras and D. Helbing, "Classification and decision trees," 2016. [Online]. Available: <https://www.ethz.ch/content/dam/ethz/special-interest/gess/computational-social-science-dam/documents/education/Spring2016/Datascience/Classification%20and%20Decision%20Trees.pdf>. [Accessed 11 2020].
- [20] S. Agarwal, "A Tutorial Introduction: Ranking Methods in Machine Learning," 2010. [Online]. Available: <http://www.shivani-agarwal.net/Events/SDM-10-Tutorial/sdm10-tutorial.pdf>. [Accessed 11 2020].
- [21] J. Hurwitz and D. Kirsch, "Machine Learning For Dummies, IBM Limited Edition," John Wiley & Sons, Inc., 2018.
- [22] A. Amidi and S. Amidi, "CS 229 — Machine Learning," 2018. [Online]. Available: <https://stanford.edu/~shervine/teaching/cs-229/>. [Accessed 11 2020].
- [23] R. Sutton and A. Barto, "Reinforcement Learning: An Introduction (Complete Draft)," 2017. [Online]. Available: <http://incompleteideas.net/book/bookdraft2017nov5.pdf>. [Accessed 11 2020].
- [24] OpenAI, "Welcome to Spinning Up in Deep RL," [Online]. Available: <https://spinningup.openai.com/en/latest/index.html>. [Accessed 11 2020].

- [25] *TutorialsPoint*, "Big Data Analytics Tutorial," [Online]. Available: https://www.tutorialspoint.com/big_data_analytics/index.htm. [Accessed 11 2020].
- [26] A. Navlani, "KNN Classification using Scikit-learn," 2018. [Online]. Available: <https://www.datacamp.com/community/tutorials/k-nearest-neighbor-classification-scikit-learn>. [Accessed 11 2020].
- [27] B. Boehmke, "University of Cincinnati - Business Analytics R Programming Guide," [Online]. Available: <http://uc-r.github.io/predictive>. [Accessed 11 2020].
- [28] L. Smith, "A tutorial on Principal Components Analysis," Department of Computer Science, University of Otago, 2002. [Online]. Available: www.cs.otago.ac.nz/research/publications/OUCS-2002-12.pdf. [Accessed 11 2020].
- [29] M. Franzese and A. Iuliano, "Hidden Markov Models," *Encyclopedia of Bioinformatics and Computational Biology*, vol. 1, pp. 753-762, 2019.
- [30] S. Cantrell, "Top 3 Most Popular Neural Networks," *Excella*, 02 2018. [Online]. Available: <https://www.excella.com/insights/top-3-most-popular-neural-networks>. [Accessed 11 2020].
- [31] A. Violante, "Simple Reinforcement Learning: Q-learning," *Towards Data Science*, 03 2018. [Online]. Available: <https://towardsdatascience.com/simple-reinforcement-learning-q-learning-fcddc4b6fe56>. [Accessed 11 2020].
- [32] *Towards Data Science*, "Applications of Reinforcement Learning in Real World," 08 2018. [Online]. Available: <https://towardsdatascience.com/applications-of-reinforcement-learning-in-real-world-1a94955bcd12>. [Accessed 11 2020].
- [33] D. Silver, T. Hubert, J. Schrittwieser and D. Hassabis, "AlphaZero: Shedding new light on chess, shogi, and Go," *DeepMind*, 12 2018. [Online]. Available: <https://deepmind.com/blog/article/alphazero-shedding-new-light-grand-games-chess-shogi-and-go>. [Accessed 11 2020].
- [34] E. Martin, "DeepMind's AlphaZero Engineering Use Cases," *Medium*, 02 2018. [Online]. Available: <https://medium.com/predict/deepminds-alphazero-engineering-use-cases-eac50cf95c4b>. [Accessed 11 2020].
- [35] R. Thomason, "Logic and Artificial Intelligence," *The Stanford Encyclopedia of Philosophy*, November 2018. [Online]. Available: <https://plato.stanford.edu/entries/logic-ai/>. [Accessed 11 2020].
- [36] *Generation5*, "Expert System in Real World Applications," 2005. [Online]. Available: <https://www.semanticscholar.org/paper/Expert-System-in-Real-World-Applications-Wai-Rahman/94a7595c51fada8ecfe0b2ecc805a6d84f8517d4>. [Accessed 11 2020].
- [37] D. Saxton, E. Grefenstette, F. Hill and P. Kohli, "Analysing Mathematical Reasoning Abilities of Neural Models," 04 2019. [Online]. Available: <https://arxiv.org/abs/1904.01557>. [Accessed 11 2020].
- [38] D. McCreary, "Knowledge Graphs: The Third Era of Computing," 2019. [Online]. Available: <https://medium.com/@dmccreary/knowledge-graphs-the-third-era-of-computing-a8106f343450>. [Accessed 11 2020].
- [39] M. Krötzsch, "Ontologies for Knowledge Graphs?," in *Proceedings of the 30th International Workshop on Description Logics, Montpellier, France, 2017*.
- [40] *Tutorials Point*, *Artificial Intelligence: Intelligent systems*, Tutorials Point (I) Pvt. Ltd, 2015.
- [41] IEC, "Artificial intelligence across industries," 2018. [Online]. Available: <https://basecamp.iec.ch/download/iec-white-paper-artificial-intelligence-across-industries-en/>. [Accessed 11 2020].

NOTE: While any hyperlinks included in this chapter were valid at the time of consultation, ILNAS cannot guarantee their long-term validity.

3

Artificial Intelligence: use cases and applications

1. Introduction

Artificial Intelligence is more hyped now than ever. Multiple applications of AI exist already on the market, and even more are to come in the future. This chapter first explores the benefits of AI and then presents a variety of application domains for AI. It finally focuses on a few most prominent application domains and specifies for them the high-potential use cases, expected benefits, existing challenges and standardization activities. It also highlights the national activities for these domains, where applicable.

2. Global market overview

Various AI-related technologies, such as AI PaaS, Explainable AI, Adaptive machine learning (ML), etc., were part of a “Gartner Hype” in 2019⁷⁶. Indeed, the AI market has seen a tremendous growth in recent years and it keeps rapidly growing. According to MarketWatch, the **AI market was valued at \$23.94 Billion in 2018 and is expected to reach \$208.49 Billion by 2025** [1].

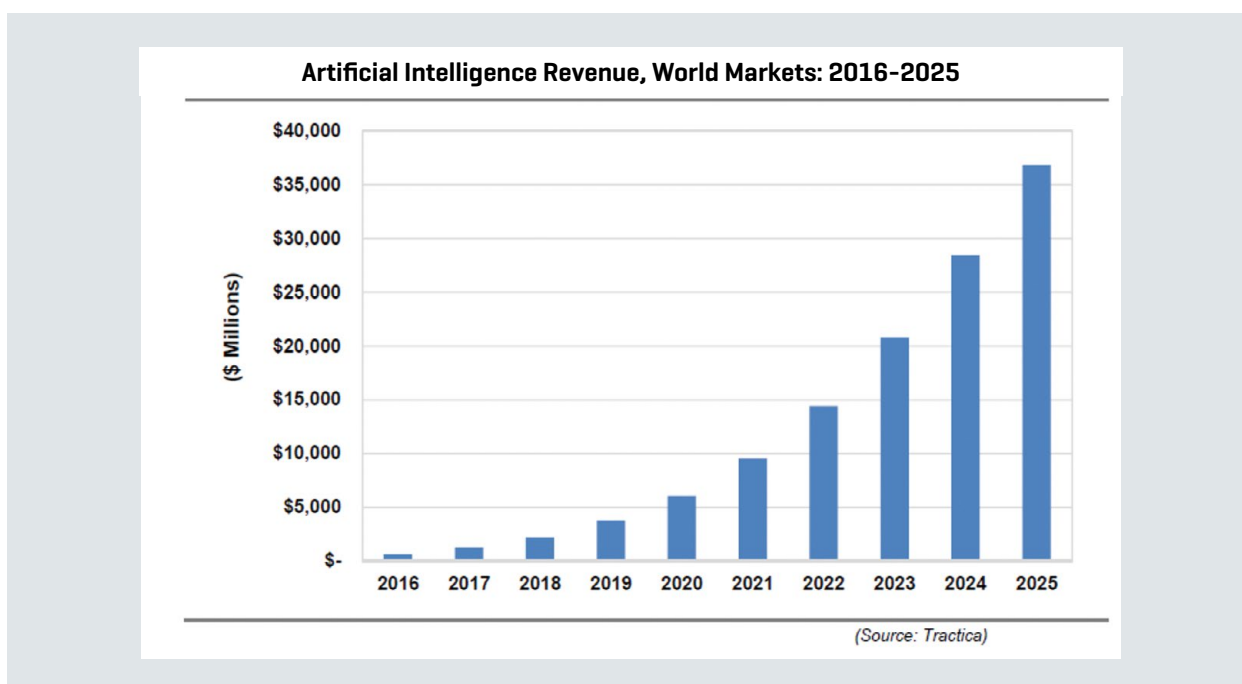


Figure 1: AI revenue worldwide [2]

According to another forecast, made by Tractica, the revenues related to the application of AI software were expected to grow from \$643.7 million in 2016 to about \$3 billion in 2018 and reach \$36.8 billion in 2025 (as shown in Figure 1) [2]. However, a similar report by the same company made two years later estimated the revenues in 2018 to be \$9.5 billion, tripling the expectations from 2016 and thus showing the increasing impact of AI on the global market [3]. Worldwide spending on systems exploiting AI software capabilities give an even more striking example of the AI impact as it was forecast to reach \$35.8 billion in 2019 and more than double that by 2022 [4].

76 <https://www.gartner.com/smarterwithgartner/5-trends-appear-on-the-gartner-hype-cycle-for-emerging-technologies-2019/>

The opportunities generated from AI are not only coming from the usage of AI-based software (as presented in Figure 2). The development AI-based products and services also contributes to economic growth [5].

This fact is supported by the World Intellectual Property Organization’s (WIPO) analysis of AI technology trends. Indeed, more than 340 000 AI-related inventions were registered between the 1950s and 2018, with more than half of them after 2013. These inventions cover various AI techniques used in different applications across multiple industries [6].

In addition, as discussed in Chapter 1, AI depends on its enabling technologies, for example the Internet of Things (IoT) for data generation or Cloud Computing to support data storage and processing. Providing AI specific infrastructure is a business opportunity, supporting economic development [5] [7].

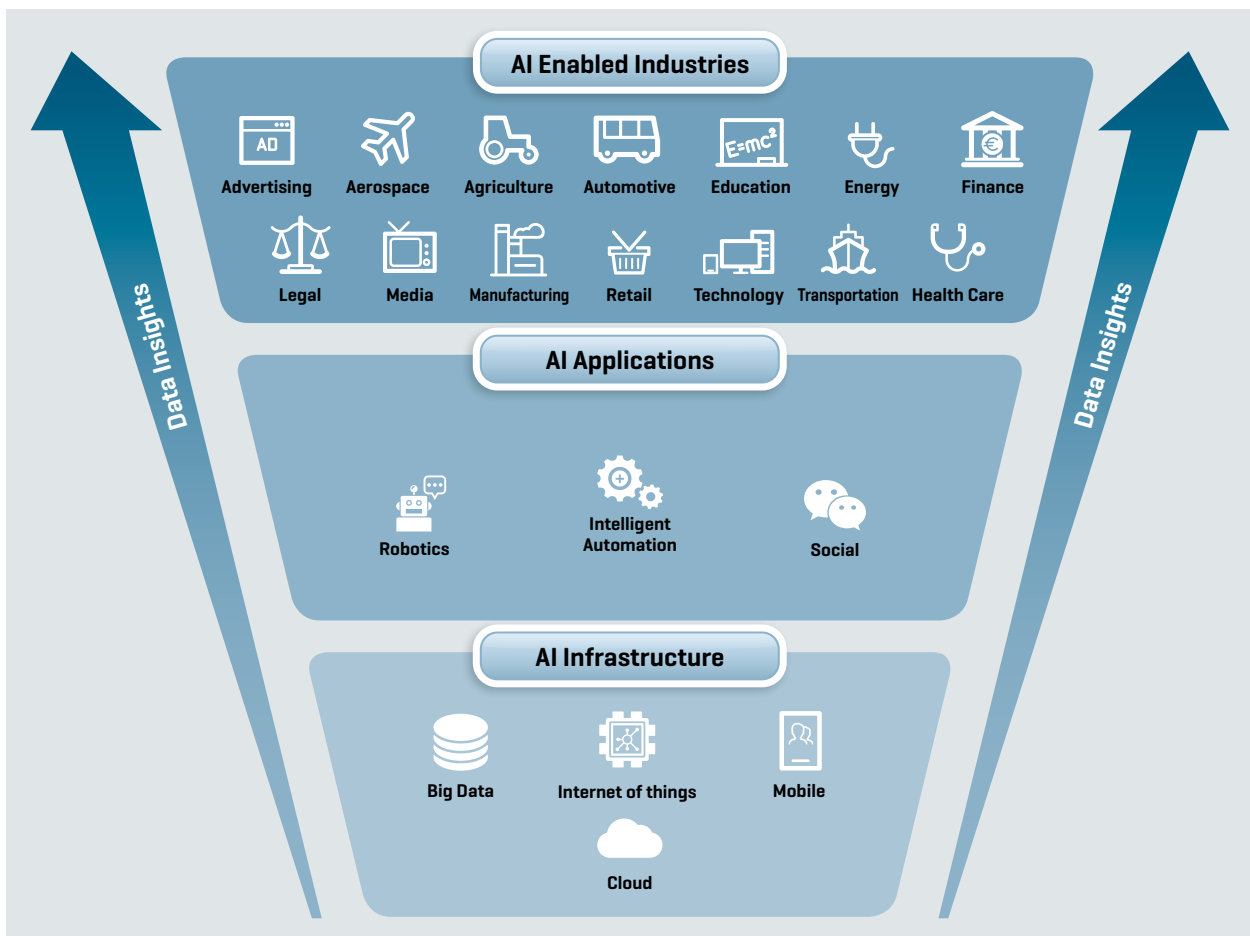


Figure 2: Business opportunities generated by AI (adapted from [5])

From the history of patenting activities, Japanese, American and Chinese companies have the leading role in AI-related innovations [6]. In terms of global AI ecosystem based on start-up data, the US is the global market leader with a 40% market share, followed by China and Israel [8]. However, in terms of global GDP that takes into account the consumption, investments and trading between countries, it is China that is expected to lead the race by 2030, followed by North America, Europe and other Asian countries (as shown on Figure 3). In total, the share of AI gains on global GDP could reach \$15.7 trillion by 2030 [9] [10].

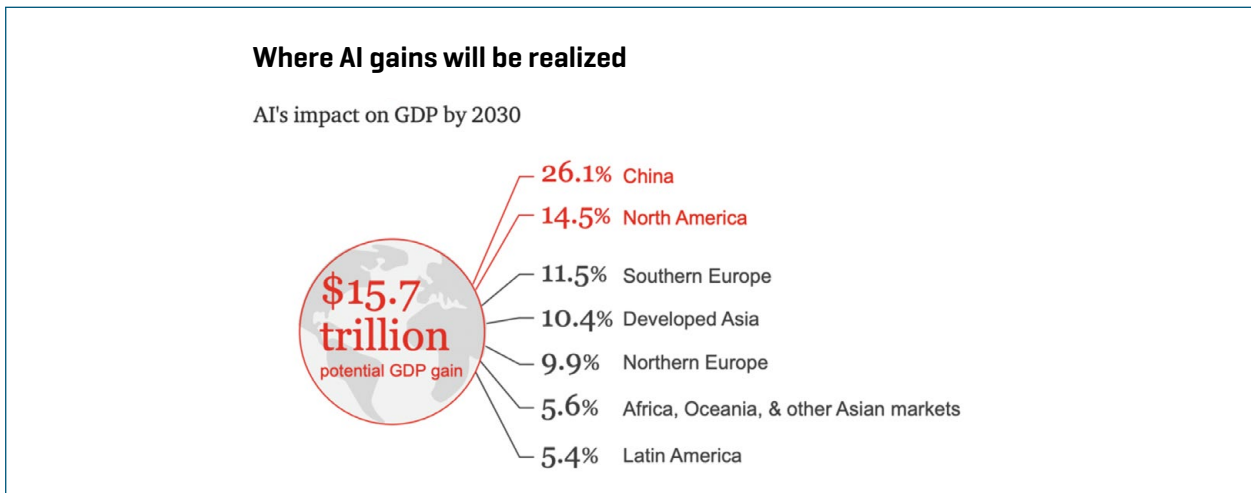


Figure 3: Gains from AI by 2030 expressed in GDP % [9]

It is interesting to mention that the gains related to AI could be measured in terms of added value to humans. Indeed, the usage of AI contributes to the quality of products and services; it helps to provide personalization, and it increases labor productivity by reducing the time spent on repetitive or burdensome tasks (see Figure 4) [10].

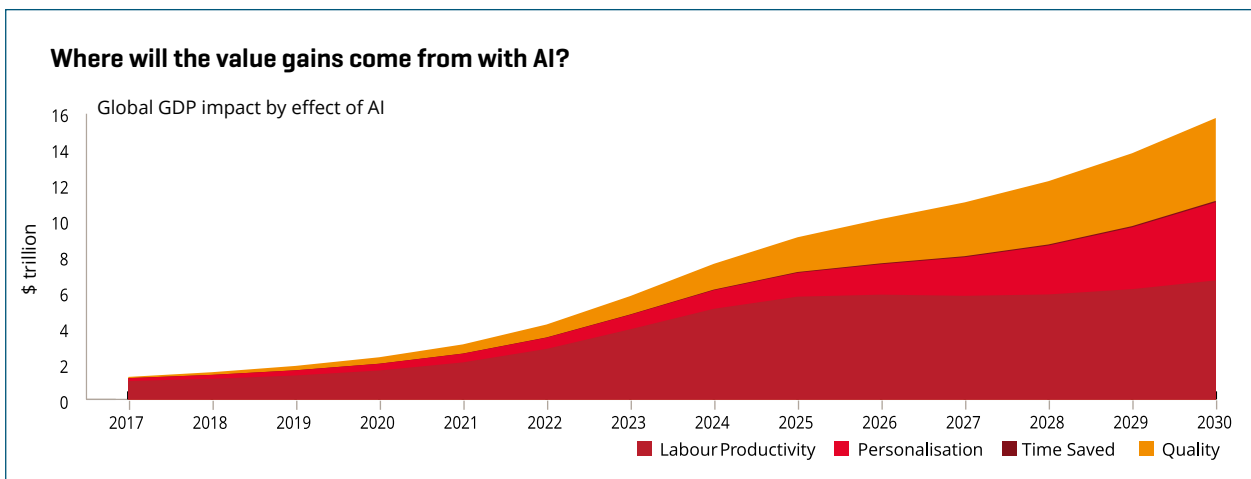


Figure 4: Value gains from AI [10]

3. European activities in AI

In Europe, the economic impact of AI is highlighted for the following three fields: robotics, autonomous vehicles and automation of knowledge. For these domains only the economic impact is estimated to reach between €6.5 and €12 trillion by 2025. And the expected benefits of AI are not limited to them. Among other examples, the European Commission mentions “better healthcare, more efficient public administration, safer transport, a more competitive industry and sustainable farming” [11].

Since April 2018, when a **Declaration of cooperation on Artificial Intelligence**⁷⁷ was signed, European countries have been leading actions towards achieving European competitiveness in the area of AI. Namely, the European approach to AI is based on 3 pillars⁷⁸:

- Encouraging technological developments in public and private sectors.
- Preparing for socio-economic changes.
- Ensuring an appropriate ethical and legal framework.

To address the first point, the European Commission has invested around €2.6 billion through its Horizon 2020 program. For example, the AI4EU⁷⁹ project received the funding of about €20 million in order to build a European-on-demand platform to foster collaboration and knowledge sharing in the area of AI. Another example are the research programs in smart robotics that received around €700 million. These public investments are complemented by private initiatives that amounted to €2.4-3.2 billion in 2016 [11] [12].

Following the **European Strategy on AI**⁸⁰, the public and private investments in AI are targeted to add up to €20 billion for the 2018-2020 period and increase by €20 billion per year from 2020 to 2025. Indeed, as part of the public spending under the next multiannual financial framework, the EU investment in AI is expected to reach at least €7 billion through the Horizon Europe and Digital Europe programs [11] [12].

Moreover, to help the population prepare for the changes that digitalization and AI are bringing, the Commission has invested specifically €2.3 billion in digital skills over 2014-2020. This will cover the design and delivery of short-term, long-term and on-the-job trainings for digital technologies [11] [12].

To get a better grasp of the current European AI ecosystem, the European Commission held a workshop during which national initiatives in academic, industrial, and governmental sectors were presented. Being part of this workshop, Luxembourg announced a number of activities showing its potential in the area of AI [13]:

- A number of research projects in the areas of ML, robotics, drones, autonomous driving, data mining, data visualization, software testing, human-machine interaction, etc.
- The Digital Luxembourg (Digital Lëtzebuerg) and ICT Luxembourg initiatives demonstrating strong government support for digitalization, including the deployment of a dedicated infrastructure, promotion of digital public services, enhancement of the innovation ecosystem, fostering of the e-skills and work on data regulation.
- A growing number of AI start-ups and increasing interest in AI from big industrial players.

77 <https://ec.europa.eu/digital-single-market/en/news/eu-member-states-sign-cooperate-artificial-intelligence>

78 <https://ec.europa.eu/digital-single-market/en/artificial-intelligence>

79 <https://www.ai4eu.eu/>

80 <https://ec.europa.eu/digital-single-market/en/artificial-intelligence>

In an attempt to involve a larger number of stakeholders in the discussion about the future of AI in Europe, including on the topics of policy-making and socioeconomics, the European Commission set up a multi-stakeholder forum, called the **European AI Alliance**. In addition, the **High-Level Expert Group on AI (AI HLEG)** was formed in order to support the EU's policy-makers by taking inputs from the European AI Alliance forum and providing guidelines and recommendations. The two major deliverables produced by the AI HLEG are the "Ethics guidelines for trustworthy AI" [14] and the "Policy and investment recommendations for trustworthy Artificial Intelligence" [15]. Both documents state the need for a regulatory framework to guarantee that a human-centric, ethical and trustworthy approach to AI is adopted. Moreover, both documents mention standardization as a co-regulatory mechanism to guarantee the quality and conformance of AI-based products and services and recommend the development of standards tailored to European needs. To address this point, the European standardization organizations established dedicated working groups: the **CEN-CENELEC Focus Group on AI**⁸¹ and the **ETSI Industry Specification Group Securing AI**⁸² (see Section 6.2 in Chapter 1 for more information on these groups).

4. AI applications and use cases

While previous sections show that the expected impact of AI is big and the investments in the technology are increasing, the current section presents concrete applications of AI. It starts with an overview of AI applications and application domains and then focuses on specific use cases in these domains. To conclude, the section provides highlights on the usage of AI in Luxembourg.

Chapter 2 already gave a hint of AI usage and introduced a few applications exploiting specialized AI systems such as natural language processing (NLP), computer vision or recommender systems. Indeed, some types of AI applications are easily identifiable as solutions to particular problem areas, and are thus commonly adopted by various industries. Such an AI application could be further adapted and tailored to the specificities of each industry, but the particular aspects of a task to accomplish allows attributing it to a particular type of AI. This is why they can be called **horizontal AI applications**. This section will illustrate the concept with a few examples rather than provide a comprehensive review of AI applications.

4.1. Horizontal AI applications

AI applications fall into the following types [7], the list being non-exhaustive and partially overlapping:

- **Operations intelligence**

This type of AI application aims at leveraging ML to optimize and automate business operations, processes and interactions. It could be used to gain insight into the operations run by a business, or visualize patterns not visible through simple dashboards, or identify potential problems, etc. that are the tasks related to business optimization. It could also be used in intelligent process automation, learning from interactions with users, applying the knowledge and incrementally improving the business process [7].

81 <https://www.cencenelec.eu/news/articles/Pages/AR-2019-001.aspx>

82 <https://www.etsi.org/committee/1640-sai>

- **Autonomous machines and robots**

This type of AI application includes autonomous vehicles, bots and systems to address a wide range of tasks. For example, these could be physical autonomous or semi-autonomous bots providing assistance in the home, the office, a manufacturing plant, etc. exploiting ML for better quality and/or personalized services. It could also be virtual bots leveraging AI for cyber-security [7].

- **Intelligent document processing**

This type of AI application is used in complex document workflows to handle multiple documents' interactions and analyze documents' content. An example of such an application is a complex contract management system. On the one hand, it would do document management (thematic annotation, classification, etc.). On the other hand, it would learn the documents' workflow by understanding the legal framework applied to similar documents [7].

- **Intelligent task assistant**

This type of AI application is intended to carry out regular routine tasks such as scheduling appointments or assisting with online ticket buying. Scheduling tasks requires managing and coordinating various activities planned by one or more participants. Intelligent personal assistants would interact with individuals, learn their preferences and help with a range of tasks in a personalized manner [7].

- **Conversational systems**

This type of AI application would exploit ML and reasoning techniques to lead a meaningful conversation. Coupled with a human-friendly interface, a conversational system could use written (in the case of a chatbot) or spoken (in case of voice assistant) language. Conversational system could use knowledge of a specific domain to better "understand" requests and provide more targeted replies. Moreover, conversational systems could gain insights from various types of interactions (such as emails, human-to-human communication, etc.) and make suggestions for future exchanges [7].

- **Predictive analytics and decision support**

This type of AI application is useful for organizations wishing to get insights from data in order to identify future trends, get assistance with multi-factor decision making, simulate various scenarios for potential outcomes, etc. Predictive analytics is helpful for preventive maintenance or for getting information on potential challenges or opportunities ahead of time. AI-enabled decision support is based on the analysis of data from a wide range of sources allowing for informed decision-making. AI-powered simulations let organizations run different scenarios in order to test products or select an optimal course of action [7].

4.2. AI application domains

With AI penetrating every area of our lives, the number of application domains where simple AI techniques as well as the aforementioned horizontal applications are used is large.

WIPO identified 20 application fields through the analysis of patents submitted for AI-based innovations. On the list one can find [6]:

- Life and medical sciences
- Cartography
- Networks (for example, social networks and smart cities)
- Military
- Security
- Industry and manufacturing
- Law, social and behavioral sciences
- Energy management
- Arts and humanities

- Banking and finance
- Computing in government
- Telecommunications
- Document management and publishing
- Entertainment
- Education
- Transportation
- Agriculture
- Business (such as customer service, e-commerce, etc.)
- Physical sciences and engineering
- Personal devices, computing and Human-Computer Interface (HCI)

For the purpose of technical standardization in the area of AI, 22 application domains were identified in ISO/IEC TR 24030 *Artificial Intelligence (AI) — Use cases*⁸³. Some of them are the same as in WIPO list (for example, Transportation, Agriculture, Education), some of them overlap partially with the WIPO domains (for example, Digital marketing, Retail vs Business, Fintech vs Banking and finance), and some of them are complementary (for example, Construction, Logistics).

The application domains can be arranged by the amount of investments in AI systems, by the level of adoption of AI in the domain, by the number of patent applications, by the potential impact of AI in terms of quality enhancement, personalization enhancement, and its potential to save time.

In terms of expenses, the top industries investing in AI-based products and services are retail, banking, manufacturing, and healthcare. They invest, for example, in automated customer service agents, recommender systems, prevention analysis, automated threat intelligence and fraud detection. Among the industries that are expected to grow rapidly there are government, education and customer services [4]. In terms of expected monetary value of AI across different sectors, the front-runners are retail, transport and logistics, travel, public and social, and the automotive sector [8].

Another report shows that current leaders of AI adoption are also likely to invest more money in it in the near future. According to this report, financial services, high-tech, advanced manufacturing, energy, entertainment, transportation, retail, health care, and education are in the AI vanguard. Travel and tourism along with the construction industry seem to be less open to AI [16]. The level of adoption of AI by different industries as well as the estimation of the future increase of their investment in AI is captured in Figure 5.

83 <https://www.iso.org/standard/77610.html>

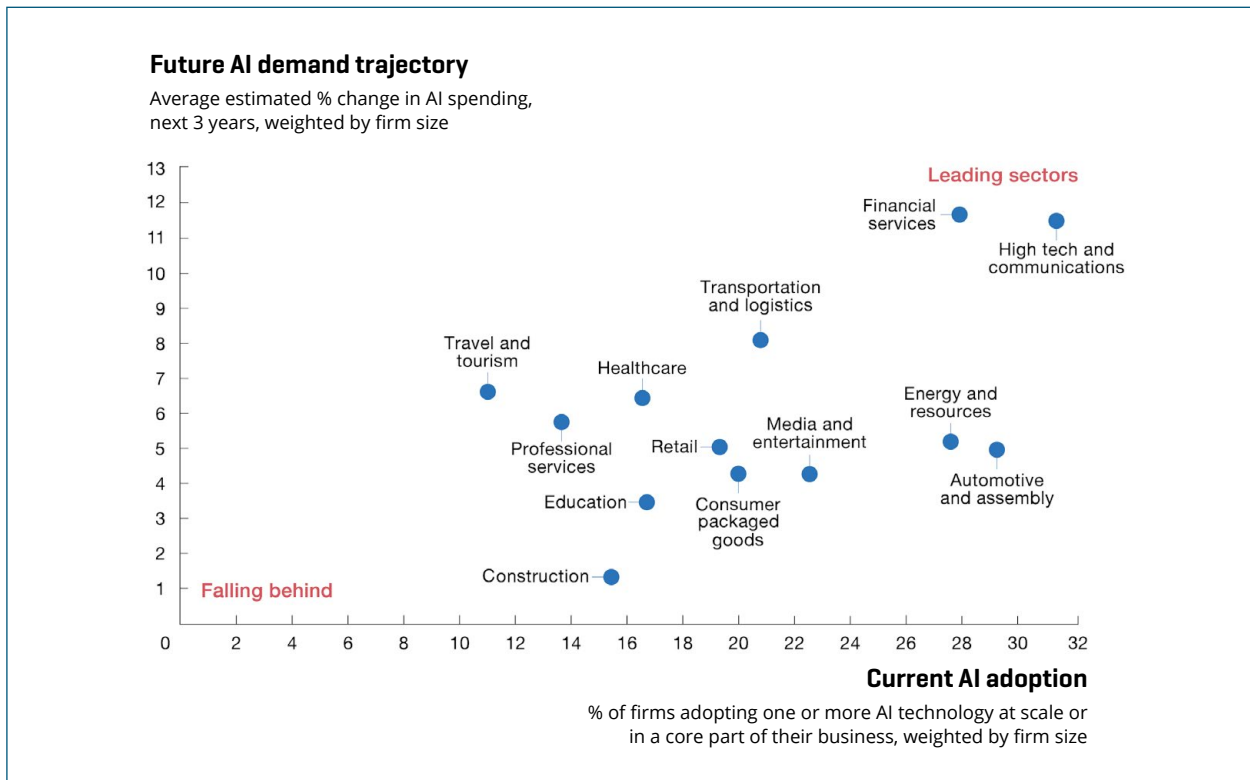


Figure 5: Adoption of AI by different industries and their estimated future investments in AI [16]

With respect to the patent applications, the forefront is kept by telecommunications, followed by the transportation sector. These two fields amount for 48% of all patent families related to AI applications. Next in line come life and medical sciences, personal devices and human-computer interface, and security. A comparison of the 20 application domains identified by WIPO is presented in Figure 6. It is interesting to note that the transportation, agriculture, government and banking sectors are the ones with the most remarkable average growth over the recent years. Another interesting fact is that 71% of all AI-related inventions address at least two distinct application domains, showing the complexity of its ecosystem [6].

The sector that is most impacted by the AI transformation in terms of quality, personalization and time gains is healthcare. Transportation, financial and high-tech will also experience significant changes [9].

Finally, the most prolific application domains in terms of use cases submitted for the analysis of standardization needs are healthcare, manufacturing, and ICT.

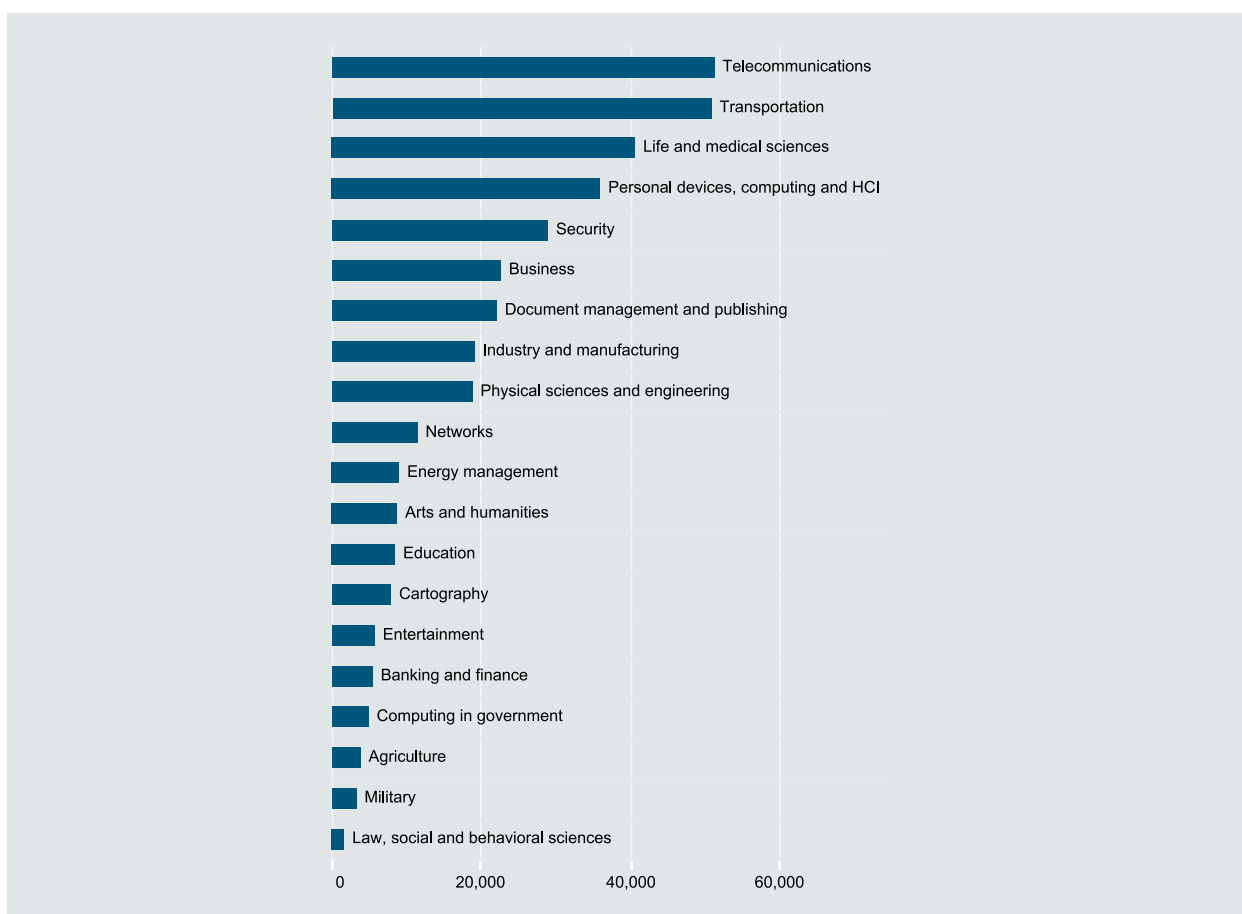


Figure 6: Distribution of number of AI patent families over application domains from 1980s until 2016

4.3. AI use cases

This section considers the application domains that were highlighted from different perspectives in the previous section, such as healthcare, banking, manufacturing, transportation, and retail. In addition, a closer look will be taken on a rapidly growing sector such as government. For these application domains, some typical use cases will be discussed, linking them to initiatives from Luxembourg when possible.

4.3.1. Healthcare

For more details on the usage of AI in healthcare, the reader can refer to [10] [16] [17] [18] [19], which were used as a source of information for this summary.

Motivation

- A growing population and need of personalized care.
- Cost reduction and healthcare accessibility.

Top high-potential use cases

- Supporting diagnostics (including images and electronic health records analysis) and treatment.

- Early identification and tracking of disease outbreaks in order to prevent further spreading and contain the epidemic.
- Optimization of non-clinical (operational) tasks (scheduling appointments, registering patients at a hospital, etc.).
- Preventive healthcare.

Currently on the market

- Smarter appointment scheduling (e.g. appointments and operations).
- Medical image processing for the identification of abnormalities.
- Activity tracking and health insurance.
- Support with personalized radiation treatment plans for cancer patients.
- Payment and claims management.

Promising technology

Due to the sensitivity of the data and regulatory complexity, most of the use of AI is currently concentrated in operations and customer service. Intelligent assistants and chatbots can be used to automate a wide range of tasks. In this context, natural language processing (NLP) including speech recognition and synthesis have a promising use in the healthcare sector.

Diagnostics often involve medical image analysis, and computer vision may sometimes recognize abnormalities that are not visible to the human eye. Diagnostics can also benefit from the comparative analysis of healthcare records of millions of patients with similar conditions. NLP could assist healthcare practitioners in this task. Moreover, a number of expert systems in various medical subfields have been developed to assist with diagnostics and recommending treatment.

A combination of hidden Markov models (HMMs) with neural networks, namely recurrent neural networks, has proven to provide good results in research aiming at predicting the patient's healthcare pathway and recommending personalized treatments.

Potential benefits

The use of AI in healthcare could help to improve care and reduce costs.

With the help of AI-based systems, a number of tasks can be automated or enhanced in the following areas: diagnostics, knowledge generation, public health, system efficiency, personalized medicine. This would lead to more accurate diagnostics, more personalized treatment, faster detection of disease outbreaks, decreased waiting time in emergency rooms and earlier-stage intervention. AI tools could accelerate a shift towards preventive medicine and remote assistance keeping the patients out of hospitals.

As a result, medical practitioners would be relieved of routine activities, patients' satisfaction would increase, and significant cost savings could be made.

Challenges to be addressed

The main concern with respect to the usage of AI in healthcare seems to be the privacy and protection of sensitive health data. This is a barrier not only for the patients whose data may be used but also for the AI application providers. Not only is it challenging for them to collect high-quality data, they also face technical difficulties of integrating data from multiple sources while dealing with strict regulations.

Another issue is complexity and variety: complexity of human biology and generated data, variety of medical equipment and data registration systems. On the one hand, this may cause a longer time-to market due to the need to fine-tune an AI solution and reach its full potential. On the other hand, it was highlighted that electronic health records cannot always be shared with patients and other providers due to the lack of interoperability between registration systems.

Evidence of effectiveness is expected to increase patients' acceptance. Indeed, a proof of a good functioning and of benefits of AI is necessary. A starting point could be the mass deployment and demonstration of the advantages of back-end operational solutions. Advanced field trials and pilots could also support user acceptance.

Since healthcare is a life-critical matter, having a tangible explanation of the results of an AI system, be it diagnostics, personalized treatment plan, or another solution, is particularly important. Thus, explainability of AI-provided results needs to be guaranteed.

Benefits of standardization

Standards can support the adoption of AI in healthcare by addressing different challenges. First of all, standards related to data quality would improve the quality of AI-based products and services. In a technical committee within the International Organization for Standardization (ISO), ISO/TC 215 *Health Informatics*⁸⁴, a few standards projects have been launched to provide guidelines on data collection, sharing and annotation. For example:

- ISO/TR 21835:2020 *Health informatics – Personal health data generated on a daily basis*⁸⁵.
- ISO/TR 21394 *Genomics informatics – Whole Genomics Sequence Markup Language (WGML)*⁸⁶.
- ISO/TS 23357 *Clinical genomics data sharing specification for next generation sequencing*⁸⁷.

Not directly related to the use of AI, but addressing the question of interoperability, the same technical committee already published a number of standards related to the communication and security requirements of electronic health records.

Moreover, ISO/TC 215 *Health Informatics* has initiated an ad-hoc group for the *Application of AI technologies in health informatics* in order to discuss the topic and identify standardization needs.

More generally, regarding the data quality aspects for ML, the technical sub-committee ISO/IEC JTC 1/SC 42 *Artificial Intelligence*⁸⁸ initiated a discussion on the topic within the big data working group. Questions of trustworthiness, such as the need of explainability, are also being explored inside this sub-committee by a dedicated working group. Specific guidelines for the trustworthiness of healthcare and other sensitive applications may also be proposed.

84 <https://www.iso.org/committee/54960.html>

85 <https://www.iso.org/standard/71923.html>

86 <https://www.iso.org/standard/75956.html>

87 <https://www.iso.org/standard/75310.html>

88 <https://www.iso.org/committee/6794475.html>

Examples from Luxembourg

In Luxembourg, research centers and hospitals are working on various applications of AI in healthcare. Using new technology, progress has been made in the areas of neurodegenerative disorders, especially Parkinson's disease, and also in handling brain tumors and diabetes.

Example 1

For example, the Luxembourg Centre for Systems Biomedicine (LCSB) is exploring molecular and cellular mechanisms of human diseases. The center is focusing on predictive, preventive and personalized medicine. One of its research areas is the detection of Parkinson's disease (PD) at early stages and suggesting an alternative treatment. For example, a link between human immune cells and the progression of PD was established. Another initiative of LCSB in collaboration with the Systems Biology Institute in Tokyo resulted in the development of a PD map⁸⁹ that allows for an efficient representation of data on various aspects of PD pathogenesis. Such a representation addresses the challenge of data quality, which is crucial for data exploration and analysis as well as knowledge generation.

Example 2

The Luxembourg Institute of Health (LIH) leads clinical trials offering new innovative therapeutic approaches. For example, one of its departments working on experimental clinical cancer research uses neuroimaging for efficient diagnosis and treatment of brain tumors. Based on multi-modal imaging systems and advanced data analysis, neuroimaging contributes to the establishment of relevant disease biomarkers used to improve disease management and patient care. Another department is working on raising the quality of life and health of people. One of its projects aims at using AI to generate digital twins of patients with diabetes in order to predict more accurately the course of the disease and develop better, personalized treatment strategies.

Example 3

One of the applications of AI in the National Health Laboratory (LNS) is for the diagnostics of tumors. For example, LNS uses Next Generation Sequencing that allows fast and less expensive deciphering of a person's DNA leading to a discovery of genetic personal information and more precise pathology diagnostics. Next Generation Sequencing has been possible thanks to AI and Cloud Computing and is typically used by LNS to run tests on samples of malignant tumors. Combined with next generation of machine-learning-based tumor classifiers, Next Generation Sequencing has a potential to transform tumor pathology.

89 https://wwwfr.uni.lu/lcsb/research/parkinson_s_disease_map

4.3.2. Banking and finance

For more details on the usage of AI in the banking and financial sector, the reader can refer to [10] [20] [21], which were used as a source of information for this summary.

Motivation

- The operating models of financial institutions are being fundamentally reshaped, switching to personalized experiences and higher customer retention.
- High competitiveness and high cost of services.
- Availability of multiple sources of data allowing for an augmented performance.
- Increased regulatory burden.

Top high-potential use cases

- Personalized financial advice and planning.
- Fraud detection and anti-money laundering.
- Front, middle and back-office optimization (client interaction, augmented risk management, regulatory compliance, etc.).
- Increasing trading speed.
- Better investment performance.

Currently on the market

- AI-powered credit scoring and automated personal and business loans.
- Fraud detection.
- Process automation supporting legal compliance, bill-payment activities, investment monitoring and reporting, etc.
- Macroeconomic trends identification through ML (for investment, trade, etc.).
- Chatbot support for sales agents.
- Dynamic customization to improve customers' financial positions, to understand investors' profiles and preferences in real-time, to evaluate trading risks, etc.

Promising technology

Various AI techniques can be used by financial institutions to improve their services, such as ML, graph theory, etc. Predictive algorithms are particularly helpful since they help financial institutions to project themselves into the future and anticipate risky situations. Augmented real-time analytics supports better customer engagement and timely response to a rapidly changing environment, as for example in trading.

Neural networks and deep learning are appreciated for their capacity to deal with vast quantities of data and identify relevant factors and their relationships. Reportedly, the outcomes of such a fine-grained analysis have already improved the accuracy of credit scoring systems.

Deploying natural language processing (NLP) is expected to support different financial services. It could help understanding regulatory documents and extracting compliance rules, thus significantly decreasing compliance costs. NLP could support automated dashboard creation by analyzing complex and non-standard financial documents, identifying relevant information and presenting it in a summarized form. Moreover, NLP is essential for chatbots to speed-up clients' onboarding and/or support. In addition, NLP is used for sentiment analysis on social media to track customers' satisfaction and foster their involvement.

Image recognition technologies enhance mobile banking thanks to user authentication via fingerprint or facial recognition using images captured by smart phone cameras.

Potential benefits

AI could benefit financial institutions in multiple ways, starting from the optimization and acceleration of existing operations to the proposal of radically new services or solutions.

AI-based process automation could simplify and improve the efficiency of business-as-usual tasks. This would lead to the cost reduction of routine processes and might improve their quality. For example, loans could be obtained online within seconds and with better conditions. Regulatory compliance may become less burdensome for financial clerks.

AI allows the analysis of large volumes of data resulting in better insights, be it for customer support, investment, insurance, trading, etc., thus supporting smarter decision-making. Moreover, data could be analyzed in near real-time, which is needed for rapid actions (for example, trading) and allows for fast anomaly detection (for example, triggering a review of potentially fraudulent activities).

AI enhances customization and a more holistic approach to products and services. On the one hand, AI-based products and services could meet unique needs of customers and offer them better conditions (for example, more credit). On the other hand, AI could increase the efficiency and productivity of financial advisors by supporting the quality, relevance and scalability of their services.

Moreover, AI allows for 24/7 services and customer support, putting all the necessary information at the fingertips of customers in their preferred format and via their preferred channel.

Finally, financial institutions could differentiate from their competitors by offering new operating modes, ways of working and brand-new products and services that take into account the habits of their customers.

Challenges to be addressed

AI may be supporting financial institutions with regulatory compliance but it also gives rise to new challenges. Legal requirements towards transparency of financial services are high in some countries, thus auditability, interpretability and explainability of AI-powered products and services should be guaranteed. Moreover, usage of various types of data from various sources by financial institutions may raise concerns about privacy. Consumers should be provided an easy way to manage consent and authorization to the use of their data.

While AI can bring advantages to some categories of customers in the financial ecosystem, other customers may be subject to discrimination. For example, if bias were introduced in an AI system (via unrepresentative data, in development or at some other stage) the decisions based on its recommendations could discriminate against or exclude marginalized groups of individuals. Ethical usage of AI by financial institutions should be promoted and closely monitored.

In the rapidly changing environment, the behavior of some AI-powered products and services may be unpredictable, which is a risk that needs to be considered and mitigated.

With process automation and real-time services supported by AI, there is evidence that new forms of cybercrime and other threats are also enabled. For example, with the implementation of a real-time payment system a more than 100% increase in fraud (referred to as “real-time fraud”) was registered in the UK.

Moreover, as in other industries where routine tasks are subject to process automation, there is a concern about AI's impact on the workforce. Workers should be trained with new skills and capabilities to be able to perform new tasks.

Benefits of standardization

Data is crucial to enable various AI-based services proposed by financial institutions. Standards for business financial documents help to ensure data quality, accuracy and availability and make verification of these documents less time-consuming. Moreover, they help to increase interoperability and facilitate data sharing among institutions. A technical committee ISO/TC 68 *Financial services*⁹⁰ has initiated a number of standards to introduce standard reference data for financial services and define financial information exchange protocol.

With respect to privacy and security protection, a *Security, Infrastructure and Trust Working Group (SIT WG)* of the Financial Inclusion Global Initiative (FIGI)⁹¹, led by ITU-T, published in late 2019 a technical report entitled “Big data, machine learning, consumer protection and privacy”. This report explores various challenges that big data and ML raise in the area of consumer and data privacy protection for the provision of digital financial services.

Finally, general considerations on the ethical and societal concerns raised by AI, which could also be applied to financial sector, are discussed in ISO/IEC TR 24368 *Artificial intelligence – Overview of ethical and societal concerns*⁹².

Examples from Luxembourg

Example 1

Various regulatory bodies — including the EBA, EU and national regulators — issue numerous regulations, circulars and updates throughout the year. This complicates compliance efforts, especially for entities that are highly regulated. Since the 2008 financial crisis, the amount of regulation for financial institutions has increased considerably, putting them in a situation where they constantly need to adapt to new or evolving regulatory requirements.

Given the regulations' complexity and volume, compliance is a financial institution's most challenging and cost-intensive area. Even with considerable effort, it is barely possible to review all regulations, retrieve the relevant parts and keep track of their implementation status. Frequent amendments and deletions result in a tedious process to identify which articles of the law are relevant and how it has evolved.

In order to facilitate the regulatory compliance practice of financial services institutions, KPMG developed a tool to analyze legal documents. With only a few keywords, their AI-powered search engine browses and bundles together relevant legal documents from various sources (i.e. EU and CSSF) that are semantically interlinked. The tool considers and retrieves the various regulations in chronological order, allowing the user to rapidly review the findings. Thus, the framework allows compliance officers to gain a holistic overview of both European and national legislation.

Behind the scenes, the combination of semantic web and natural language processing (NLP) technologies allow to:

- Identify all regulatory requirements that apply to the client's unique business,
- Find cross-references among regulations,
- Automatically classify and tag regulations.

90 <https://www.iso.org/committee/49650.html>

91 <https://www.itu.int/en/ITU-T/extcoop/figisymposium/Pages/FIGISITWG.aspx>

92 <https://www.iso.org/standard/78507.html>

Moreover, the tool's user-friendly interface allows users to browse various regulations, select relevant excerpts and extract them into a new document to complete a specific compliance task. To do so, KPMG put in place the following mechanisms:

- The automatic retrieval and categorization of legal documents using ML classifiers,
- The structuring of the retrieved regulations thanks to metadata and semantic annotations using NLP,
- The mapping of keyword searches to the semantics of the text to retrieve relevant information and recompile it to a new document.

The tool is currently at prototype stage and already shows the following benefits:

- It accelerates the regulatory compliance exercise by collecting the relevant sections of regulations;
- It provides a complete retrieval of information that could otherwise only be manually achieved by exhaustive search across various web sources.

To complement the regulatory compliance tool, and in case of litigation, KPMG developed an AI-powered platform for legal search in Luxembourg case-law. The platform allows its users (including lawyers, law firms, tax advisors, in-house counsels and legal departments) to search in the judicial conclusions of similar past cases and make an informed professional judgment for each new case. The solution predicts court case outcomes, case durations, judicial tendencies and legal trends.

Example 2

The processing of tax documents is a routine task in many organization. It could be rather time-consuming if done by human operators, and does not bring much added value to the employees to whom it is assigned. To address the issue, KPMG developed a tool for digitalization of tax documentation process.

The tool allows the automated extraction of relevant tax information from paper documents and its storage on a dedicated platform. Such process automation eliminates the risk of incorrect input due to manual human data entry error. At the same time, it helps to accelerate the process. After automation, the human operators stay in the loop to control the quality and manage potential exceptions, bringing their added value to the process. They also have more time to focus on other important to organization activities.

The processing of tax documents requires the accomplishment of the following steps, if done by humans:

- Collect tax documents provided via emails or file exchange platforms,
- Extract information from a scanned PDF and manually input it into selected fields of an external database/platform that acts as a communication channel with external parties,
- Save the PDF document to the external platform.

KPMG fully digitalized this process. Intelligent Optical Character Recognition (OCR) technology that incorporates ML algorithms was used to extract data. Robotic Process Automation (RPA) was used to insert the relevant data and documentation into the external platform. The process is sufficiently robust to cope with different templates (putting relevant information on different pages) thanks to ML algorithms.

4.3.3. Manufacturing

For more details on the usage of AI in manufacturing, the reader can refer to [10] [16] [22] [23] [24], which were used as a source of information for this summary.

Motivation

- Absorbing increasing demand, especially coming from emerging countries.
- A growing need for flexibility due to market fragmentation based on consumers' taste for customization.
- Safer/better work conditions.
- Lower production and maintenance costs.

Top high-potential use cases

- Enhanced monitoring and auto-correction of manufacturing processes.
- Optimization of supply chain and stock management.
- On-demand production.
- Predictive maintenance using fault prediction and fault localization.
- Collaborative robots automatically adapting to the environment.

Currently on the market

Manufacturing is one of the most advanced in terms of the adoption of AI. Multiple applications have already been deployed by AI pioneers in the domain:

- Using digital twins in combination with AI for more efficient equipment maintenance and performance optimization,
- Using ML for better end-to-end supply chain and stock management,
- Using adaptive collaborative robots for a large number of production processes to ease the burden on human workers,
- Using ML and simulated environments for product design to improve quality and speed-up the production process,
- Predictive maintenance and personalized customer service packages.

Promising technology

A wide range of AI techniques can be used in manufacturing, both ML-based and logic-based. Expert systems with domain-specific rules can be used in maintenance. ML can be applied in maintenance but also in supply chain management, performance prediction, collaborative robots, etc. Linear regression, Monte Carlo or ensemble methods (such as random forest models) are frequently used in this domain, with neural networks gaining in popularity. The type of techniques to choose would depend on the amount of data available, on the quality of that data and on the type of problem. Currently, the training of ML models is mainly done on historic data but with the improvement of technology and its increasing adoption by organizations, real-time data processing would be integrated in decision support, performance measurement and service optimization.

With respect to specialized AI systems, computer vision and robotics are the most impactful in manufacturing. Coupled with the IoT as data source, computer vision can be used in the assembly line for production optimization and predictive maintenance. It can also be exploited by collaborative robots for gesture recognition and consecutive behavior adaptation. Thus, collaborative robots could become more than simple process automation support.

Potential benefits

Using AI can bring innovation to the whole manufacturing process, from product design to after-sales services, and significantly cut costs.

Through the analysis of defective products and customers' feedback, new, better and more attractive products could be designed. This will not only improve customer satisfaction but also lower quality control costs.

Production efficiency could be increased thanks to predictive insights of demand, assembly line optimization and the additional control gained over the supply chain. Production will thus be faster, more flexible and scalable, and waste would be reduced.

Process automation and the use of collaborative robots could significantly improve working conditions, especially when it comes to hazardous activities. At the same time, it could help avoid human error and improve quality.

Predicting defects or production interruption could lead to timely maintenance. This could not only reduce maintenance costs but also the time and money potentially lost because of the interruption. Thus, after-sales services could be more efficient, timely and cost-efficient. Moreover, it would also contribute to the safety of the work environment.

Challenges to be addressed

Although multiple applications are already in operation, many companies are still considered unprepared to embrace AI. For them to benefit fully from AI, a collaboration between supply chain professionals and manufacturers is required, as well as them being equipped with the necessary technology. Only the largest and best-resourced organizations are currently up to speed. Others need to update their infrastructure, and prepare and implement the change management strategy.

One of the concerns is related to unemployment due to process automation. Routine and low-skilled jobs would disappear. However, companies would need more qualified individuals to work with the new technology. Thus, it is recommended that they invest in training their workers.

Another challenge related to automation and collaborative robots is safety. Robots must not cause injury to humans, and this becomes particularly challenging in complex environments with several robots interacting with humans and each other.

One more concern against automation is the potential loss of human control resulting in the abuse of the manufacturing process. It must be ensured that people retain control over machines at all times and that the algorithms behave as intended. Robustness of data acquisition and algorithms is thus of high priority.

Finally, real-time automation could result in legal liability questions in case an accident occurs. The algorithms should be transparent and explainable in order to identify the source of any accident and the responsible party.

Benefits of standardization

The International Electrotechnical Commission (IEC) has a currently active technical committee IEC/TC 65 *Industrial-process measurement, control and automation*⁹³ that develops, among others, standards related to the issues of security and safety raised by manufacturing automation. So far, AI was not considered in those standards. However, given the recent developments in the domain, the group plans on exploring the use of AI for industrial system safety and security. Moreover, the committee initiated standards related to the use of IoT in industrial facilities and, more generally, describing a "digital factory" framework. Another committee, ISO/IEC

93 https://www.iec.ch/dyn/www/?p=103:7:::;FSP_ORG_ID:1250

JTC 1/SC 41 *Internet of things and related technologies*⁹⁴, also develops standards for industrial IoT systems and devices in order to guarantee their compatibility and interoperability and provide a framework for real-time data exchange.

Another technical committee under responsibility of ISO, ISO/TC 299 *Robotics*⁹⁵, develops a number of standards related to the use and assessment of robots specifically in manufacturing, since this is considered a high-potential use case. For example, the topics under development include the manipulation of industrial robots, safety requirements for industrial robots, performance criteria for service robots, etc. Robots could have AI components in them or not, which is why ISO/TC 299 is connected with ISO/IEC JTC 1/SC 42 *Artificial Intelligence*⁹⁶ in order to integrate the relevant materials in the robotics standards.

Preventive maintenance and other services could rely on ad-hoc software with AI components. A sub-committee ISO/IEC JTC 1/SC 7 *Software and systems engineering*⁹⁷ discusses the additional requirements and constraints that should be applied to AI-based software as compared to traditional software. Experts from ISO/IEC JTC 1/SC 42 joined these efforts and established a number of dedicated working groups. Moreover, the issues of safety of AI systems based on neural networks are being addressed in an ongoing project ISO/IEC TR 24029-1 *Artificial Intelligence (AI) — Assessment of the robustness of neural networks — Part 1: Overview*⁹⁸.

Examples from Luxembourg

Example 1

In Luxembourg, Tarkett uses ML algorithms in the production line of floor paint.

In the process of creation of a floor paint, the same product (color) must be produced several times per month. The challenge is to obtain the same shade each time the paint is produced since even minor differences in color cause difficulties in the exploitation of the final product.

To address this challenge and improve the quality of the product, it is possible to use standardization of color references. Inks will be managed at the beginning of the process to create a specific color using the standard of the International Commission on Illumination - CIE (CIE 1976 / ISO 11664-4:2008⁹⁹). On the one hand, it allows to encode each color in a three-dimensional color space, which can be easily processed by computers. On the other hand, it allows to reduce differences between colors that are perceptible to the human eye.

At the beginning of a new production process, a computer automatically defines the paint color based on a sample from the previous production. It calculates the right amount of every composing color to be mixed in order to reach the perfect shade. The second step consists in managing the color in the production of the paint. In order to do so, the paint sample is produced and used on vinyl sheets. The painted vinyl is compared to the one from the previous production. The main idea is then to analyze the color composition of both vinyls using a k-means algorithm, which allows measuring the distances between colors in the three-dimensional color space. The color composition is then adjusted to reduce these distances, resulting in a new paint having no difference in color visible to the human eye.

94 https://www.iec.ch/dyn/www/?p=103:7:0:::FSP_ORG_ID:20486

95 <https://www.iso.org/committee/5915511.html>

96 <https://www.iso.org/committee/6794475.html>

97 <https://www.iso.org/committee/45086.html>

98 <https://www.iso.org/standard/77609.html>

99 <https://www.iso.org/fr/standard/52497.html>

4.3.4. Transportation and Automotive sector

For more details on the usage of AI in transportation, the reader can refer to [10] [22] [23], which were used as a source of information for this summary.

Motivation

- Increasing traffic, leading to congestion and pollution.
- Saturated infrastructure.
- Multiple fatalities arising from human error.

Top high-potential use cases

- Traffic management and congestion control.
- Enhanced safety and security.
- Autonomous fleets for ride sharing.
- Semi-autonomous features such as driver assistance.
- Engine monitoring and predictive, autonomous maintenance.

Currently on the market

- Automated driver assistance systems (for example, parking assist, lane centering, adaptive cruise control, etc.).
- Metro trains and/or monorails running driverless.
- Automated scheduling of infrastructure maintenance.

Promising technology

Autonomous vehicles rely on a combination of various technologies, such as intelligent path planning, computer vision and global positioning systems. Real-time traffic control and prediction of traffic situations allow for route optimization, which is beneficial for passengers but also helps to avoid congestions and reduce emissions. Computer vision allows for the identification and analysis of surrounding objects (for example, vehicles, pedestrians, road signs, etc.), which is essential for the safe operation of autonomous vehicles.

Computer vision, including image and video processing, can be used for the detection of violations of rules and regulations by both pedestrians and vehicles. It can be coupled with voice alerts to prevent such violations or support traffic management.

Predictive algorithms could help transport companies and public administrations to optimize the maintenance of vehicle fleets, equipment and infrastructure.

Potential benefits

AI can contribute to traffic efficiency. Analyzing real-time traffic conditions, it could select optimal roads, assist in traffic signal management, suggesting alternative means of public transport, etc. Thus, traffic volumes could be reduced, congestion could be avoided and people could move around faster and with greater flexibility. This would not only improve the quality of living but also reduce air pollution and contribute to environment protection.

Moreover, with the support of assisted and eventually autonomous driving, people would be relieved from stress and fatigue, especially for a long-distance journey.

AI could contribute to road safety. It could assess the condition of the car, the surrounding environment and the status of the driver and then support the driver or switch to auto-pilot when the human is incapable of driving. Traffic optimization also contributes to road safety since it reduces the risk of an accident. Moreover, AI could help with the identification of violations of traffic regulations by means of road surveillance. AI embedded into a traffic robot could support law enforcement on challenging road intersections: on the one hand, it could guide pedestrians (for example, via voice alerts and safety messages); on the other hand, it could deal with vehicles (for example, by giving arm and light signals).

Finally, AI could assist the public organization of transportation through preventative fleet-and-infrastructure maintenance and thus increase systems' safety and reliability. Indeed, thanks to the collection of data from multiple sources (sensors, human observations, etc.) AI could help with the identification and prevention of likely system failures or infrastructure deterioration.

Challenges to be addressed

Safety and security are the major concerns for users of smart transportation. Despite the potential that AI has in terms of improving road safety and security, documented accidents of autonomous cars have also shown its vulnerability. The technology still needs development to perform safely under certain conditions (weather, time of the day, surroundings, etc.), and correctly identify individuals and objects.

And even if the technology is in place, it would need to gain consumer trust and regulatory acceptance. There has been progress regarding the authorization of some types of driver assistance. Nevertheless, for fully autonomous vehicles there is a number of considerations, including among others a liability issue (who is responsible in case of errors and/or fatalities) and data privacy (knowing individuals' travel history, etc.).

One more challenge is related to the infrastructure required for an efficient autonomous fleet. The volume of data generated by one single autonomous vehicle is huge, and for multiple autonomous vehicles on the road, there is a challenge to transmit, store and analyze all the generated data efficiently. Moreover, traditional general-purpose CPUs lack the necessary performance, and dedicated accelerators are needed.

Benefits of standardization

A goal of standards is to set up common grounds for the levels of security required for secure transmission, storage and processing of data by autonomous vehicles. To this end, the technical committee ISO/TC 204 *Intelligent transport systems*¹⁰⁰ has initiated a study related to the usage of Big Data and AI in transport systems. In particular, the committee is considering the multi-parted standard ISO/IEC 20547 on *Big data reference architecture*¹⁰¹ and the potential impact of AI on intelligent transport systems. European technical committees, ETSI *Intelligent Transport Systems (ITS)*¹⁰² and CEN/TC 278 *Intelligent transport systems*¹⁰³, also work on data sharing, interoperability and security issues in the intelligent transport systems that are aligned with European legislation.

Moreover, in order to gain public trust and provide the necessary support to regulators, a benchmark comparing the performance of AI with a careful human driver could be defined. To address this point, ITU-T established in late 2019 a Focus Group *AI for autonomous and assisted driving (FG-AI4AD)*¹⁰⁴. The objective of the group is to determine a minimal performance threshold for vehicles using AI as a driver in accordance with the 1949 and 1968 Convention on Road Traffic of the UNECE Global Forum for Road Safety.

100 <https://www.iso.org/committee/54706.html>

101 <https://www.iso.org/standard/71277.html>

102 <https://www.etsi.org/technologies/automotive-intelligent-transport>

103 https://standards.cen.eu/dyn/www/f?p=204:7:0:::FSP_ORG_ID:6259&cs=1EA16FFFE1883E02CD366E9E7EADFA6F7

104 <https://www.itu.int/en/ITU-T/focusgroups/ai4ad/Pages/default.aspx>

Examples from Luxembourg

Example 1

In Luxembourg, the Interdisciplinary Center for Security, Reliability and Trust (SnT) of the University of Luxembourg has been working on various aspects related to the safety of autonomous vehicles. Among the topics addressed, one finds detection and avoiding of obstacles (pedestrians or vehicles), testing performance and robustness of Deep Neural Networks used in the context of automated driving systems, etc. Moreover, SnT developed one of the autonomous vehicles being tested under the cross-border mobility project between Luxembourg, France and Germany. Indeed, the three countries set up a testbed in order to offer the developers of connected and autonomous vehicles an opportunity to do trials on open roads under real-life conditions and in a cross-border environment. The testbed will allow to make progress on various challenges related to the adoption of autonomous vehicles, such as uninterrupted connectivity for flawless data capture and transmission (for example, the 5GCroCo¹⁰⁵ project), varying rules for speed limits, etc.

Example 2

On a similar note, Luxembourg city, along with Copenhagen, Geneva and Lyon, is involved in full-scale trials of autonomous vehicles under the umbrella of the European project entitled AVENUE¹⁰⁶. The project aims at the validation of the usage of autonomous vehicles as a means to complement the public transport in urban and suburban regions.

4.3.5. Government and public sector

For more details on the usage of AI in public sector and government services, the reader can refer to [21] [25] [26], which were used as a source of information for this summary.

Motivation

- Reduce the burden of repetitive tasks and length of procedures.
- Improve public service productivity and efficiency.
- Support all groups of population, including the most vulnerable.
- Support sustainable development goals¹⁰⁷.

Top high-potential use cases

- Making legal documents and other governmental information more accessible to users.
- More agile and effective interactions with public authorities.
- Improving the efficiency of law enforcement, by detecting and anticipating criminal activities and/or by identifying at-risk residents.

¹⁰⁵ <https://5gcroco.eu/>

¹⁰⁶ <https://h2020-avenue.eu/>

¹⁰⁷ <https://sdgs.un.org/goals>

Currently on the market

- Suggesting patrol routes for security checks for police officers.
- Connecting vulnerable residents with the services that could support them.
- Targeted inspection of businesses to ensure the quality of service.
- Annotated legal documents/public websites for easier information access.
- Pre-processing of citizens' requests to accelerate decision-making.

Promising technology

Natural language processing (NLP) is a powerful technique when processing and analyzing written documents. It can support automatic question answering and information extraction. On the one hand, this technology can be applied to governmental documents (legal, informative, etc.) in order to facilitate the access to information by citizens. On the other hand, governmental organizations could use it to accelerate the processing of citizens' requests and decision-making related to them.

Classification techniques can be used to analyze businesses' behavior and highlight the ones that need inspection. These types of techniques also can be applied to citizens' data in order to identify if they belong to a vulnerable group of society, represent a danger to others, etc. Classification techniques could help maintain consistent and human-friendly web services for citizens by providing recommendations on the best placement of new documents and information. In this case, the available data is not labeled and thus cannot be used for classification, clustering techniques could be applied in order to identify groups of businesses or citizens with similar behavior.

In case of surveillance, neural networks provide promising results for image recognition, which is not limited to facial recognition but also uses alternative information (such as height, postures, clothing, etc.) about subjects. Moreover, to cope with poor image quality, image-sharpening technologies are used: neural networks are trained on typical features of physical objects (for example, hair, building materials, etc.) and then use complementary information about these features (for example, texture) for better recognition.

Potential benefits

AI could benefit the public sector by increasing its efficiency. By supporting the simplification and speeding-up of its processes, it allows at the same time to enhance the public sector's productivity. With the support of AI, some tedious and repetitive tasks can be done automatically and free the time of public servants for more interesting and complex tasks or those requiring human interaction. When dealing with large volumes of paperwork, AI could help reduce errors and improve the quality of results.

When it comes to the security, quality, injury, fraud or any other type of check, AI could optimize issue detection by suggesting new strategies for checks. It could also help prevent some issues by identifying patterns in large volumes of data.

With the support of AI, it is also possible to strengthen social welfare programs by attaining optimal inventory levels at health and social services and allowing for more human interaction, and thus more personalization.

Challenges to be addressed

The use of AI raises issues with respect to privacy protection. The public sector being by design the one to protect the privacy of the citizens via policies and legislation, it should be particularly careful when dealing with personal data itself.

Moreover, when it comes to the decision making supported by AI, especially in judicial affairs, there are high concerns related to bias and transparency. There have been cases where bias towards people of color was introduced in the police system. These cases did not help with building trust towards the use of AI, especially in critical situations. Thus, transparency and explainability of AI-based results are crucial points to be addressed when using AI in the public sector.

Benefits of standardization

At a high level and transversal level, a standard ISO/IEC 38507 on the *Governance implications of the use of AI by organizations*¹⁰⁸ could support any type of organization, including governmental ones, with the identification of relevant stakeholders along with their roles and responsibilities. This information could then be used to determine the liability of parties using AI.

Moreover, ISO/IEC 23894 on *AI Risk Management*¹⁰⁹ could support governmental organization when dealing with risks associated with AI.

In order to deal with bias, an ISO/IEC TR 24027 *Artificial Intelligence (AI) — Bias in AI systems and AI aided decision making*¹¹⁰, can be considered as a reference of good practices to be applied.

Finally, when it comes to the automatic processing of documents, the standards related to metadata could be helpful. For example, there are two series of standards developed by ISO/IEC JTC 1/SC 32 *Data Management and interchange*¹¹¹: ISO/IEC 19583 on *Concepts and usage of metadata*¹¹² and ISO/IEC 11179 on *Metadata registries*¹¹³. They aim at facilitating and harmonizing the description of all types of data. For particular types of data, there exist specialized metadata frameworks. An example of such a specialized framework is the *European Legislation Identifier (ELI)*¹¹⁴, originating in Luxembourg and offering a standardized labeling template for legal documents in order to access, exchange and reuse legal data across borders.

Examples from Luxembourg

Example 1

The Central Legislative Service (SCL) of the Ministry of State has set as an objective to help citizens and legal experts interact with laws and regulations in Luxembourg. With this objective in mind, SCL teamed up with the Interdisciplinary Center for Security, Reliability and Trust (SnT) of the University of Luxembourg to combine legal and IT expertise and provide a tool that would facilitate the demonstration of legal compliance.

The first challenge the team faced was the transformation of existing legal documents, typically in print-only formats such as PDF, into web-enabled and machine-analyzable formats that would allow more advanced and computer-assisted forms of interaction. The complexity of this task is due to the flexibility of natural language, variations in the organization of legal texts and multiple cross-references that link different provisions. Given the volumes of existing legal texts, doing this task manually while maintaining the quality would be very time-consuming. This is why this task was automated by using such technologies as natural language processing (NLP), Semantic Web and information retrieval.

108 <https://www.iso.org/standard/56641.html>

109 <https://www.iso.org/standard/77304.html>

110 <https://www.iso.org/standard/77607.html>

111 <https://www.iso.org/committee/45342.html>

112 <https://www.iso.org/standard/67365.html>

113 <https://www.iso.org/standard/61932.html>

114 https://ec.europa.eu/isa2/solutions/european-legislation-identifier-eli_en

Transforming legal documents into machine-analyzable formats means annotating them with legal metadata. There exist different types of legal metadata: administrative (creation date, authors, etc.), provenance (origins, such as parliamentary discussions, etc.), usage (application in case law, jurisprudence, etc.), structural (hierarchical organization of the text), and semantic (fine-grained information about meaning and interpretation of legal provisions, such as application modalities, exceptions, sanctions, etc.). Most important for the task pursued by SCL are the last two types, structural and semantic metadata.

In order to perform structural referencing and annotation, the European Legislation Identifier (ELI) is used. Luxembourg (SCL) was at the origins of ELI, but the initiative is EU-endorsed and used by multiple countries. ELI defines an intuitive framework for the labeling of parts of legal documents. It attributes to each such document a universal resource Identifier (URI)¹¹⁵, metadata and RDF¹¹⁶ information to render the data computer-readable. SnT developed a tool allowing the annotation of legal documents following the ELI standard as well as the resolution of cross-references across documents.

The next step was the semantic annotation of documents. Indeed, semantic metadata is a prerequisite for the automatic extraction of legal requirements for compliance. However, there was no harmonized view on semantic metadata types. Thus, SnT proposed a conceptual model for semantic metadata types that would allow for the annotation of pertinent legal requirements. They then suggested and implemented automated extraction rules for these metadata types based on advanced linguistic analysis exploiting the full potential of existing NLP techniques. As a result, a selection of legal documents could be annotated with relevant semantic metadata and meaning.

The third step is concerned with extending the knowledge base. SCL had already chosen an RDF implementation, a conceptual model for semantic metadata types. RDF is a popular Resource Description Format that was proposed by W3C in the framework of their Semantic Web recommendations. It has the advantage of being complemented by RDF Query Language (SPARQL)¹¹⁷, a standard that was also developed by W3C. SPARQL is used to extract relevant information from the RDF-annotated text, but it is not user-friendly for those with no technical IT background.

The legal requirements extraction was tested on the Income Tax Law of Luxembourg and Traffic Regulations. The outcomes of this test showed that the usage of legal metadata is promising for smart legal search. The main challenge that needs to be addressed here is how to extract and represent all the metadata information in a form that would enable users to query the information with simple questions. As an example, in the context of Traffic Regulations, the users should be able to ask questions about the penalties and fines associated with different traffic offences.

The existing portal of the Official Journal of Luxembourg¹¹⁸ integrates some of the results of the above-described work. The portal currently features multiple legal documents that are enriched with varying amounts of legal metadata.

115 <https://www.w3.org/Addressing/URL/uri-spec.html>

116 <https://www.w3.org/RDF/>

117 <https://www.w3.org/TR/rdf-sparql-query/>

118 <http://legilux.public.lu/>

4.3.6. Retail

For more details on the usage of AI in retail, the reader can refer to [8] [16], which were used as a source of information for this summary.

Motivation

- Development of e-commerce and on-line consumers' presence.
- Need for competitive advantage compared to other retailers.
- High expectations for personalized products and services.

Top high-potential use cases

- Personalized design and production.
- Better shopping experience (recommender systems, chatbots, etc.).
- Automated delivery.
- Optimized operations (warehousing, logistics, prices setting, advertisement, etc.).

Currently on the market

- Personalized product recommendations.
- Targeted promotion campaigns.
- Sales predictions resulting in better stock management and waste.
- Automated vehicles (trolleys, drones, etc.) for logistics (warehouse navigation, etc.).

Promising technology

Computer vision and natural language processing (NLP) could play an important role in innovation in the retail sector. On the one hand, computer vision could change the functioning of physical retail outlets. It could be used to identify repeat customers and send them real-time advertisement (recommendations, sales, etc.). It could also allow for a seamless automated "check-out" thanks to facial recognition and the linking of products taken from shelves to identified customers. Additionally, it could be used to collect feedback from customers by analyzing their facial expressions and recognizing their emotions. On the other hand, the analysis of images posted online (for example in social media) could help retailers understanding people's styles and preferences and send them personalized commercials.

NLP could be used to improve user experience when shopping online. Virtual assistants and chatbots could get better insights on the user's needs and provide personalized recommendations in a human-friendly manner based on a conversation with the user. Virtual home assistants could alert users that they are about to run out of a product and suggest buying more. Both language understanding and generation (including speech synthesis) would be useful in the retail sector.

Autonomous vehicles could be used to facilitate the preparation of goods in the warehouses, but also to deliver them to consumers, which is one of the challenges currently being addressed by retailers.

With the increasing volumes of data collected about consumers, deep learning could be used to analyze the data, detect patterns, and predict consumers' behavior.

Potential benefits

First, using AI could help retailers cope with the increasing demand for personalized products and services from consumers, be it tailored on-demand consumables, targeted promotion campaigns or online recommender systems. This would bring more customer satisfaction and increased profits due to more sales.

Real-time forecasting could help master the dynamic market environment. Based on the insights gained from data, retailers could adjust the assortments of their products and better manage supply. This could help reduce waste, especially in the food industry, where expiration dates are critical. Moreover, it could be used to set or adapt pricing. Generally, real-time predictions would allow cost reductions and potentially profits increase.

Optimized and automated operations, mainly in the warehouse and delivery process, could improve the quality of services, make the working conditions better for humans and reduce manual labor costs. Using computer vision in physical stores could accelerate the shopping process and thus serve more customers in the same amount of time, resulting in customer satisfaction and additional profits.

Challenges to be addressed

Adapting to the agile and tailored approach in product design and production would require an important effort from most retailers.

High levels of automation and switching to online services might lead to the significant reduction of human employers in the domain. Retailers may need to rethink their processes and train their teams to new skills.

Using high volumes of data from consumers results in the need to pay more attention to privacy protection and regulatory requirements in this domain. The way businesses deal with data could either build trust in their service or ruin their reputation.

Benefits of standardization

Given the challenge of privacy preservation in personalized services, standards related to data protection can be used. The technical sub-committee ISO/IEC JTC 1/SC 27 *Information security, cybersecurity and privacy protection*¹¹⁹ has a number of projects dedicated to data protection. They have also initiated a study regarding the impact of AI on privacy.

Moreover, a working group of ISO/IEC JTC 1/SC 42 *Artificial Intelligence*¹²⁰ collected a few use cases related to the retail sector and will analyze them to identify the standardization needs in the domain.

5. Key barriers to AI adoption

According to a study carried out among European companies in 2018, the main barriers to AI adoption are related to technical capacity, AI policy and regulatory risks, and business and social acceptance. Among different barriers, five were identified as the most constraining (see Figure 7) [8]:

- Lack of necessary skills or resources to manage or deploy AI,
- Data compliance, data privacy and data protection issues,
- Actual or expected employee resistance to deploying AI,
- Actual or perceived immaturity of the technology,
- Lack of widespread trust in AI's capability.

119 <https://www.iso.org/committee/45306.html>

120 <https://www.iso.org/committee/6794475.html>

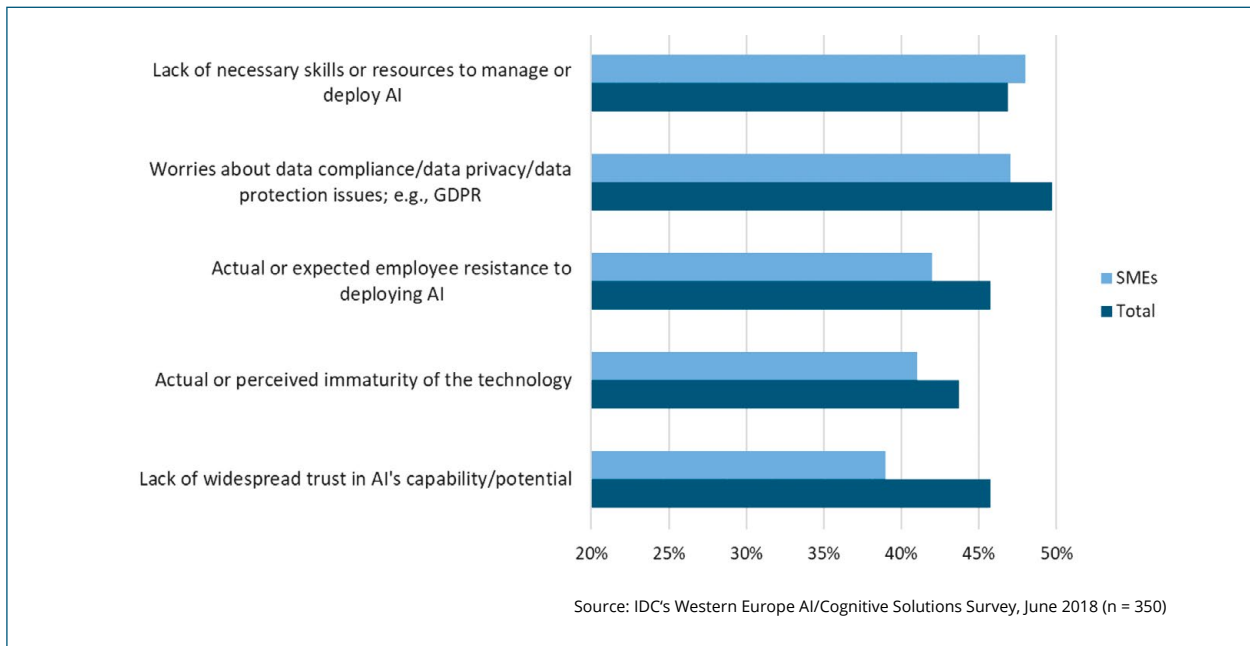


Figure 7: Key barriers to AI adoption in Europe [8]

As one of the actions towards addressing these barriers, the High-level Expert Group on AI produced a “Policy and Investment Recommendations for Trustworthy AI”. One part of these recommendations focuses on empowering humans by increasing knowledge about AI, raising awareness on AI and providing relevant education and training. Another part targets building a data and AI ecosystem, including research and industry capabilities as well as supporting infrastructure. Finally, an important role is given to establishing an appropriate governance and regulatory framework [15].

The investments in AI mentioned in previous sections (namely, Section 3) aim at bringing these recommendations to life. Regarding the regulatory initiatives, the European Commission has initiated a draft working document on liability in emerging technologies¹²¹ that states the challenges related to the use of these technologies, such as AI, and analyses the need of an evolution of existing laws, namely the Product Liability Directive¹²² and the Machinery Directive¹²³.

Standardization in the area of AI could also be considered as a means to achieve trust in the technology. At the international level, the technical committee ISO/IEC JTC 1/SC 42 *Artificial Intelligence* has established a dedicated working group on Trustworthiness (as already mentioned in Chapter 1). This group currently has 5 projects under development to provide an overview of trust-related challenges of AI and to detail some specific topics, such as bias and robustness. At the European level, ETSI has established a working group to address the security challenges of AI, while CEN-CENELEC has conducted a study regarding the need of specific European standards for AI. The next chapter will provide more information about the trust-related challenges of AI and the respective standardization activities.

121 <https://ec.europa.eu/digital-single-market/en/news/european-commission-staff-working-document-liability-emerging-digital-technologies>

122 Directive 85/374/EEC

123 Directive 2006/42/EC

References

- [1] MarketWatch, "Artificial Intelligence Market 2019 Share, Trends, Segmentation and Forecast to 2025," 08 2019. [Online]. Available: <https://www.marketwatch.com/press-release/artificial-intelligence-market-2019-share-trends-segmentation-and-forecast-to-2025-cagr-of-362-2019-08-12>. [Accessed 09 2019].
- [2] Public-Private Analytic Exchange Program, "AI: USING STANDARDS TO MITIGATE RISKS," 2018.
- [3] Tractica, "Artificial Intelligence Software Market to Reach \$118.6 Billion in Annual Worldwide Revenue by 2025," 04 2019. [Online]. Available: <https://web.archive.org/web/20190915133724/https://www.tractica.com/newsroom/press-releases/artificial-intelligence-software-market-to-reach-118-6-billion-in-annual-worldwide-revenue-by-2025/>. [Accessed 12 2020].
- [4] IDC, "Worldwide Spending on Artificial Intelligence Systems Will Grow to Nearly \$35.8 Billion in 2019," 03 2019. [Online]. Available: <https://web.archive.org/web/20191213191931/https://www.idc.com/getdoc.jsp?containerId=prUS44911419>. [Accessed 12 2020].
- [5] Allianz global Investors, "2018 Outlook for Global Artificial Intelligence," 2018.
- [6] WIPO, "WIPO Technology Trends 2019: Artificial Intelligence," World Intellectual Property Organization, Geneva, 2019.
- [7] Cognilytica, "Classification of the AI Vendor Ecosystem," 01 2019. [Online]. Available: <https://www.cognilytica.com/2019/01/16/cognilyticas-classification-of-the-ai-vendor-ecosystem-overview-and-bottom-3-layers/>. [Accessed 11 2020].
- [8] L. Delponte, "European Artificial Intelligence (AI) leadership, the path for an integrated vision," European Parliament, Policy Department for Economic, Scientific and Quality of Life Policies, 2018.
- [9] PwC, "2018 AI predictions: 8 insights to shape business strategy," 2018.
- [10] PwC, "Sizing the prize: What's the real value of AI for your business and how can you capitalise?," 2017. [Online]. Available: <https://www.pwc.com/gx/en/issues/analytics/assets/pwc-ai-analysis-sizing-the-prize-report.pdf>. [Accessed 11 2020].
- [11] European Commission, "Factsheet: Artificial Intelligence for Europe," 07 2019. [Online]. Available: http://ec.europa.eu/newsroom/dae/document.cfm?doc_id=51610. [Accessed 11 2020].
- [12] European Commission, "Trustworthy AI - Brochure," 09 2019. [Online]. Available: <https://ec.europa.eu/digital-single-market/en/news/trustworthy-ai-brochure>. [Accessed 11 2020].
- [13] European Commission, "The European AI Landscape," 2018. [Online]. Available: https://ec.europa.eu/knowledge4policy/publication/european-ai-landscape_en. [Accessed 11 2020].
- [14] High-Level Expert Group on Artificial Intelligence, "Ethics Guidelines for Trustworthy AI," European Commission, 2019.
- [15] High-Level Expert Group on Artificial Intelligence, "Policy and Investment Recommendations for Trustworthy AI," European Commission, 2019.
- [16] McKinsey Global Institute, "Artificial Intelligence: The next digital frontier?," McKinsey & Company, 2017.
- [17] I. Joshi and J. Morley, "Artificial Intelligence: How to get it right. Putting policy into practice for safe data-driven innovation in health and care.," NHSX, London, United Kingdom, 2019.
- [18] S. O'Neill, "Transforming medicine through AI-enabled healthcare," The Alan Turing Institute, [Online]. Available: <https://www.turing.ac.uk/research/impact-stories/transforming-medicine-through-ai-enabled-healthcare>. [Accessed 11 2020].
- [19] B. S. Abu-Nasser, "Medical Expert Systems Survey," International Journal of Engineering and Information Systems, vol. 1, no. 7, pp. 2018-224, 2017.
- [20] World Economic Forum (in collaboration with Deloitte), "The New Physics of Financial Services: Understanding how artificial intelligence is transforming the financial ecosystem," 2018.
- [21] OECD, "Artificial Intelligence in Society," OECD Publishing, Paris, 2019.
- [22] The Economist (Commissioned by Google), "Risks and rewards: Scenarios around the economic impact of machine learning," The Economist Intelligence Unit Limited, 2017.
- [23] IEC, "Artificial intelligence across industries," 2018. [Online]. Available: <https://basecamp.iec.ch/download/iec-white-paper-artificial-intelligence-across-industries-en/>. [Accessed 11 2020].
- [24] Partnership on AI, "AI, Labor, and Economy: Case Study Compendium," [Online]. Available: <https://www.partnershiponai.org/compendium-synthesis/>. [Accessed 11 2020].
- [25] IBM Center for the business of government and the Partnership for Public Service, "The Future Has Begun: Using Artificial Intelligence to Transform Government," 2018. [Online]. Available: <http://www.businessofgovernment.org/blog/future-has-begun-using-artificial-intelligence-transform-government>. [Accessed 11 2020].
- [26] "A guide to using artificial intelligence in the public sector," UK's Government Digital Service and Office for Artificial Intelligence, 2019. [Online]. Available: <https://www.gov.uk/government/collections/a-guide-to-using-artificial-intelligence-in-the-public-sector>. [Accessed 11 2020].

NOTE: While any hyperlinks included in this chapter were valid at the time of consultation, ILNAS cannot guarantee their long-term validity.

4

Towards trustworthy Artificial Intelligence

1. Introduction

Lack of **trust** is listed as one of the key barriers to Artificial Intelligence (AI) adoption in Chapter 3. From a software perspective, trust could be defined as the “degree to which a user or a stakeholder has confidence that a product or system will behave as intended” [1]. However, when talking about AI, and especially about continuously learning systems, the intended behavior could be difficult to define. Moreover, it is only one of the components that lead to the instauration of trust in AI; indeed, it is complemented by the ethical use of the technology, its explainability, its robustness, etc. This chapter will talk about **trustworthiness** of AI as the ability to meet stakeholders’ expectations in a verifiable way. Trustworthiness can be assessed through various characteristics and applied to services, products, technology, data and information as well as in organizations [2].

Given the multitude of angles by which to inspect trustworthiness, this document considers a layered architecture of trust. The layers include infrastructure, software and systems, and organizations.

The chapter first provides an overview of these three layers in order to introduce the concept to the reader. Then, the trustworthiness concerns pertinent to each of these layers are explored (for example, robustness, transparency, and bias are covered in the software and systems layer). The document describes both the concerns and possible ways forward to address them.

Standards play an important role in building trust in any technology, AI not being an exception. On the one hand, they serve as a technical reference to stakeholders. On the other hand, they describe the relevant processes to put in place by organizations. In some cases, standards can be used as an assessment basis to show compliance with legislation. Throughout this chapter, references are provided to the existing or under development standards to set the scene for the reader.

2. Layers of trust

It is common to talk about functional layers of an IT system in ICT world. It is natural to explore the risks, threats and vulnerabilities that could undermine the trustworthiness of each layer, and thus of an IT system as a whole. For example, in the ITU-T report on Trust provisioning, three **layers of trust** are distinguished: physical trust, cyber trust and social trust [3].

Physical trust relates to the safety and the reliability of physical things, such as the hardware used in the system, etc. For the purpose of this document, the computational environment and architecture are considered as contributing to the physical trust. Thus, the related issues include the misbehavior of sensors used as a data source, a faulty computing environment, etc. [2] [3] [4]. Physical trust is achieved at the **infrastructure layer**, covered in Section 3.

Cyber trust reflects various trust aspects related to cyber objects, such as AI system (software, data, etc.) [3] [4]. The AI-based system should satisfy a number of software quality criteria to be trustworthy, including reliability, verifiability, safety, etc. Some of these quality criteria, such as fairness, transparency, privacy, etc., also help to address ethical and societal concerns [2]. These issues, in turn, often relate to the quality of data that therefore contributes to the establishment of cyber trust. Various cyber trust challenges are addressed at the **software and system layer**, discussed in Section 4.

Social trust rest upon the notion of user trust, which is a subjective expectation of a person, or a physical entity, about the behavior of another entity [3]. This behavior could translate into non-technical measures put in place

by organizations or legislators in order to install users' trust. The organizations using AI could put in place clear and formal guidance for AI governance, AI-related risk management systems and ethical principles to adhere to. Legislators could provide a legal framework that will guarantee AI users' protection. Certification schemes can be introduced so that organizations using AI are able to demonstrate compliance and assure their clients of a certain level of quality [4] [5]. For the purpose of this document, these aspects are reflected at the **organizational layer** and are described in the Section 5.

3. Infrastructure layer

Different AI systems rely on the **heterogeneous resources** that require on-demand provisioning. These resources include memory and storage, compute logic, and networking [6].

This section explores the infrastructure-related aspects of AI trustworthiness. It starts with the challenges related to some applications, such as autonomous vehicles, where fast data-processing and decision-making are critical to guarantee the safety of users. On the one hand new computing paradigms including cloud and edge computing are introduced. On the other hand, progress on making hardware that would be optimized for machine learning (ML) operations is described, which is important for business adoption but also may be safety-critical in some situations.

Then, the section introduces the topic of infrastructure security. It discusses the issues of hardware faults and hardware security issues and proposes some mitigation measures.

3.1. Infrastructure optimization

Infrastructure optimization allows for faster learning process and model application. In some applications, like in healthcare, accelerating learning could help save lives, for example with new treatment proposals for an illness. In the others, like in autonomous cars, it is crucial to be able to make fast decisions, by applying the model to real-time data, to avoid accidents on the road.

3.1.1. New computing paradigms

Most modern AI systems require heterogeneous infrastructure resources that are provided on demand, and are coordinated and managed in an automated manner [7]. In this respect, **cloud computing** *"enabling network access to a scalable and elastic pool of physical or virtual resources with self-service provisioning and administration on demand"* [8] has proven efficient. Indeed, it is a paradigm that allows for the storage and processing of large data sets, which is essential for some parts of the AI life cycle [7]. This way, the training of ML models on large volumes of data happens in an accelerated way and allows for optimized decision making in some domains.

However, for some applications this is not sufficient. For example, in the case of autonomous vehicles or robotic devices, decisions need to be taken in real time to guarantee the safety of the users and the environment in which they are operating. Sending data to the cloud could raise privacy concerns and bring significant latency to decision-making [9]. In this context, the **edge computing** paradigm is introduced. Edge computing entails processing and storage taking place at or near the boundary between pertinent digital and physical entities [10], which can be AI systems and sensors collecting data, respectively.

However, the devices involved in computing at the edge would typically have restrained processing, storage and network capacity. Thus, building a robust AI system relying solely on edge computing is not plausible. It is commonly the case that cloud and edge computing are combined in order to take advantage of both paradigms. Multiple AI system designs exist that combine edge and cloud, one of which showing the training of a ML model in the cloud and its usage at the edge is illustrated in Figure 1 [11].

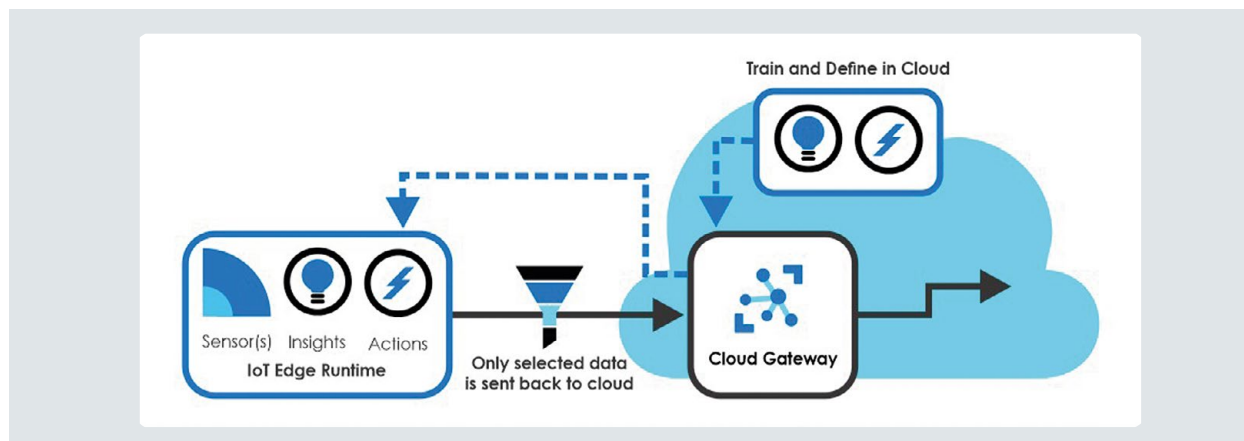


Figure 1: ML system design combining Edge and Cloud computing [11]

More information on edge and cloud computing could be found in ISO/IEC TR 23188:2020 *Information technology – Cloud computing – Edge computing landscape*¹²⁴, while ISO/IEC 22989 *Artificial intelligence – Concepts and terminology*¹²⁵ will explain how these are used in AI. In addition, there exist a number of standards that provide guidance on security measures in cloud computing, among which ISO/IEC 27017:2015 / ITU-T X.1631 *Information technology – Security techniques – Code of practice for information security controls based on ISO/IEC 27002 for cloud services*¹²⁶ that provides guidelines supporting the implementation of information security controls for cloud service customers and cloud service providers.

To support the uptake of AI in Europe, the European Commission proposed to invest more than €4 billion in high-performance and quantum computing, including edge computing and AI, data and cloud infrastructure [12].

3.1.2. Hardware optimization

AI systems rely on heterogeneous resources including memory and storage, compute logic, and networking [6]. All of these elements are evolving to meet the increasing needs of specialized AI applications.

AI systems have an increased need in **data storage** when they refine their algorithms during the training phase. Thus, **new forms of non-volatile memory** are being provided that combine the advantages of traditional memory, such as dynamic random access memory (DRAM), and traditional storage, such as NAND flash. These include higher density, better performance and better power consumption. To handle vast amounts of stored data, AI applications require high-bandwidth memory. Currently, memory is typically optimized for central processing units (CPUs) but new architectures for specialized memory are being proposed, such as **high-bandwidth memory (HBM)** and **on-chip memory** [6].

Moreover, in order to reduce energy consumption while preserving throughput and model accuracy, new architectures with **optimized memory hierarchies**, such as spatial array, are being proposed. Such architectures allow to reduce data movements, which is the main source of energy consumption, and develop data flows with increased data reuse at the low-cost levels of the memory hierarchy [9].

¹²⁴ <https://www.iso.org/standard/74846.html>

¹²⁵ <https://www.iso.org/standard/74296.html>

¹²⁶ <https://www.iso.org/standard/43757.html>

Different types of AI systems have different requirements for **computational resources**. CPUs are not sufficient for some compute-intensive workloads (for example, some neural network topologies), bringing the need for **hardware accelerators** such as **graphic-processing units** (GPUs), **field programmable gate arrays** (FPGAs), **application-specific integrated circuits** (ASICs), etc. Typically, heterogeneous computing solutions are used by AI systems, and the optimal hardware architecture depends on their compute requirements [6].

AI systems may need many servers during the training phase. If the network connecting servers is slow, the training may not succeed. One way of dealing with this issue involves **data-center hardware**. An alternative solution is **programmable switches** that can route data in different directions. This could accelerate the (re)-synchronization between servers during training with updated model parameters and significantly increase the training speed [6].

3.2. Infrastructure security

“Interest in hardware security is increasing because at the end of the day, hardware is the root of trust” [13]. **Faults in hardware** can result in important errors in ML development and applications, causing low accuracy of AI applications. Hardware faults can be accidental, a result of defective components, or a result of a malicious act [2]. Beyond hardware faults, **security attacks** on AI can apply to both hardware and software and target the integrity of data, data privacy, ML model confidentiality and robustness [14]. Mitigation measures should be taken to minimize or avoid hardware errors and attacks.

3.2.1. Hardware faults

Problem description

As with any other software, AI-based systems are prone to common random **hardware faults**. These faults are typically temporary state changes of memory cells or logic components that can be caused by an external source, such as high-energy radiation, electromagnetic noise, etc., internal cross talk between conductor paths or component parts, or malicious injection of perturbations such as clock glitches. An example of a hardware fault are bit-flips occurring in RAM or CPU registers and leading to data corruption [2].

Different fault models would account for the possible sources of error and subsequently for effective fault tolerant strategies. In some system architectures, for example such as using GPUs, the diagnostics could be difficult [2].

Mitigation measures

Robust and fault tolerant systems can be built using different methods based on the architecture and detailed design of the hardware but also the whole development process. One of these methods is to exploit **hardware, information or time redundancy** to mask or otherwise work around failures and keep the desired level of functionality. For example, hardware redundancy consists in incorporating extra hardware into the design to either detect or override the effect of a failed component [2]. A comprehensive description of common methods and processes for implementing fail-safe hardware and a description of certifiable functional safety levels can be found in the IEC 61508 series of standards on *Functional safety*¹²⁷. As a complement, ISO/IEC TR 24028 *Artificial intelligence — Overview of trustworthiness in artificial intelligence*¹²⁸ provides a highlight of AI-related hardware faults and mitigation measures.

¹²⁷ <https://www.iec.ch/functionalsafety/>

¹²⁸ <https://www.iso.org/standard/77608.html>

3.2.2. Hardware security and verification

Problem description

In Section 3.1, it was mentioned that some types of **hardware and compute architecture** allow for better data and privacy protection, since they limit data movement and thus make it less prone to external attack [9]. Various specialized hardware that enables faster execution of AI systems was also introduced in Section 3.1. While AI systems are often run on **specialized hardware**, so-called trusted execution environments (or secure enclaves) are currently available only on non-specialized hardware [4]. There is a need to find ways to ensure that AI infrastructure responds to the requirements of confidentiality, integrity and availability.

Mitigation measures

Secure hardware components

In some cases, hardware components, for example accelerators, can be **paravirtualized or emulated**, or benefit from using specifically assigned devices in cloud-based applications [2]. This helps to increase the **segregation of memory** used by AI applications and thus increase the level of security.

Providing a **secure enclave**, a trusted execution environment combining hardware and software features, would be an additional step to reduce the risk of unauthorized access to sensitive data or interference with an AI system. Since a secure enclave is a setting that provides an isolated execution environment, it gives additional guarantees regarding the applications running inside the enclave even if malicious actors manage to access the system outside it. However, the execution of ML on specialized hardware with secure enclaves is not common and comes with additional costs and performance overhead [4].

Research for secure AI hardware is ongoing. Since AI systems are often run on specialized hardware, it could be an incentive to develop secure hardware for specific applications rather than develop generic secure hardware. An interesting feature for such a secure AI hardware would be to **prevent copying** a trained AI model off a chip without the original copy first being deleted. This feature would allow to limit the number of AI systems that could cause damage if used on a large scale by malicious actors. **Tamper-proof hardware**, which already exists in the context of cryptography, could be useful in the AI context, preventing outsiders from discovering the inner workings of an AI system in case the hardware is stolen, compromised, etc. Moreover, developing a **reference model for secure AI-specific hardware**, including hardware-level access restrictions and audits, could be considered [15].

Hardware verification

Formal verification of devices on which an AI application is run could help to ensure that privacy and security requirements are met. However, given the complexity of AI systems, it may be difficult or impossible to provide an end-to-end verification framework. It appears more feasible to increase the security of the system's components. For example, **hardware input/output (IO) memory management capabilities** could be used to limit a given device to the processes it is supposed to run. In addition, formal methods, which are widely used in the hardware industry, could be adopted [2] [15].

4. Software and System layer

Trustworthiness of an AI application depends strongly on the trustworthiness of the underlying AI system and its hardware and software parts. While hardware aspects were discussed in Section 3, this section addresses the issues related to software components and software-based systems.

Quality of a system and software, as defined in ISO/IEC 25010:2011 *Systems and software engineering — Systems and software Quality Requirements and Evaluation (SQuaRE) — System and software quality models*¹²⁹, contributes considerably to their trustworthiness, since it allows to define and measure the degree to which the product satisfies stated and implied needs, with multiple trustworthiness characteristics (such as security, reliability, etc.) being typically part of those needs [2]. However, AI systems and software cannot be fully evaluated using only the traditional systems and software quality characteristics: new requirements have been identified for AI, such as fairness, explainability, etc. [16].

Since traditional systems and software quality characteristics are still applicable to AI, this section first introduces the standard system and software quality model while specifying which characteristics may need to be adapted to define AI quality. After that, some major new quality characteristics of AI systems are discussed.

4.1. Quality of AI system

A **quality model** for classical software is introduced in the ISO/IEC 25000 series of standards for *System and Software Quality Requirements and Evaluation (SQuaRE)*¹³⁰. It specifies the quality of computer systems/software products, quality in use and data quality, through quality characteristics and sub-characteristics. ISO/IEC 25059 on *Quality Model for AI-based systems*¹³¹ is a new part of the SQuaRE series under development that will complement the traditional software quality model with new characteristics specific to AI systems. This section outlines various quality characteristics both introduced in the SQuaRE series and specific to AI systems.

¹²⁹ <https://www.iso.org/standard/35733.html>

¹³⁰ <https://iso25000.com/>

¹³¹ <https://www.iso.org/standard/80655.html>

4.1.1. Product quality

Product (system or software) quality in ISO/IEC 25010:2011 is defined through 8 quality characteristics, as summarized in Figure 2. Each characteristic is described further in this section while highlighting the challenges related to AI.

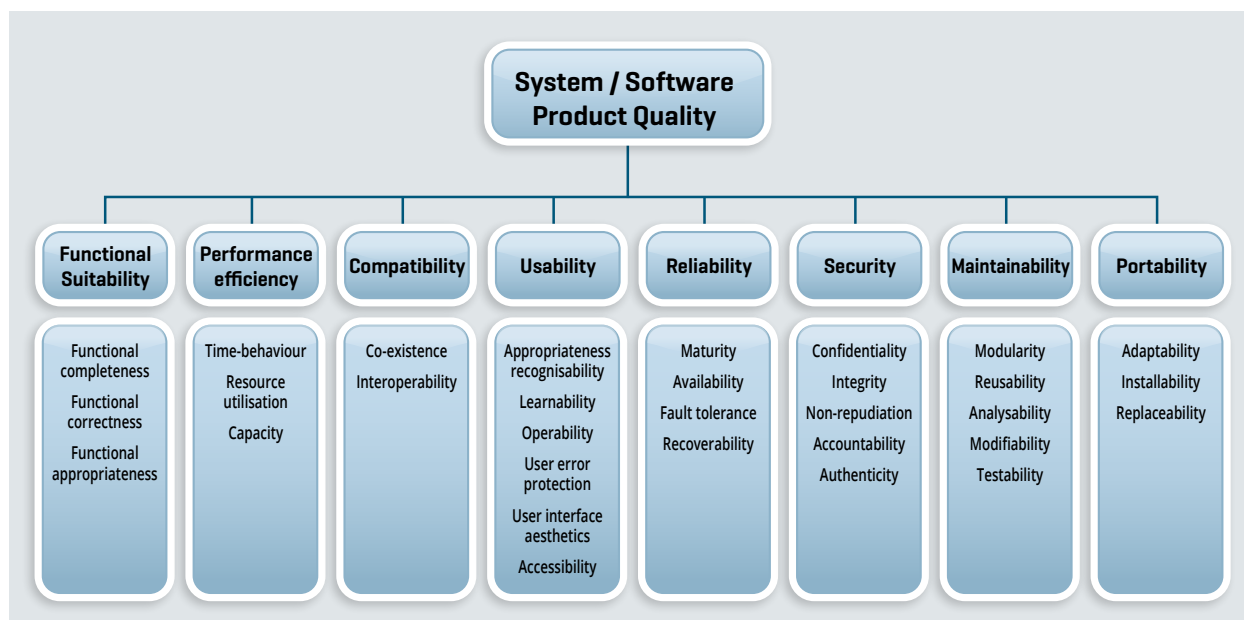


Figure 2: Product quality characteristics according to ISO/IEC 25010:2011 [1]

Functional suitability is about meeting the stated implied needs. It is composed of functional completeness, correctness and appropriateness to define the coverage of the specified tasks and objectives, the correctness of the provided results and the facility to accomplish the specified tasks and objectives [1]. With respect to AI, the functional completeness may sometimes be fully reached only during the operation phase, like in the case of personal assistants that learn through experience. Similarly, the correctness may change over time due to the evolution of the self-learning AI system. In general, the ability to learn can have an important impact on the functional suitability of an AI system.

Performance efficiency concerns the performance of the system with respect to the usage of resources under stated conditions. It involves time required by the system to perform its functions, usage of resources (hardware, software, etc.), capacity of a system to reach the maximum limits of its parameters (such as number of concurrent users, throughput of transactions, database size, etc.) [1]. Performance efficiency could be critical for some types of AI applications, as was discussed in Section 3, along with resource-related challenges and mitigation measures.

Compatibility refers to the ability of the product/system/component to exchange information with other products/systems/components and use that information, and/or share the same hardware/software while performing its functions. It thus relates to the interoperability and co-existence capabilities of the product/system/component [1]. Given the complexity of some AI systems and potential provenance of components from different organizations, compatibility could be a particular challenge in AI-based products as compared to traditional ones. High reliance of some AI systems on data and the provenance of data from multiple sources could be an additional challenge to interoperability.

Usability reflects the degree to which a product or system can be used by specified users to achieve specified goals with effectiveness, efficiency and satisfaction in a specified context of use. It is characterized by the ability of the user to recognize the system's appropriateness, the level of ease with which one can learn to use it, the

presence of attributes allowing to easily operate and control the system, the protection of users against making errors, the aesthetics of the user interface, and the system's accessibility to people with the widest range of characteristics and capabilities [1]. With respect to different levels of automation, human oversight and control have become the important characteristics of the usability of AI systems, also related to security, safety and accountability. With the complexity and lack of transparency of some AI systems - the so-called "black-box" systems - explainability is an important quality characteristic supporting the usability of AI. When it comes to embodied AI systems, such as for example care robots, the usability comes to a new level of complexity with new ways of human-machine interactions.

Reliability defines the extent to which the product/system/component performs specified functions under specified conditions for a specified period of time. It is qualified by the maturity of the system/product with respect to the level of reliability it has under normal operation, by its availability (being accessible and operational) when it is required for use, by its ability to maintain the intended level of operation despite the presence of hardware or software faults, and by the ability of the product/system to recover the data and re-establish the desired state of operation in case of a failure or interruption of service [1]. With some AI systems being particularly sensitive to the specified conditions of use as well as being subject to various attacks, the robustness of the system is an important characteristic of reliability. Moreover, users need to be sure that the system operates with respect to ethical principles (fairness, privacy protection, etc.).

Security concerns the information and data protection within the system/product from unauthorized tampering with respect to different types of users (persons or other products/systems) and the level of authorization they have. System and software security is characterized by confidentiality, integrity, non-repudiation, accountability and authenticity [1]. Authenticity of data used in a ML system could be breached and cause significant deviations in the outputs of the system. With the complexity of AI systems and dependencies between its components, accountability could be hard to achieve. New security attacks target integrity and confidentiality of systems/products, with some of them also creating a threat to privacy. Security of AI systems is a challenging topic, with new threats and mitigation measures, and it was given more attention in Section 4.2.1.

Maintainability characterizes the effectiveness and efficiency with which the product/system can be modified to improve, correct or adapt to changes in the environment and in requirements. Maintainability is defined by modularity (composition of a system where the discrete components have minimal impact on each other), reusability, analyzability, modifiability and testability [1]. Maintainability of AI systems is rather complex. Given the complexity of an AI system, a change in one component would often require adapting the whole system. Moreover, some AI systems are supposed to evolve over time while in operation in order to improve their performance, but malicious actors could use this property to degrade the quality of a product/system. Additionally, it is typically the case that an AI system is developed to perform a specific task and it will not be efficient in performing other tasks (if it is trained to recognize dogs and cats, it cannot be expected to perform well in recognizing road signs), although the developments in transfer learning are currently addressing the issue. Given the lack of transparency of some AI systems, it could be difficult to analyze the system and detect the deficient component. In order to test an AI system that will operate in unknown conditions (for example an autonomous vehicle, or an adaptable chatbot), a wide range of tests should be used, including field testing or testing in a simulated environment.

Portability defines the effectiveness and efficiency with which a system/product/component can be transferred from one hardware, software or other operational or usage environment to another. It is characterized by adaptability to a different hardware, software or other operational or usage environment, efficient and effective installability in a specified environment and capability of being replaced by another product for the same purpose in the same environment [1]. In line with limited reusability of an AI system, changing its operational or usage environment can have a significant impact on system performance. For example, if an autonomous vehicle was trained to drive in exclusively dry conditions, adapting it to drive under the rain could require significant effort.

4.1.2. Quality in use

Quality in use in ISO/IEC 25010:2011 refers to the human interaction with a system. It is defined through 5 characteristics, as summarized in Figure 3, which are further described in this section along with AI-specific challenges.

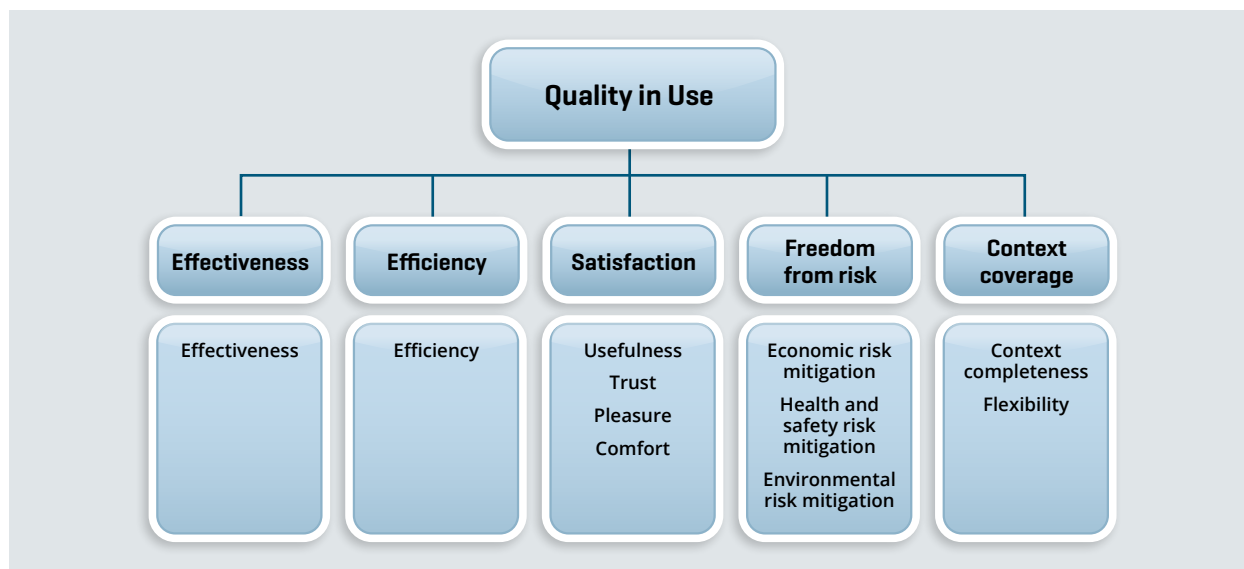


Figure 3: Quality in use characteristics according to ISO/IEC 25010:2011 [1]

Effectiveness, efficiency and satisfaction are directly related to the usability of the product.

Effectiveness refers to the accuracy and completeness of results which the users achieve [1]. There exist various metrics to measure the accuracy and completeness of results of an AI system/product, depending on the underlying model and algorithms.

Efficiency specifies the resources expended in relation to the accuracy and completeness with which users achieve goals [1]. Multiple efforts are directed towards making AI products and systems as human-friendly as possible (for example, emotion manifestation in caring robots), thus targeting the efficiency of their use.

Satisfaction represents the user's level of satisfaction when a product/system is used in a specified context. Satisfaction is manifested through usefulness, trust, pleasure and comfort. Usefulness measures the user's satisfaction with their perceived achievement of pragmatic goals. Pleasure and comfort are respectively about the level of pleasure and physical comfort achieved in the process of usage of the product/system. Trust here refers to a degree to which a stakeholder has confidence that a product/system will behave as intended [1]. In the context of AI, trust is a much more complex concept that is not limited to the stakeholder satisfaction. Transparency, explainability, respect of ethical considerations, etc. all contribute to the establishment of trust in AI systems and the current document attempts to capture various facets of AI trustworthiness.

Freedom from risk refers to the mitigation of potential risks to economic status, human life, health or the environment. Economic risks are the risks to financial status, efficient operation, commercial property, reputation or other resources in the intended contexts of use. Health and safety risks are the risks to people in the intended contexts of use. Environmental risks are the risks to property or the environment in the intended context of use [1]. In the context of AI, health and safety risks are not limited to physical aspects (injuries, etc.) but also relate to mental health and human well-being [12] [17]. Environmental risks are not limited to the intended context of use, and the environmental impact of the technology through the usage of resources (energy, water, etc.)

throughout the life cycle of the AI system should be considered¹³². Moreover, ethical risks reflecting the respect of ethical principles have been identified as critical for AI systems [17]. With respect to the freedom from risks, transparency and explainability of AI system play an important role in risk mitigation since they allow avoiding accidental misuse of the system/product. Similarly, robustness and resilience of an AI system allow mitigating the risks related to the security attacks (potential unexpected behavior of a system causing physical harm, privacy breach, etc.).

Context coverage reflects the level of effectiveness, efficiency, freedom from risk and satisfaction achieved through the use of the product/system in both specified contexts of use and in contexts beyond those initially explicitly identified. Thus, it is characterized by context completeness for the usage in the specified contexts of use and flexibility for the usage in a context beyond that initially specified in the requirements [1]. AI systems can be very sensitive to the specified context of usage, thus limiting their flexibility. At the same time, it could be difficult if not impossible to specify all the contexts of use of some AI systems (for example, autonomous vehicles), making them flexible by default.

4.2. AI-specific attributes for trustworthy and ethical AI systems

4.2.1. Security

Problem description

AI could play different roles in **cybersecurity**. On the one hand, it could render the detection of, and response to intrusion more efficient. On the other hand, malicious actors could use AI to find new ways to, or lower the cost of the infiltration of a target system. Finally, the AI systems themselves could expose new kinds of cybersecurity weaknesses through their use, in addition to the vulnerabilities of traditional computer systems [15] [18]. This section focuses on the attacks specific to AI systems. Typically, the attacks would target the data that is used by ML systems (in either the training or inference stage), the ML model itself (either in a training or deployment environment), or the underlying infrastructure [5] [19]. Both good quality data and trained ML models could represent an important competitive advantage for the organizations, thus their protection is important.

Common **security attacks** exploiting the vulnerabilities of AI systems are as follows:

- **Adversarial examples**

Adversarial examples are obtained by adding slight perturbations to the original legitimate dataset and are intended to trick a benign ML system into misclassifying the input samples at the inference phase or reducing the confidence level of correct classification. These changes are typically not noticeable to a human eye but could affect greatly the output of a ML model, for example by identifying one traffic sign as another [2] [14] [15] [18] [20].

- **Data poisoning**

Poisoning attacks aim at corrupting the process during which the AI system is created so that the resulting system malfunctions in a way desired by the attacker. Poisoning attacks can target data, algorithms or trained models. Regardless of the attack target, the outcome is a model with a hidden weakness or a backdoor [21] [22].

¹³² <https://www.itu.int/en/ITU-T/focusgroups/ai4ee/Pages/default.aspx>

Data poisoning is the most straightforward and frequent way to poison the model. Data poisoning targets the data used for the training of a ML model. Changing (contaminating) training data by adding carefully crafted erroneous samples eventually leads to training an inaccurate model. This is particularly problematic for systems subject to continual training and retraining based on user input, such as recommendation systems or personalized health care. For example, adding malicious training data could result in a model recommending a significant drug dosage change for patients [2] [14] [15] [18] [21].

- **Model stealing/model extraction**

Model stealing, also called model extraction, consists in the replication of the internal functioning of a model. It is done by interacting with a model, sending it the input queries and using the outputs, as well as other information about the system to train another model. Open AI-as-a-Service (AlaaS) APIs are particularly sensitive to these attacks [2] [14].

One particular model stealing attack on neural networks explores power consumption of the neural networks' computations by running specific computational tasks with known input data in order to reveal the secret weight values that characterize the network [13].

Going one step further, the attackers could use the extracted models to generate adversarial examples for the original model. It was shown that as long as the training data of both models is the same, the adversarial examples generated against one model can also deceive the other. The adversarial attacks using an extracted model are called transferability or black-box attacks, since the exact parameters of the original model are not known [14] [19].

- **Model inversion**

Model inversion consists in interacting with a ML model in a deployment environment in order to recover the private features used, including private training data, thus gaining access to potentially secret or sensitive information [18] [19] [20].

Mitigation measures

Mitigation measures to prevent security attacks could be general-purpose, or addressing specific types of attack. An example of a general-purpose measure is the **isolation of the functional modules** of an AI system and the setting up of access control mechanisms between them, thus limiting the surface of the attack [14]. It could be complemented by **continuous monitoring, automatic detection of remote vulnerability scanning** and estimation of the current risk level [14] [15]. The formal verification of the respect of security protocols in the development process could also increase the protection against certain types of attacks [15]. The enhancement of model robustness by exploiting various mechanisms such as **model verifiability, model explainability** or using **automatic suggestions** is another means of defense [14] [15] [20] [22].

Some specific measures are summarized in Table 1 and further described in the text below.

Security attacks	Defense measures
Adversarial examples	Adversarial training Adversarial detection Input reconstruction Feature squeezing
Data poisoning	Data filtering/sanitization Regression analysis Ensemble analysis
Model inversion	Differential privacy
Model stealing/model extraction	Differential privacy PATE Model watermarking Masking

Table 1: Summary of defense measures against security attacks on AI (adapted from [13] [14] [20])

Adversarial training: generating adversarial examples using known attack methods and using them as a subset of a training data set during the training stage, thus making the model resistant to the potential attack perturbations [14] [20].

Adversarial example detection: identification of adversarial examples before they reach the model used for the inference. Different criteria could be used to detect the adversarial examples, such as detecting differences between input samples and normal reference data, etc. [14].

Input reconstruction: deforming input samples by adding noise, using automatic encoders, etc. in order to protect against adversarial attacks in the model inference stage [14].

Feature squeezing: technique of detecting adversarial inputs by comparing the model's performance on original inputs and on "squeezed" inputs, where the samples with many different feature vectors are combined into a single sample [20].

Data filtering/sanitization: the control of training datasets and implementation of detection and purification of training data. For example, it could identify the poisoned data points based on label characteristics and filter those points out during retraining [14] [20].

Regression analysis: detecting noise and abnormal values in datasets based on statistical methods, for example using the distribution characteristics of data [14].

Ensemble analysis: using multiple independent sub-models trained on different datasets help to increase the ability of a final model to defend against poisoning attacks [14].

Differential privacy: adding noise to data or models while minimizing effects on performance [14] [15].

PATE (Private Aggregation of Teacher Ensembles): segmenting training data into multiple datasets and training of multiple independent models that are then used to jointly train a student model, thus limiting the possibility to reveal the information of a particular dataset [14].

Model watermarking: in neural networks, embedding special neurons into the original model during the training stage that enable to check, by providing a special input sample, whether some other model was obtained based on the original one [14].

Masking: for neural networks, splitting intermediate computations into two randomized shares that change each time the same intermediate computation occurs in the network. The split shares only get combined at the final step before producing a result, thus preventing a malicious actor from analyzing intermediate computations and deducing power consumption patterns [13].

European standards for AI security

Security of AI is an important topic in Europe. Thus, an Industry Specification Group was established by ETSI, entitled **Securing AI**¹³³, to work on dedicated standards. The group currently addresses the following subjects [23]:

- Problem Statement, to define and prioritize potential AI threats;
- Threat Ontology for AI, to define AI threats and how they differ from threats to traditional systems;
- Data supply chain, to focus on data issues and risks in training AI;
- Mitigation strategy, to summarize and analyze existing and potential mitigation measures against threats for AI-based systems;
- Security testing of AI, to identify methods and techniques for security testing of AI-based components;
- The role of hardware in security of AI.

Other groups in ETSI also consider AI and security, but with a focus on specific application domains, such as e-health (EP *eHEALTH*¹³⁴) or network infrastructure (ISG *ZSM*¹³⁵, TC *CYBER*¹³⁶) [23].

4.2.2. Privacy and data protection

Problem description

Privacy protection in AI is important not only to gain users' trust but also to remain compliant with the legislation, such as General Data Protection Regulation. However, with the increasing usage of data, particularly personal data, by AI systems, multiple risks and challenges related to privacy protection arise, such as:

- **Interception** of personal data, when in transit from multiple data sources: moving data from one location to another always increases the risk of it being intercepted by malicious entities [9].
- **Re-identification** of personal data: even if the collected data was de-identified, when crossed with other data, links could be established resulting in re-identification. If one data set with health records was de-identified by removing the name and keeping only information such as birth date, postal code and sex to identify the patient, and another data set, for example residents' registry, would contain name, birth date, postal code and sex, then by crossing the two datasets the health records could be directly attributed to the persons [24].
- **Inference** of personal data: ML model could be trained to infer personal data from other, not necessarily personal, data. For example, family situation may be inferred based on the shopping list, presence of some disease may be inferred through eating or physical exercise history, etc. [24].
- **Profiling, influence and manipulation:** identification of trends in personal data could result in the categorization of persons in specific groups and potentially differential treatment, or influencing a person's behavior (targeted advertising, creating stimuli to elicit some reaction, etc.) [24].
- **Unfair/discriminatory decision making** based on profiling, that could be amplified by the lack of transparency of some AI systems and automation of decision-making process [24].

133 <https://www.etsi.org/committee/1640-sai>

134 <https://www.etsi.org/committee/1396-ehealth>

135 <https://www.etsi.org/committee/1431-zsm>

136 <https://www.etsi.org/committee/cyber>

- **Surveillance capitalism or surveillance state:** with the multitude of data sources and the risks of re-identification and inference, big companies or governments that have access to the data might use it to create market opportunities, anticipate market transformation, govern populations, identify and eliminate in a timely manner threats, deliver valuable social services, etc. [24].

As these examples show, the risk of privacy exposure is present at different stages of an AI system's development and usage: at the data acquisition stage, at the data processing and modeling stage, or when the model is deployed and used for inference (as also discussed in Section 4.2.1 on security attacks) [2] [25] [26].

Mitigation measures

Since the risks of privacy violation occur at different stages in an AI system's life cycle, the mitigation strategy should rely on various techniques that guarantee different levels of privacy protection [25] [26] [27]:

- **De-identification:** data transformation that consists in removing the association between individuals and personal identifiers, thus preventing a personal identity from being revealed.
- **Privacy-preserving data mining/ML:** data processing techniques (mining/ML) that minimize the possession of/access to real personal data without losing the functionality of a system.
- **Synthetic data generation:** creation of a dataset that is similar to the original data, but where some or all of the resulting data elements are generated and do not map to actual individuals.

For de-identification and privacy-preserving data mining, a number of techniques could be used, some of which are summarized in Table 2. Synthetic data could be either randomly generated by respecting a number of statistical properties, or produced by the transformation of a real dataset using a set of de-identification techniques. For further reference, ISO/IEC 20889:2018 *Privacy enhancing data de-identification terminology and classification of techniques*¹³⁷ contains a comprehensive list of de-identification techniques, while [28] and [29] discuss privacy-preserving ML.

137 <https://www.iso.org/standard/69373.html>

Defense level	Techniques	Techniques description
De-identification	Structural	Statistical methods that introduce structural changes to the dataset. An example of structural techniques is sampling, a selection of a representative sub-set of a larger dataset.
	Suppression	Removing selected identifiers from the data. One technique, called masking, consists in removing all direct identifiers.
	Pseudonymization	Replacing identifiers with specifically created indirect identifiers, called pseudonyms. The pseudonyms can be randomly generated or cryptographically derived from the original values.
	Generalization	Reducing the granularity of information contained in a selected identifier or in a set of related identifiers. One generalization technique is rounding and consists in rounding the values up or down to the nearest multiple of rounding base.
	Randomization	Modification of identifiers' values so that their new values differ from their original values in a random way. One possibility is to add random values, so-called random noise, to the values of selected identifiers.
Privacy-preserving data mining/ML	Homomorphic encryption	One of the approaches to secure the computation. It allows to encrypt the data and perform operations on the encrypted data without accessing private information.
	Differential privacy	Differential privacy is one of the data-perturbation-based privacy approaches that adds systematic randomized modification to data or algorithm. It helps to reduce individually recognizable information, while preserving the global statistical distribution of a dataset and minimizing effects on algorithm performance.
	Federated learning	Distributing copies of ML algorithms to the devices where the data is kept, performing the training locally and returning the results of the computation, for example updated model parameters, to a central repository to update the main model/algorithm. Thus, the main model has no access to the private information that is only kept locally.

Table 2: Description of privacy protection techniques in AI [28] [29] [30]

As discussed in Section 4.2.1, some techniques could be used to improve both security and privacy (for example, differential privacy).

To be able to measure the degree of privacy protection, privacy measurement models were introduced. One of the well-known models is **k-anonymity**, along with its variations, **l-diversity** and **t-closeness**. K-anonymity identifies a number of k records within which the identity of each individual is not distinguishable from at least k-1 other individuals [26] [30].

4.2.3. Robustness

Problem description

Robustness could be defined as the ability of a system to maintain its level of performance under any circumstances. In other words, when in production, a robust AI system is supposed to continue to operate as intended by design regardless of the input data (also called production data) it receives or any external interference or harsh environmental conditions [2]. However, the evidence shows that robustness is not easy to achieve: an autonomous car may hit a pedestrian due to poor visibility, a chatbot may display racist behavior by interacting with ill-intentioned users, etc.

As robustness guarantees the proper operation of a system, it is directly connected to the safety of its users/stakeholders in a given context/environment [2].

Robustness comprises various system's abilities related to proper operation of a system [2] [4] [5] [12] [31]:

- **Resilience:** typically is referred to as an ability of a system to resist to external or internal interferences, such as hardware failures, accidents or attacks, thus providing a measurement of an AI system's error tolerance [31]. AI systems should be protected against such interferences since the resulting change of a system's behavior could cause the system to make incorrect decisions or shut down altogether. Resilience can also refer to the ability of the system to recover operational condition quickly following an incident [2].
- **Accuracy:** degree to which a product or system provides the correct results with the needed degree of precision [2] [5]. The consequences of low accuracy could be a misclassification of a medical image, incorrect prediction of car trajectory, wrong recommendation of a next product of interest. Depending on the criticality of an AI application, high levels of accuracy are more or less crucial. When it is impossible to achieve high accuracy, it is important to know the likelihood of errors [5].
- **Reliability:** the ability of a system or an entity within that system to perform its required functions under stated conditions for a specific period of time. In other words, a reliable AI system produces the same outputs for the same inputs consistently [2] [5]. It is similar to a concept of predictability, meaning the system will work as intended when used under specified conditions [31]. Reliability is an important feature when it comes to the definition of a legal framework for liability [32].
- **Reproducibility:** in a narrow sense it is about exhibiting the same behavior given the same initial conditions (data, context, etc.), that is, repeatability [4] [5]. More generally, reproducibility refers to a similar or better performance when an AI system is implemented and/or used in a different context [4].

The main challenges in achieving robustness of AI systems as compared to traditional systems is their less deterministic nature (an AI system may produce different outcomes in the same situation because it has learned from the experience) and the need to deal with uncertainty (after deployment, a ML system typically operates on new data that it has not seen before).

Mitigation measures

Various techniques exist that allow to achieve and measure the robustness of AI systems. A combination of multiple approaches is typically required to achieve the desired level of robustness.

As discussed in Section 3.2.1, one way to improve **resilience** of AI systems is through component redundancy [31]. Mitigation measures against security attacks, also contributing to resilience, are discussed in Section 4.2.1.

Accuracy could be achieved through a well-formed development and evaluation process [5]. For example, one well-established practice to measure performance is to split the available data into training and test sets [33]. Depending on the type of problem that is being solved using AI systems, different metrics could be used to evaluate the performance of the system to ensure good results on both training data and the new input data (R-squared, F1 score, etc.) [34].

Reproducibility is a key to enable the verification of claims about an AI system's properties, thus encouraging the reproducibility of reported results could improve the robustness of the system. One way to recognize reproducibility is through the artifact evaluation badges of Association for Computing Machinery (ACM)¹³⁸, which is a system of badges that measures the reproducibility based on evaluated information, available information and reported results. Another effort in this area is the Reproducibility Challenge¹³⁹ [4].

In order to make a **reliable** system and guarantee the correct system's behavior against specific requirements, various **validation and verification methods** could be used. According to definitions in ISO/IEC/IEEE 24765:2017 *Systems and software engineering — Vocabulary*¹⁴⁰, validation would guarantee the right system was built, while verification would guarantee that the system was built correctly. The following validation and verification methods could be used for AI systems:

- **Formal methods:** provide guarantees regarding the functional behavior of a system using mathematical methods. They are often mandatory in safety-critical applications to complement functional testing. An example of a formal method is an uncertainty metric that allows to check if a generalization of neural network does not introduce unstable behavior. However, some sources consider formal methods to be still in their infancy [2] [4].
- **Empirical testing:** is an alternative to formal verification, but cannot fully guarantee its claims. Empirical testing includes benchmarking (measuring performance of competitive systems on a specifically designed datasets), expert panels (experts reviewing the results of the AI systems), metamorphic testing (running multiple iterations of tests and comparing test results) [2] [4].
- **Practical verification:** using scientific protocols to characterize data and performance of AI systems. Training data can be checked for representativeness, absence of unwanted bias, etc., while performance can be evaluated by measuring generalization, heterogeneity across population subsets, etc. [4].
- **Field trials:** a way to test and improve the quality of the deployed system by testing its performance, efficiency, durability or acceptability to human users in actual operating conditions [2].
- **Testing in simulated environment:** when an AI system is intended to perform physical actions in the environment, it is crucial to test its performance in a variety of situations. Physical tests under extreme conditions in the controlled environment allow to evaluate systems' performance and determine their limit operating conditions. Since it may be difficult to reproduce all possible environments in physical world, complementary tests/simulations could be run in virtual environments [2].

For further reference, ISO/IEC TR 24029-1 *Artificial Intelligence (AI) – Assessment of the robustness of neural networks – Part 1: Overview*¹⁴¹ will provide an overview of techniques that guarantee robustness of the neural networks, while another technical specification under development ISO/IEC TS 4213 *Artificial Intelligence – Assessment of machine learning classification performance*¹⁴² will provide a benchmark for the assessment of classification performance of ML.

4.2.4. Safety

Problem description

The concept of safety is related to the use of product/system and associated risks to physical or mental health [32]. One of the major safety risks is the risk of accidents in ML systems, where accidents can be defined as unintended and harmful behavior that may emerge from poor design or implementation of the real-world AI systems. There has been a lot of research on accidents, particularly on those involving embedded reinforcement learning-based systems. This topic is tightly linked to robustness [35].

138 <https://www.acm.org/publications/policies/artifact-review-and-badging-current>

139 <https://reproducibility-challenge.github.io/neurips2019/>

140 <https://www.iso.org/standard/71952.html>

141 <https://www.iso.org/standard/77609.html>

142 <https://www.iso.org/standard/79799.html>

In its report on the safety and liability implications of AI and related technologies, the European Commission identifies various challenges to the European safety framework specific to AI [32]:

- **Connectivity:** connectivity may directly compromise the safety of the product, since it can be hacked, leading to security threats and affecting the safety of users. Moreover, the risk of loss of connectivity of emerging digital technologies may also entail risks related to safety (for example, an autonomous car that would not be able to detect an obstacle and alter its trajectory).
- **Autonomy and evolution throughout the life cycle:** there may be situations where the intended use initially foreseen by the manufacturer is modified due to the evolution and autonomous behavior of the AI system. Thus, new safety risks may be introduced.
- **Opacity:** depending on the methodological approach, AI-based products and systems can be characterized by various degrees of opacity. Thus, human users may be not able to understand and/or foresee the decisions taken by the system, which could lead to unexpected use or misuse of a system resulting in dangerous situations.
- **Data dependency:** if the quality of input data that an AI system receives while in use is lower than expected, this could lead to poor performance and potential safety issues. It is important that AI system producers anticipate the impact of data quality on the accuracy of the system and respective safety functions during the design and testing phases.
- **Mental health:** safety is typically considered in the context of physical harm, but some “behavior” of autonomous AI systems could also trouble the mental health of its users (for example, causing excessive stress or discomfort).
- **Complexity of the system and respective value chain:** AI systems are typically composed of various components, often produced by different providers. The interaction between the components within one system, or between the systems, could be an additional risk factor and should be considered during the risk assessment.
- **Presence of software:** software is an essential component of an AI system. Typically it is required that software faults do not lead hazardous situations, same as software updates. However, software updates may significantly modify the system’s behavior, and thus should result in new risk assessment.

Mitigation measures

Human oversight is considered as an essential measure towards the safety of AI products and systems and should be present from the design phase and throughout the life cycle of the product/system [32].

Given the identified challenges, some of the measures towards AI safety are addressed in sections dedicated to security (Section 4.2.1), robustness (Section 4.2.3), data dependency (Section 4.2.5), transparency and explainability (Section 4.2.7).

It is required by legislation that AI systems integrate mechanisms to ensure they are verifiably safe at every step [32]. For traditional safety-critical systems there exist already **frameworks and tools to assess compliance with safety requirements**. For example, compliance of audit trails can be assessed using Goal Structuring Notation (GSN) or Claims-Arguments-Evidence (CAE) frameworks, while Assurance and Safety Case Environment (ACSE) helps both the auditor and the auditee manage compliance claims and corresponding evidence [4].

Safety requirements for specific applications can be defined in **standards**. For example, IEC 61508 is a basic functional safety series of standards required by many industries [4]. A new technical report ISO/IEC TR 5469 *Artificial Intelligence – Functional safety and AI systems*¹⁴³ is being produced by ISO/IEC JTC 1/SC 42 *Artificial Intelligence*¹⁴⁴ to cover the specificities of functional safety requirements for AI systems.

¹⁴³ <https://www.iso.org/standard/81283.html>

¹⁴⁴ <https://www.iso.org/committee/6794475.html>

Risk management is a useful mechanism to minimize potential harm. With generic Risk Management guidelines, introduced in ISO/IEC 31000:2018 *Risk management*¹⁴⁵ and AI-specific practices described in ISO/IEC 23894 *Artificial Intelligence – Risk Management*¹⁴⁶, the combination of two standards could help organizations to mitigate the risks arising from their use of AI, thus contributing to the development of safe systems and the establishment of trust into the organization.

4.2.5. Data dependency

Problem description

Data is the fuel of AI systems, particularly those based on ML. However, a number of challenges arise around the data for AI systems:

- **Data collection and provenance:** data is often possessed by big market players, collecting it via online platforms. On the one hand the practices of collecting the data need to respect the existing regulation (informed consent, sharing permission, permitted usage, etc.), but even when implemented properly the exact volumes, processing purpose, etc. of collected data could remain vague for the data holders. On the other hand, having the data in the hands of big players creates market imbalances and may slow down research and innovation. Moreover, data provenance is of high importance: on the one hand to guarantee the quality of collected data, on the other hand to be sure that it was not altered [36] [37].
- **Data sharing and processing:** when it comes to data sharing and processing, there is a need for a trusted infrastructure with the capacity to process high-quality data. The European market often experiences problems with multi-cloud interoperability, in particular data portability [36].
- **Data interoperability:** the combination of data from different sources is typically a challenge given the lack of compatible formats and protocols for the gathering and processing of multi-source data in a coherent and interoperable manner [36].
- **Data quality:** good quality AI systems can only be achieved with good quality data. When the data is collected it may contain inaccuracies, errors, mistakes, biases, etc. This has to be addressed prior to training an AI system [5]. Data quality for traditional software can be described by 15 characteristics, as presented in ISO/IEC 25012:2008 *Software product Quality Requirements and Evaluation (SQuaRE) — Data quality model*¹⁴⁷, such as accuracy, completeness, consistency, credibility, currentness, accessibility, compliance, efficiency, precision, traceability, understandability, availability, portability and recoverability [38]. Each of these characteristics is relevant for the data used by AI systems. For example, having accurate, complete, consistent and credible data is crucial for AI systems, and these characteristics typically need to be verified during the design and development of AI systems, at a data pre-processing stage. And the currentness of data in AI systems is directly linked to the validity of predictions made by the systems, especially those making real-time decisions (such as autonomous vehicles).
- **Data security and privacy:** data needs to be protected against unauthorized access and security attacks when it is being exchanged, stored or processed [36].

145 <https://www.iso.org/standard/65694.html>

146 <https://www.iso.org/standard/77304.html>

147 <https://www.iso.org/standard/35736.html>

Mitigation measures

Measures for data security and privacy protections are discussed in Section 4.2.1 and Section 4.2.2 respectively.

Measures towards **interoperability** are being taken at the European level. For example, the European Interoperability Framework¹⁴⁸ provides recommendation to improve interoperability of public digital services. Moreover, the European Commission intends to fund EU-wide common interoperable data spaces and the interconnection of cloud infrastructures that aim at removing technical and legal barriers to data sharing across organizations [36].

A **verifiable data audit** could establish confidence in logs of data interactions and usage [4]. In addition, ITU-T Y.3602 *Big data – Functional requirements for data provenance*¹⁴⁹ describes a **model for big data provenance** as well as operations to consider in this context (how to provide transparency and ensure reliability of collected data). ITU-T Y.3601 *Big data – Framework and requirements for data exchange*¹⁵⁰ provides a **framework for data exchange** in a big data ecosystem covering multiple processes, data types and formats. Finally, standards from ISO/IEC JTC 1/SC 38 *Cloud Computing and Distributed Platforms*¹⁵¹ could help to formalize and facilitate the **data sharing between cloud services** or across the distributed platforms:

- ISO/IEC 19944-1:2020 *Cloud computing and distributed platforms – Data flow, data categories and data use – Part 1: Fundamentals*¹⁵²,
- ISO/IEC TR 23186:2018 *Cloud computing – Framework of trust for processing of multi-sourced data*¹⁵³,
- ISO/IEC 23751 *Cloud computing and distributed platforms – Data sharing agreement (DSA) framework*¹⁵⁴ (in development).

In order to achieve the desired data quality, a number of techniques can be applied, such as **data cleaning** (removing noise, handling the missing values, detecting outliers, etc.), **data transformation** (normalization, numerical transformation, etc.), **data reduction** (sampling, etc.), etc. [39]. For a brief introduction to the topic, ISO/IEC 23053 *Framework for Artificial Intelligence (AI) Systems Using Machine Learning (ML)*¹⁵⁵, currently in preparation, will describe ML process including steps and techniques for data preparation. For more details, a series of standards ISO/IEC 5259 *Data quality for analytics and ML*¹⁵⁶, is being developed.

4.2.6. Bias and fairness

Problem description

AI systems are supposed to respect human rights, such as right to equality, right to non-discrimination, social and cultural rights, etc. [40]. Respecting these rights is often associated with a principle of fair treatment, avoiding stereotypes based on some attributes such as gender, age, ethnicity, etc. [31].

Bias, in its turn, can be defined as disparate judgement about some things, people or groups as compared to others [2] [41]. Although bias could have a positive impact and support **fair treatment** (for example, training and testing an autonomous car in snow conditions would improve its reliability and safety for Nordic countries), it is typically the **unfair bias** that results in violation of human rights that preoccupies stakeholders. This section outlines the sources and types of unfair bias and then describes mitigation measures towards unbiased and fair AI.

148 https://ec.europa.eu/isa2/eif_en

149 <https://www.itu.int/rec/T-REC-Y.3602-201812-I/en>

150 <https://www.itu.int/rec/T-REC-Y.3601-201805-I/en>

151 <https://www.iso.org/committee/601355.html>

152 <https://www.iso.org/standard/79573.html>

153 <https://www.iso.org/standard/74844.html>

154 <https://www.iso.org/standard/76834.html>

155 <https://www.iso.org/standard/74438.html>

156 <https://www.iso.org/standard/81088.html>

Bias can arise from different sources: data bias, statistical bias, human bias, etc. For example, historical data could not reflect the nowadays situation and be biased against gender or ethnicity, such as loan history data, police criminal records, etc. Some groups could be less represented than others, thus resulting in statistically biased models. AI is developed by humans who could do so only based on their personal experience and knowledge, thus limiting the generalization of the system to other social and cultural populations [2] [17] [40] [41] [42].

Thus, bias can manifest itself in different stages throughout the AI system life cycle:

- **Data collection and pre-processing:** the collected data could be incomplete, inaccurate, non-representative of a larger population, etc. Moreover, bias could be introduced to the data when it is being prepared for AI systems: introducing missing values, correcting supposedly erroneous data, annotating the data, etc. Data annotation for supervised ML could indeed introduce bias since it is subjective and dependent on the social and cultural background of human annotators [31] [41] [42] [43].
- **Modeling:** during the modeling phase, inclusion of some features (such as gender, age, race, etc.) could lead to a biased model. But even if these features are omitted, other features (for example, driving at night for those working night shifts or going to parties), called proxy variables, could still impact the model's behavior and result in incorrect correlations and unfair bias. Moreover, statistical models may amplify the disparities present in training data [41] [42] [43].
- **In operation:** blind application of suggestions made by AI systems or actions of autonomous AI systems can generate new data that could perpetuate or even amplify the original bias [42] [43].

Finally, bias can be classified as [44]:

- **Pre-existing:** reflecting biases in society, social institutions, practices and attitudes.
- **Technical:** resulting from hardware or software limitations.
- **Emerging:** appearing when a system is deployed due to changing societal standards.

Mitigation measures

When dealing with bias, it is important to be able to measure its presence in order to take adequate measure towards removing it. One of the existing metrics is a confusion matrix. It allows, for example, to check the distribution of false positives across sensible classes of data [41].

Various approaches exist to diagnose and mitigate bias, among which:

- **Observational fairness strategy:** analyze the dataset and apply methods to the data aiming the detection of bias against individuals or groups typically based on sensitive attributes, such as race, gender, socioeconomic status [42].
 - **Anti-classification strategy:** declares a ML model fair if it does not depend on sensitive attributes in the data set. This typically requires omitting all the sensitive attributes and correlated proxy attributes [42].
 - **Classification parity:** declares a model fair when its predictive performance is equal across various groups of population that are defined by sensitive attributes. This may result in decreasing accuracy for some groups in order to match that of the others [42].
 - **Calibration strategy:** ensures that the outcomes of an AI system do not depend on the sensitive attributes [42].
- **Causal reasoning algorithms:** tools allowing to audit the causal factors behind AI decisions and justify the outcomes [44].
- **Rapid prototyping, formative evaluation, and field testing** with well-conceived populations of users [44].

On a more practical note, various tech companies have recently released tools allowing to measure bias and to use fair algorithms, such as [42]:

- “AI Fairness 360”, an open-source tool kit by IBM, including algorithms and fairness metrics;
- “Fairness Flow”, a tool released by Facebook that helps to detect bias in ML models;
- “What-if” and “Facets” tools by Google, allowing to visualize the effects of different bias mitigation strategies and metrics and to recommend a fairness metric to use in a decision-making process;
- “Fairlearn”, a Python package released by Microsoft that helps to assess and improve the fairness of ML models.

For further reference, a technical report ISO/IEC TR 24027 *Artificial Intelligence (AI) – Bias in AI systems and AI aided decision making*¹⁵⁷ currently in development in ISO/IEC JTC 1/SC 42 will provide an overview of bias types and sources, as well as techniques to measure and mitigate bias in AI systems and decision-making process. Moreover, IEEE P7003 *Algorithmic Bias Considerations*¹⁵⁸ describes specific methodologies that help to eliminate bias.

4.2.7. Transparency, accountability and explainability

Problem description

AI systems are characterized by various degrees of opaqueness [2] [12]. **Transparency** is thus one of the requirements to understand, predict and trust an AI system [5] [17]. It is also necessary for **accountability**^{159,160}, to be able to track the actions of the whole system to its components and underlying requirements, and then to the individuals or organizations responsible [45].

It could be requested at different levels [2] [45]:

- In the context of organizational practices of data collection, system development, etc.;
- In the context of data collection regarding data provenance;
- In the context of system development with regards to the functionality of system components and model interpretability;
- In the context of usage with regards to the interaction with an AI system, intended usage conditions, data exploitation, expected impact, etc.

The question of **interpretability** and **explainability** arises particularly in the context of system development and the usage of AI systems. On the one hand, it could help developers to build robust, reliable and unbiased systems (by testing them and modifying their behavior if necessary). On the other hand, since the development of AI systems requires technical skills, its inner functioning and outcomes/decisions made are typically not obvious for the layman, bringing up the need to explain them. Finally, transparency and explainability help to prove regulatory compliance of AI systems [17] [31] [45].

There are main four manifestations of transparency [17]:

- **Traceability**: transparency of data usage and system engineering process during implementation, so that the requirements could be traced through the process to the final system.
- **Verifiability**: being able to see how the requirements were implemented helps to verify that the decisions the system makes match those requirements. Verification could be done through formal methods or based on a log of ethical reasoning.

157 <https://www.iso.org/standard/77607.html>

158 <https://standards.ieee.org/project/7003.html>

159 <https://www.iso.org/obp/ui/#iso:std:iso-iec:2382:ed-1:v1:en:term:2126250>

160 <https://www.iso.org/obp/ui/#iso:std:iso:ts:21089:ed-1:v1:en:term:3.3.2>

- **Non-deception:** AI systems are expected to be honest, in their design and communication. Honest design in this context could imply the “physical” appearances of a system corresponding to its functionality (for example, only having “ears” when it is supposed to process sound), the ability to reason about a topic if brought to the conversation by the system itself, etc. Honest communication implies the AI system would not lie or hide information, unless explicitly asked to do so.
- **Intelligibility:** a system should be able to explain its reasoning to a user and do so in a comprehensible way, not with complex technical details. In case the system cannot explain itself, its designers/developers must be able to do it.

Additionally, transparency is required for data usage to guarantee privacy protection to data owners [17]. Section 4.2.2 discusses in more details the challenges related to data protection and privacy.

Mitigation measures

The limitations and assumptions of an AI system, as well as data-related processes are often not properly documented [17]. The comprehensive documentation of AI system implementation processes is an important step towards system transparency (traceability).

The explanations could be process-based (demonstrating the follow-up of the good governance processes and best practices) or outcome-based (clarifying the results of a specific decision) [33] [43]. Also, the explanations could be either destined to technical experts (helping them to improve the system) or to a larger group of stakeholders (having a touch of social science in them) [24].

Technical explanations cover typically [24] [33]:

- **Model explanation:** the overall explanation of an opaque AI system through an interpretable and transparent model that fully captures the logic of the system.
- **Model inspection:** a representation that makes it possible to understand some specific properties of an opaque model or of its predictions.
- **Outcome explanation:** listing the choices of the system that led to a particular outcome/decision.

User-friendly explanations could be based on the following elements [24]:

- **Contrastive explanation:** indicating the inputs that made a difference when taking a particular decision over other possible alternatives.
- **Selective explanation:** provides elements that are considered the most relevant according to human judgement.
- **Social explanation:** an interactive approach wherein the explanation is adopted based on the recipient's beliefs and comprehension capacity.

Some existing technical methods and frameworks for explanations are:

- **Sensitivity analysis:** assumes that the most relevant input features are those to which the output is most sensitive, could provide features that should be changed to modify the decision [33] [46].
- **Layer-wise relevance propagation (LRP):** typically used in image recognition, and explains the decisions by decomposition, that is specifying how much each pixel contributes to a prediction. There exists an LRP toolbox¹⁶¹ that provides an implementation of the method in Python [46].
- **Local Interpretable Model-Agnostic Explanations (LIME):** explaining ML predictions based on isolating parts of a given input that contribute the most to a classification prediction, with an implementation package¹⁶² available [46].

¹⁶¹ https://github.com/Scitator/lrp_toolbox

¹⁶² <https://github.com/marcotcr/lime>

- **Quantitative Input Influence (QII):** a technique to measure influences of individual inputs or groups of correlated inputs on the outputs of the system, by providing causal relations [46].

The DARPA program for Explainable AI (XAI) is a remarkable effort to create transparent and explainable AI. The final delivery of the program is expected to be a “toolkit library consisting of machine learning and human-computer interface software modules that could be used to develop future explainable AI systems” [47] [48].

Another interesting development is IEEE P7001 *Transparency of Autonomous Systems*¹⁶³. This standard will describe measurable and testable levels of transparency, so that autonomous systems can be objectively assessed.

4.2.8. Autonomy and controllability

Problem definition

AI is a technology allowing the development of systems with different **levels of autonomy**. There have been different frameworks to define autonomy of engineered systems, ranging from no system assistance (human does and controls everything), to full system autonomy (system makes decisions without any human input), through intermediary stages with more or less human control [49] [50] [51]. It is thus evident that the concepts of autonomy and **controllability** are tightly connected. With increasing AI system capabilities, its potential autonomy increases and controllability decreases [52]. The **lack of control** over an AI system can result in hazardous situations (for example, an autonomous car without human driver misinterpreting the images of the environment and hitting the pedestrian) and bring additional complexity to the question of **accountability**. The issue of controllability is well-known but yet not well defined [52].

Mitigation measures

Measures to increase transparency could help with controlling the development and operation of AI systems [52].

Moreover, the concept of **human oversight** was introduced in order to avoid the situation where an AI system undermines human autonomy. Three types of human oversight are discussed by the European High-level Expert Group (HLEG) on AI: human in the loop, human on the loop, and human in command. **Human in the loop** refers to the scenario where the human intervenes in every decision cycle of the system. **Human on the loop** refers to the scenario where the human monitors the system’s operation, and human intervention remains possible. **Human in command** refers to a broader context of AI use, where the human can oversee the overall activity of the AI system, estimate its impact (economic, societal, etc.), and decide about when and how to use it in the given context [5]. The level of human oversight should be decided at the organizational level, but the system should be then developed in a way that would allow for the human oversight to take place as intended.

The issue of controllability is currently being discussed in ISO/IEC JTC 1/SC 42 in order to propose a standard framework of addressing it.

¹⁶³ <https://standards.ieee.org/project/7001.html>

5. Organizational layer

As mentioned in Section 2, the users' trust could be achieved by technical and non-technical measures [3] [31]. For example, transparency could be ensured by technical means via explaining the results of the AI systems, as discussed in Section 4.2.7, but these measures should be complemented by active communication and explanation of the decision-making process inside the organization [31]. This section discusses non-technical measures to achieve trustworthiness, which are covered by Responsible AI practices. These measures include an ethical framework, end-to-end governance and regulations [48] [53].

5.1. AI ethical framework

Problem description

The ethics of AI and autonomous systems has recently been at the center of attention and debate. If it has been shown that some universals may exist worldwide for solving ethical dilemmas, quantitative variations remain substantial across countries [54]. In this context, multiple organizations have listed a number of **universal ethical principles** to which organizations that use AI should adhere: OECD¹⁶⁴, European Commission [5], IEEE [17], Future of Life Institute¹⁶⁵, global tech companies, etc. Among the most common ethical principles [55] one can find:

- Privacy and security,
- Fairness and non-discrimination,
- Accountability,
- Transparency, openness and responsible disclosure around AI,
- Robustness, reliability, and safety throughout the life cycle,
- Common good, sustainability, well-being,
- Human oversight, control, auditing,
- Respect the rule of law and human rights.

However, it is only through the implementation of those principles and demonstration of the evidence of such implementation that trust could be gained [4] [31] [55].

Mitigation measures

In addition to technical measures discussed in Section 4.2, some institutional practices can be put in place to improve the functioning of AI systems and achieve compliance with ethical principles, such as **red teaming** (establishing dedicated internal to organization team tasked to find flaws and vulnerabilities) or **bias and safety bounties** (compensation for discovering bias and safety vulnerabilities), etc. [4]. For **guidance on the implementation of ethical principles**, organizations could turn to standards. For example, a technical report ISO/IEC TR 24368 *Artificial intelligence — Overview of ethical and societal concerns*¹⁶⁶ outlines the existing ethical frameworks and good practices of applying ethics into decision making. IEEE P7000 *Draft Model Process for Addressing Ethical Concerns During System Design*¹⁶⁷ will describe a model process for addressing ethical concerns during system design.

¹⁶⁴ <http://www.oecd.org/going-digital/ai/principles/>

¹⁶⁵ <https://futureoflife.org/ai-principles/>

¹⁶⁶ <https://www.iso.org/standard/78507.html>

¹⁶⁷ <https://standards.ieee.org/project/7000.html>

Verification of implementation could be done via an assessment either internally or by a third party. To this end, the High-Level Expert Group (HLEG) on AI put in place by European Commission proposes an assessment list for trustworthy AI for self-assessment that is a series of yes/no questions targeting various aspects of 7 key requirements to trustworthy AI [56]. The group working on Fairness, Accountability, and Transparency in Machine Learning (FAT ML) provides a list of questions to assess the implementation of 5 principles of algorithmic accountability as well as a guidance towards addressing these questions¹⁶⁸. AI Ethics Impact Group suggests an ethical label that, similar to energy efficiency label, allows to measure on a scale from A to G the implementation of 6 culture-independent ethical principles by organizations, taking into account socio-technical nature of AI and diversity of stakeholders [31]. IEEE has been working on an Ethics Certification Program for Autonomous and Intelligent Systems (ECPAIS) that will focus on transparency, accountability and algorithmic bias¹⁶⁹.

Proper governance would help to orchestrate the implementation of ethical principles and its assessment at different levels of organization.

5.2. AI Governance

Problem description

The rapid advancement of technology, as well as the increasing competition between companies and even countries, may hinder responsible and sustainable AI development thus resulting in harmful effects on society and potentially prohibitive legal frameworks. In this context, to avoid unwanted consequences, 2019 has seen a rise of interest in **AI governance** with multiple countries and organizations establishing **policies, strategies** and other initiatives [57].

Mitigation measures

AI Governance is happening at different levels, and both the public and private sectors are concerned. More specifically, three modes of AI governance could be distinguished: governance based on governmental bodies, governance based on technologies and governance based on humanistic values [57]. In the current context, a mix of two or three modes is often observed.

An example of a comprehensive **governmental AI policy is provided by the European Commission** [57], including ethical guidelines, research and investment recommendations, supporting education and digital skills development, and the exploration of a legal framework¹⁷⁰. One also finds the **OECD AI Principles**¹⁷¹ and **Policy Observatory**¹⁷², which show steps towards global AI governance. However, it is still too early to talk about real international governance [57].

Regarding the governance of technology, the companies using AI are putting in place **governance frameworks** covering design, development, evaluation, deployment and monitoring of AI to tackle societal concerns and provide better services to society [48] [53].

Governance based on humanistic values aims at embedding human values into AI by means of ethical frameworks, already discussed in Section 5.1.

Standards could be considered as a means to establish global AI governance [58], covering both technical and non-technical issues. For example, a standard ISO/IEC 38507 *Governance of IT – Governance implications of the*

¹⁶⁸ <https://www.fatml.org/resources/principles-for-accountable-algorithms>

¹⁶⁹ <https://standards.ieee.org/industry-connections/ecpais.html>

¹⁷⁰ <https://ec.europa.eu/digital-single-market/en/artificial-intelligence>

¹⁷¹ <https://www.oecd.org/going-digital/ai/principles/>

¹⁷² <https://oecd.ai/>

use of Artificial Intelligence by organizations¹⁷³ is currently being developed to provide guidance to the governance bodies of the organizations using AI.

For effective implementation of governance, an AI management system (possible solution introduced in Section 5.3) is a key.

5.3. Legal framework and certification schemes

Problem description

Being lawful is one of the principles of trustworthy AI [5] [40]. Indeed, a **legal framework** is needed to establish clear paths for **accountability** and introduce requirements for **human safety and human rights**. The current European legislation seems to only partially cover these questions around the implementation of AI-based products.

Mitigation measures

In this context, the European Commission published a number of documents analyzing the current **EU legal framework** and highlighting potential challenges of its application in the context of AI [12] [32] [59]. The current legal framework under consideration is illustrated in Figure 4.

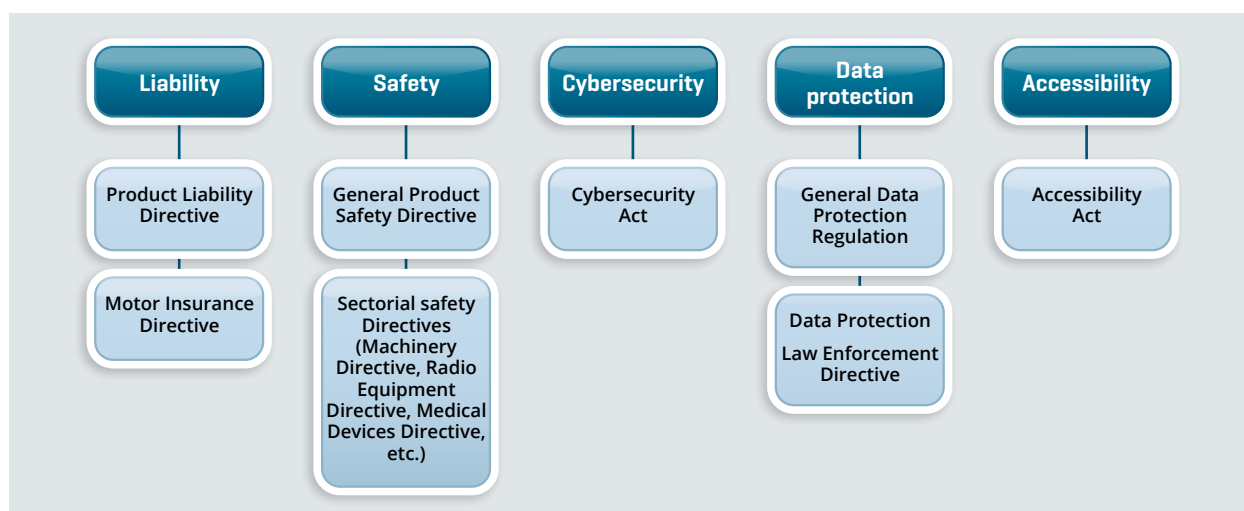


Figure 4: EU legislation relevant to AI [12] [32] [59]

The European Commission launched a public consultation on the topic and the collected results will help to further shape European legislation in the area of AI.

At the same time, **compliance mechanisms** are explored. If for high-risk applications a **mandatory conformity assessment scheme** is imposed, lower-risk applications are offered to apply a **voluntary labelling scheme** [12]. Whatever the risk level of the application, **harmonized standards**, developed at the mandate of the Commission by the European Standards Bodies, could support organizations in their quest for compliance.

Moreover, an international standard ISO/IEC 42001 *Artificial intelligence – Management system*¹⁷⁴ has been initiated to provide guidance for establishing **AI management systems** within organizations. Management system standards are typically subject to a conformity assessment, such as for example a certification. Publication of ISO/IEC 42001 is expected by 2023.

¹⁷³ <https://www.iso.org/standard/56641.html>

¹⁷⁴ <https://www.iso.org/standard/81230.html>

Finally, as mentioned in Section 5.1, a **certification program** from IEEE, ECPAIS, will focus on transparency, accountability and algorithmic bias.

6. Conclusion

Trustworthiness of AI is an important issue that can significantly support the adoption of the technology by a variety of stakeholders. As this Chapter shows, it should be tackled at different levels in organizations and all along the AI-based systems life cycle. However, it is only through the understanding of the whole AI ecosystem, including its history (Chapter 1), techniques (Chapter 2) and applications (Chapter 3) that the trustworthiness can be fully achieved. Covering all these aspects, standards will play an important role in AI adoption. Moreover, standards can provide a basis for assessment of AI trustworthiness that is helpful for users, developing organizations and legislators. Throughout this chapter the relevant standardization activities are highlighted, and for the readers' convenience are summarized in Annex: Standards for Trustworthy AI.

Annex: Standards for Trustworthy AI

Various standards development organizations work on standards with the objective to increase trust in AI. Most of this work was mentioned through this chapter, in order to connect directly each topic with the relevant standards. Table 3 below presents a non-comprehensive summary of these different activities.

Topic	Standard title	Notes
Computing paradigm	ISO/IEC TR 23188:2020 <i>Cloud computing – Edge computing landscape</i> (published)	Introduction to edge computing.
	ISO/IEC 17788: 2014 <i>Cloud computing – Overview and vocabulary</i> (published)	Cloud computing concepts.
	ISO/IEC 22989 <i>Artificial intelligence – Concepts and terminology</i> (in progress)	Introducing concepts of cloud and edge computing in the context of AI, in addition to other AI-related concepts.
Trustworthiness	ISO/IEC TR 24028:2020 <i>Artificial intelligence – Overview of trustworthiness in Artificial Intelligence</i> (published)	Overview of various concepts contributing to the trustworthiness of AI-based systems and products.
Security	ISO/IEC 27017:2015/ ITU-T X.1631 <i>Security techniques – Code of practice for information security controls based on ISO/IEC 27002 for cloud services</i> (published)	Guidelines for implementation of information security for cloud services.
	ITU-T X.1601 (10/2015) <i>Cloud computing security – Overview of cloud computing security – Security framework for cloud computing</i> (published)	Overview of security threats and challenges in the cloud computing environment and possible mitigation measures.
	ITU-T X.1641 (09/2016) <i>Cloud computing security – Overview of cloud computing security – Guidelines for cloud service customer data security</i> (published)	Generic security guidelines for the cloud service customer data in cloud computing throughout the data security life cycle.
	ETSI DGR/SAI-001 <i>AI Threat Ontology</i> (in progress)	Overview of AI threats and how they differ from threats to traditional systems.
	ETSI DGR/SAI-003 <i>Security Testing of AI</i> (in progress)	Overview of methods and techniques for security testing of AI-based components.
	ETSI GR SAI 004 <i>Securing Artificial Intelligence (SAI); Problem Statement</i> (published)	Overview and prioritization of potential AI threats.
	ETSI DGR/SAI-005 <i>SAI Mitigation Strategy report</i> (in progress)	Summary of existing and potential mitigation measures against threats for AI-based systems.
	ETSI DGR/SAI-006 <i>Hardware in SAI</i> (in progress)	Explaining the role of hardware in security of AI.

Privacy	ISO/IEC 20889:2018 <i>Privacy enhancing data de-identification terminology and techniques</i> (published)	Overview of techniques for data de-identification and privacy preservation.
	ISO/IEC 20547-4:2020 <i>Big data reference architecture – Part 4: Security and Privacy</i> (published)	Guidance on security and privacy operations for big data applicable to stakeholders and functional components.
Functional safety	IEC 61508 series on <i>Functional safety of electrical/electronic/programmable electronic safety-related systems</i> (published)	Methods for application, design, deployment and maintenance of automatic protection of safety-related systems.
	ISO/IEC TR 5469 <i>Artificial intelligence – Functional safety and AI systems</i> (in progress)	Impact of AI on safety related functions and related properties, risk factors, methods and processes.
Robustness	ISO/IEC 24029 series on <i>Assessment of the robustness of neural networks</i> (in progress)	Overview of existing techniques to achieve the robustness of neural networks in Part 1, with particular focus on formal methods in Part 2.
	ISO/IEC TS 4213 <i>Artificial Intelligence – Assessment of classification performance for machine learning models</i> (in progress)	Overview of methodologies for measuring classification performance of ML models, systems, and algorithms.
Software quality	ISO/IEC 25010:2011 <i>Systems and software Quality Requirements and Evaluation (SQuaRE) – System and software quality models</i> (published)	Quality characteristics allowing to specify, measure and evaluate software products and computer systems quality.
	ISO/IEC 25059 <i>Systems and software Quality Requirements and Evaluation (SQuaRE) – Quality Model for AI-based systems</i> (in progress)	Extension of ISO/IEC 25010 describing quality characteristics specific to AI-based systems.
	ISO/IEC 29119-11:2020 <i>Software and systems engineering – Software testing – Part 11: Testing of AI-based systems</i> (published)	Overview of available testing techniques for AI-based systems.
Data	ISO/IEC 25012:2008 <i>Software product Quality Requirements and Evaluation (SQuaRE) – Data quality model</i> (published)	Defining quality characteristics of data retained in a structured format within a computer system and used by humans and systems.
	ISO/IEC 23053 <i>Framework for Artificial Intelligence (AI) Systems Using Machine Learning</i> (in progress)	Describing ML ecosystem, including ML pipeline, covering among others data collection and data preparation steps.
	ISO/IEC 5259 series on <i>Data quality for analytics and ML</i> (in progress)	Multi-parted standard describing data quality requirements, metrics, governance and management principles.
	ETSI DGR/SAI-002 <i>Data Supply Chain Report</i> (in progress)	Focus on data issues and risks in training AI.

	ITU-T Y.3601 (05/2018) <i>Big data – framework and requirements for data exchange</i> (published)	A framework for data exchange in a big data ecosystem covering multiple processes, data types and formats.
	ITU-T Y.3602 (12/2018) <i>Big data – Functional requirements for data provenance</i> (published)	A model for big data provenance as well as operations to consider in this context in order to ensure transparency and reliability of collected data.
Bias	ISO/IEC TR 24027 <i>Artificial Intelligence – Bias in AI systems and AI aided decision making</i> (in progress)	Describing types and sources of bias in AI systems, as long as the mitigation techniques.
	IEEE P7003 <i>Algorithmic Bias Considerations</i> (in progress)	Describing methodologies helping to address and eliminate the so-called negative bias in the creation of the algorithms.
Transparency	IEEE P7001 <i>Transparency of Autonomous Systems</i> (in progress)	Describing measurable, testable levels of transparency, so that autonomous systems can be objectively assessed and levels of compliance determined.
Ethical concerns	ISO/IEC TR 24368 <i>Artificial Intelligence – Overview of ethical and societal concerns</i> (in progress)	Existing ethical frameworks and their application to AI.
	IEEE P7000 <i>Draft Model Process for Addressing Ethical Concerns During System Design</i> (in progress)	A set of processes by which organizations can include consideration of human ethical values throughout the stages of concept exploration and development, supporting both management and engineering teams.
Risk management	ISO/IEC 23894 <i>Artificial Intelligence – Risk Management</i> (in progress)	Extension of ISO 31000, Risk management – Guidelines, dealing with specificities of AI and related processes.
Governance	ISO/IEC 38507 <i>Governance of IT – Governance implications of the use of Artificial Intelligence by organizations</i> (in progress)	Explaining the role of AI in the organizations, outlining its advantages and risks, allowing to efficiently evaluate, direct and monitor its introduction and use.

Table 3: Standards for Trustworthy AI

References

- [1] ISO/IEC JTC 1/SC 7, "ISO/IEC 25010:2011, Systems and software engineering – Systems and software Quality Requirements and Evaluation (SQuaRE) – System and software quality models," 2011.
- [2] ISO/IEC JTC 1/SC 42, "ISO/IEC 24028:2020, Information technology – Artificial intelligence – Overview of trustworthiness in artificial intelligence," 2020.
- [3] ITU-T, "Recommendation ITU-T Y.3052 "Overview of trust provisioning in information and communication technology infrastructures and services"," ITU-T, 2017.
- [4] B. Miles and e. al, "Toward Trustworthy AI Development: Mechanisms for Supporting Verifiable Claims," 04 2020. [Online]. Available: <https://arxiv.org/abs/2004.07213>. [Accessed 11 2020].
- [5] High-Level Expert Group on Artificial Intelligence, "Ethics Guidelines for Trustworthy AI," European Commission, 2019.
- [6] G. Batra, Z. Jacobson, S. Madhav, A. Queirolo and N. Santhanam, "Artificial-intelligence hardware: New opportunities for semiconductor companies," McKinsey&Company, 2018.
- [7] IEC, "Artificial intelligence across industries," 2018. [Online]. Available: <https://basecamp.iec.ch/download/iec-white-paper-artificial-intelligence-across-industries-en/>. [Accessed 11 2020].
- [8] ISO/IEC JTC 1/SC 38, "ISO/IEC 17788:2014, Information technology – Cloud computing – Overview and vocabulary," 2014.
- [9] S. Vivienne, Y.-H. Chen, J. Emer, A. Suleiman and Z. Zhang, "Hardware for Machine Learning: Challenges and Opportunities," Oct 2017. [Online]. Available: arxiv.org/abs/1612.07625. [Accessed 11 2020].
- [10] ISO/IEC JTC 1/SC 38, "ISO/IEC TR 23188:2020, Information technology – Cloud computing – Edge computing landscape," 2020.
- [11] Microsoft, "What is AI@Edge?," 10 2019. [Online]. Available: <https://microsoft.github.io/ai-at-edge/docs/aiatedge/>. [Accessed 11 2020].
- [12] European Commission, "White paper on Artificial Intelligence - A European approach to excellence and trust," 2020.
- [13] J. Hsu, "Preventing AI From Divulging Its Own Secrets," IEEE Spectrum, 05 2020.
- [14] Huawei, "AI Security White Paper," HUAWEI TECHNOLOGIES CO., LTD., 2018.
- [15] B. Miles and e. al, "The Malicious Use of Artificial Intelligence: Forecasting, Prevention, and Mitigation," 02 2018. [Online]. Available: <https://arxiv.org/abs/1802.07228>. [Accessed 11 2020].
- [16] H. Kuwajima and F. Ishikawa, "Adapting SQuaRE for Quality Assessment of Artificial Intelligence Systems," 2019. [Online]. Available: <https://arxiv.org/abs/1908.02134v1>. [Accessed 11 2020].
- [17] IEEE, "Ethically aligned design: a vision for prioritizing human well-being with autonomous and intelligent systems (version 2)," 2018.
- [18] B. Buchanan, "A National Security Research Agenda for Cybersecurity and Artificial Intelligence," Center for Security and Emerging Technology, 2020.
- [19] S. Herpig, "Securing Artificial Intelligence. Part 1: The attack surface of machine learning and its implications.," Stiftung Neue Verantwortung, 2019.
- [20] A. Marshall, J. Parikh, E. Kiciman and R. S. S. Kumar, "Threat Modeling AI/ML Systems and Dependencies," November 2019. [Online]. Available: <https://docs.microsoft.com/en-us/security/engineering/threat-modeling-aiml>. [Accessed 11 2020].
- [21] M. Comiter, "Attacking Artificial Intelligence: AI's Security Vulnerability and What Policymakers Can Do About It," Belfer Center for Science and International Affairs, Harvard Kennedy School, 2019.
- [22] National Science and Technology Council, "AI and Cybersecurity: Opportunities and Challenges," 2020.
- [23] ETSI, "Artificial Intelligence and future directions for ETSI," 2020.
- [24] European Parliament Research Service, "Study: The impact of the General Data Protection Regulation (GDPR) on artificial intelligence," European Parliament, 2020.
- [25] S. Esmailzadeh Dilmaghani, M. Brust, G. Danoy, N. Cassagnes, J. Pecero and P. Bouvry, "Privacy and Security of Big Data in AI Systems: A Research and Standards Perspective," in 2019 IEEE International Conference on Big Data (IEEE BigData 2019), 2019.
- [26] ILNAS - University of Luxembourg, "Smart ICT: Gap Analysis Between Scientific Research and Technical Standardization," 2019.
- [27] NISTIR 8053, "De Identification of Personal Information," 2015. [Online]. Available: <http://dx.doi.org/10.6028/NIST.IR.8053>. [Accessed 11 2020].
- [28] G. Kaissis, M. Makowski, D. Rückert and R. Braren, "Secure, privacy-preserving and federated machine learning in medical imaging," Nature Machine Intelligence, vol. 2, pp. 305-311, 2020.

- [29] J. Mancuso, "Privacy-Preserving Machine Learning 2019: A Year in Review," January 2020. [Online]. Available: <https://medium.com/dropoutlabs/privacy-preserving-machine-learning-2019-a-year-in-review-123733e61705>. [Accessed 11 2020].
- [30] ISO/IEC JTC 1/SC 27, "ISO/IEC 20889:2018, Privacy enhancing data de-identification terminology and classification of techniques," 2018.
- [31] AI Ethics Impact Group, "From Principles to Practice: an interdisciplinary framework to operationalise AI ethics," 2020.
- [32] European Commission, "Report from the Commission to the European Parliament, the Council and the European Economic and Social Committee: Report on the safety and liability implications of Artificial Intelligence, the Internet of Things and robotics," 2020.
- [33] R. Hamon, H. Junklewitz and I. Sanchez, "Robustness and Explainability of Artificial Intelligence - From technical to policy solutions, EUR 30040," Publications Office of the European Union, 2020.
- [34] CSSF, "Artificial Intelligence : opportunities, risks and recommendations for the financial sector," 2018.
- [35] D. Amodei, C. Olah, J. Steinhardt, P. Christiano, J. Schulman and D. Mané, "Concrete Problems in AI Safety," 07 2016. [Online]. Available: <https://arxiv.org/abs/1606.06565>. [Accessed 11 2020].
- [36] European Commission, "A European strategy for data," 2020.
- [37] K. Alwani and M. Crawford Urban, "The Digital Age: Exploring the Role of Standards for Data Governance, Artificial Intelligence and Emerging Platforms," CSA Group, 2019.
- [38] ISO25000.com, "ISO/IEC 25012, Software product Quality Requirements and Evaluation (SQuaRE) – Data quality model," [Online]. Available: <https://iso25000.com/index.php/en/iso-25000-standards/iso-25012>. [Accessed 11 2020].
- [39] L. Canchen, "Preprocessing Methods and Pipelines of Data Mining: An Overview," 06 2019. [Online]. Available: <https://arxiv.org/abs/1906.08510>. [Accessed 11 2020].
- [40] OECD, "Artificial Intelligence in Society," OECD Publishing, Paris, 2019.
- [41] Partnership on AI, "Report on Algorithmic Risk Assessment Tools in the U.S. Criminal Justice System," 04 2019. [Online]. Available: <https://www.partnershiponai.org/report-on-machine-learning-in-risk-assessment-tools-in-the-u-s-criminal-justice-system/>. [Accessed 11 2020].
- [42] M. Whittaker, K. Crawford, R. Dobbe, G. Fried, E. Kazianas, V. Mathur, S. Myers West, R. Richardson, J. Schultz and O. Schwartz, "AI Now Report 2018," AI Now Institute, 2018.
- [43] ICO and The Alain Turing Institute, "Explaining decisions made with AI," 2020.
- [44] R. Yampolskiy, "Current State of Knowledge on Failures of AI Enabled Products," Consortium for Safer AI, 2018.
- [45] European Parliament Research Service, "A governance framework for algorithmic accountability and transparency," European Parliament, 2019.
- [46] S. Wojciech, W. Thomas and M. Klaus-Robert, "EXPLAINABLE ARTIFICIAL INTELLIGENCE: UNDERSTANDING, VISUALIZING AND INTERPRETING DEEP LEARNING MODELS," ITU Journal: ICT Discoveries, no. 1, 2017.
- [47] D. Gunning, "Explainable AI Program Update," 11 2017. [Online]. Available: [www.darpa.mil > XAIProgramUpdate](http://www.darpa.mil/XAIProgramUpdate) (<https://www.darpa.mil/program/explainable-artificial-intelligence>). [Accessed 03 2019].
- [48] Accenture, "Responsible AI: a framework for building trust in your solutions," 2018.
- [49] T. Sheridan and W. Verplank, "Human and Computer Control of Undersea Teleoperators," Man-Machine Systems Laboratory, Massachusetts Institute of Technology, Cambridge, Massachusetts, 1978.
- [50] NIST Ad hoc Working Group Autonomy Levels for Unmanned Systems (ALFUS), "Autonomy Levels for Unmanned Systems (ALFUS) Framework. Volume 2: Framework models," NIST, 2007.
- [51] Society of Automotive Engineers (SAE), "Taxonomy and Definitions for Terms Related to Driving Automation Systems for On-Road Motor Vehicles," 2018.
- [52] R. Yampolskiy, "On Controllability of AI," July 2020. [Online]. Available: <https://arxiv.org/abs/2008.04071>. [Accessed 11 2020].
- [53] PwC, "A practical guide to Responsible Artificial Intelligence (AI)," 2019.
- [54] E. Awad, S. Dsouza, I. Rahwan and J.-F. Bonnefon, "Universals and variations in moral decisions made in 42 countries by 70,000 participants," in *Proceedings of the National Academy of Sciences*, 2020.
- [55] T. Hagendorff, "The ethics of AI ethics: an evaluation of guidelines," *Minds and Machines*, vol. 30, pp. 99-120, 2020.
- [56] High-Level Expert Group on Artificial Intelligence, "The Assessment List for Trustworthy Artificial Intelligence (ALTAI)," European Commission, 2020.
- [57] "AI GOVERNANCE IN 2019: a year in review," Shanghai Institute for Science of Science, 2020.

- [58] P. Cihon, "Standards for AI Governance: International Standards to Enable Global Coordination in AI Research and Development," *Future of Humanity Institute, University of Oxford*, 2019.
- [59] Expert Group on Liability and New Technologies, "Liability for Artificial Intelligence and other emerging digital technologies," *European Commission*, 2019.

NOTE: While any hyperlinks included in this chapter were valid at the time of consultation, ILNAS cannot guarantee their long-term validity.

Conclusions and outlook

Having been born more than 60 years ago, AI has now reached a point of maturity where it can benefit society and the economy. As with any technology, it comes with challenges that need to be addressed.

These can be technical or societal and manifest themselves in different ways. The complexity of the AI supply chain and its dependence on data and computational resources is one of them. The fact that there is no one-size-fits-all AI solution and that businesses face the issue of applying the appropriate techniques to solve their business problems is another. This white paper attempts to provide an overview of AI techniques and business problems and to connect them, to show which techniques can be applied to which business problems and application domains.

However, having a solution that works for any given problem may not be sufficient for full market adoption. Additional factors play an important role, such as solution robustness, transparency, or the respect of ethical principles. These are all building blocks of trustworthy AI. Indeed, trust is key to the adoption of AI and should be built into different layers, covering all the components of the AI ecosystem. There are different means of achieving trust in AI that were introduced in a dedicated chapter of this white paper, including technical measures and organizational policies, with technical standardization as a means of recording and communicating good practices in this area.

A great deal of activities in AI standardization currently take place in ISO/IEC JTC 1/SC 42. But, other standards developing organizations and industrial fora and consortia also have their domain and industry-specific standardization activities that can help developers and users of AI systems. It is beneficial for organizations to not only use standards, but to contribute to their development. Throughout this white paper, standardization activities are introduced, with the objective to make national stakeholders aware of current developments in this area and facilitate the identification of projects that could help them in building trusted AI systems. ILNAS, as the national standards body, vividly encourages the involvement of the national market in the technical committees responsible for such developments in order to gain knowledge in the domain and ensure that the national perspective is taken into account in future international standards.

Standardization developments in AI are of strategic importance to ILNAS, which is reflected in the National Standardization Strategy 2020-2030¹⁷⁵ and the related Policy on ICT Technical Standardization 2020-2025¹⁷⁶. Indeed, in addition to this white paper, ILNAS leads different projects covering, inter alia, AI in order to give national stakeholders an opportunity to extend their knowledge of AI technical standardization:

1. National Standards Analysis “Smart Secure ICT”¹⁷⁷: the national standards analysis of the Smart Secure ICT sector is a report published yearly by ILNAS in order to provide a “snapshot” of the Smart Secure ICT standardization landscape, including AI, Blockchain and Distributed Ledger Technologies, Cloud Computing and the Internet of Things (IoT) as well as related Digital Trust standards developments. Through this overview of international standardization activities, national stakeholders can easily identify technical committees developing standards relevant for their businesses and decide whether they would have an interest to participate in the development of these standards.
2. Knowledge transfer: ILNAS, with the support of ANEC GIE, currently holds the presidency of the national mirror committee for ISO/IEC JTC 1/SC 42¹⁷⁸, which gathers all the national stakeholders participating in the development of AI international standards. In parallel, a continuous monitoring of all standardization activities related to AI is carried out (including projects of CEN, CENELEC, ETSI and ITU-T), allowing ILNAS to

¹⁷⁵ <https://portail-qualite.public.lu/content/dam/qualite/publications/normalisation/2020/strategie-normative-luxembourgeoise-2020-2030.pdf>

¹⁷⁶ <https://portail-qualite.public.lu/content/dam/qualite/publications/normalisation/2020/policy-on-ict-technical-standardization-2020-2025.pdf>

¹⁷⁷ <https://portail-qualite.public.lu/dam-assets/publications/normalisation/2020/smart-secure-ans-tic-september-2020.pdf>

¹⁷⁸ <https://portail-qualite.public.lu/fr/normes-normalisation/secteurs/tic/artificial-intelligence.html>

transfer relevant information to the market, notably through the publication of this white paper or through the organization of events dedicated to the AI topic. In this frame, ILNAS also encourages all interested stakeholders in Luxembourg to get involved in the standards development process by becoming national delegates in standardization¹⁷⁹ and directly access the relevant information.

3. Education: ILNAS, in collaboration with the University of Luxembourg and the Luxembourg Lifelong Learning Center, has developed the Master degree MTECH “Technopreneurship: mastering smart ICT, standardization and digital trust for enabling next generation of ICT solutions”¹⁸⁰, which is an innovative part-time program mainly designed for professionals in the Smart ICT driven economy sectors. Thus, digital trust, which is crucial for AI and other Smart ICT, will be at the heart of the program, along with technical standardization that is considered an important tool in building trustworthy products and services.
4. Research: ILNAS also has a strong relationship with the University of Luxembourg and SnT in order to foster the exchanges between technical standardization and research and innovation. Through this partnership, two research programs have been developed with the focus on digital trust for smart ICT¹⁸¹. The first research program (2017-2020) “Technical Standardization for Trusted Use in the Field of Smart ICT”, with its core pillars - Big Data & Analytics (also extended to AI), IoT and Cloud Computing - has already resulted in a number of scientific publications and contributions to technical standards. A new joint research program (2021-2024), “Technical Standardization for Trustworthy ICT, Aerospace, and Construction”, is a logical follow-up of the first one. In line with the National Standardization Strategy 2020-2030, it will address three key sectors for the Luxembourg economy, namely ICT, Aerospace and Construction and will focus on the aspects of reliability, security and privacy protection. In the ICT sector, it will deepen the research work carried out on AI in the first research program, in order to assess, analyze and develop aspects of confidentiality, trust and security of AI systems.

179 <https://portail-qualite.public.lu/fr/normes-normalisation/participer-normalisation/experts-normalisation.html>

180 <https://portail-qualite.public.lu/fr/normes-normalisation/education-recherche/education-normalisation.html>

181 <https://portail-qualite.public.lu/fr/normes-normalisation/education-recherche/normalisation-recherche.html>





ILNAS

Institut Luxembourgeois de la
Normalisation, de l'Accréditation, de la
Sécurité et qualité des produits et services

ANEC

Agence pour la Normalisation
et l'Economie de la Connaissance

Southlane Tower I · 1, avenue du Swing · L-4367 Belvaux · Tel. : (+352) 24 77 43 -70 · Fax : (+352) 24 79 43 -70 · E-mail : info@ilnas.etat.lu

www.portail-qualite.lu