TRUSTWORTHY AI

A STRATEGIC OPPORTUNITY FOR INDUSTRIAL AND DIGITAL SOVEREIGNTY

Julien CHIARONI and Arno PONS

THINK-TANK DIGITAL NEW DEAL

June 2022

DIGITAL NEW DEAL COLLECTION "TRUSTWORHTY DIGITAL"

BY IMPOSING TRUSTWORTHY AI AS A STANDARD VALUE, THE EU CAN SET ITS OWN EXTRATERRITORIAL CONDITIONS TO THE GLOBAL MARKET.

SUMMARY

PREFACE

The Covid crisis has shed a harsh light on the "Deindustrialization of Europe"[1], highlighting the need to revive a strong and sovereign industry. This new industrial sovereignty is partly linked to digital sovereignty which already lays out a path for this transformation (initiated in 2015) and will continue to do so more and more with artificial intelligence: *"Artificial intelligence is an existential issue for our nations, those who do not master AI will be vassalized and will lose their sovereignty"* Bruno Le Maire[2].

To do this, we must focus on the continental level, by strengthening Europe's position as a pole of excellence in the field of AI, and as a beacon for the world in terms of the values that AI must convey:

For AI to be a guarantor and vector of these values, we must demonstrate high ethical and political standards towards it. The "AI of the Enlightenment" as understood by the Digital New Deal think-tank, is an "AI of trust" that refuses to let ideological biases - libertarian or authoritarian - guide its definition today, only to be translated into technological biases with a wide range of impacts: economic, societal, or ecological. Hence the importance of translating our values into global standards through the European regulatory package[3] that creates this space of trust for the benefit of citizens and companies (a foreign company will be obliged to respect this framework of trust if it wants to market its algorithms within the EU).
For our excellence in AI research to be translated into industrial leadership capacity, regulation alone is not enough. We must succeed in creating industrial cooperation at the European level, in line with our ambitions, to guarantee this digital sovereignty and to develop our industrial competitiveness. Writing this new chapter by 2030 is entirely possible. France for instance has already launched several ambitious programs that this report proposes to complete to increase its continental dimension and guarantee its impact.

In the following report, we present a method for achieving this at the European level. On the one hand, we propose making trustworthy AI the arrowhead of an ecosystem of trust and guaranteeing our strategic autonomy, and, on the other hand, we propose capitalising on our industrial culture (critical systems) to seize this historic opportunity of conquering the global AI market.

Olivier Sichel,
Chairman, Digital New Deal

[1] *Desindustrialisation of France:* 1995-2015, Nicolas Dufourcq, Odile Jacob, 2022
[2] *Statement by the Minister of Economy at the inauguration of the Interdisciplinary AI Institute*, octobre 2019
[3] GDPR (General Data Protection Regulation), DSA (Digital Services Act), DMA (Digital Markets Act), DGA (Data Governance Act), DA (Data Act), AI Act (Artificial Intelligence Act).

THE VOCATION OF EUROPE IS TO DEFINE AN AI OF ENLIGHTENMENT: HUMANISTIC, TRANSPARENT AND RESPONSIBLE.

I. IDENTIFY THE CHALLENGES TO BE MET

Artificial Intelligence is at the heart of concerns and fantasies, and its simple definition is a challenge in itself. It is therefore crucial that we bring our own, in order to offer the world a European (i.e. humanist) vision of this matrix subject which is becoming increasingly structuring in the economic, social, ecological and democratic fields. Europe must not repeat past mistakes by allowing a "siliconian" definition of AI to be imposed on it, as it did for the Internet. We must be able to visualise and operate an "AI of the Enlightenment" that is humanistic, transparent, and responsible. Humanism, as we understand it in the 21st century (i.e. human beings, their values, but also their interactions with their natural environment), is at the center of everything. Humanism has always aimed for the development of Man, and his confidence in his ability to evolve in a positive way while respecting his environment, an environment understood today in the broadest and most ecological sense of the term.

1.1. WHAT IS AI?

To simplify this report, we will use the term **"artificial intelligence" or "AI"** to describe AI-based systems, or systems based on AI-based technologies.

1.1.1. The humanistic challenges of AI

AI is unfolding in all aspects of our lives

The potential of artificial intelligence (AI) and its developments fuels a powerful imagination, reflecting the anxieties caused by major technological upheavals.

This potential responds to the anxiety of dispossession in the face of a technology that will be deployed everywhere health, finance, industry, security, transport, commerce, public service, all areas of our lives are concerned. Even if sometimes exaggerated, the potential of these disruptive technologies is nevertheless very real and already perceptible in many areas of our daily lives.

In this "algorithmic life"[4], **artificial intelligence already goes far beyond university research programs** and is gradually becoming an element that conditions the organisation of knowledge, the production of goods and services, and even decision-making, to the point of becoming a factor in the organisation of communities and a vector of power for States[5].

Exponential advances in information technologies (Moore's law[6]), ever-expanding connectivity (Metcalfe's law[7]), as well as access to ever-increasing masses of data (Big Data), continuously accelerate the progress of AI.

[4] *Algorithmic Life. Critique of digital reasoning*, Éric Sadin, L'Echappée editions, 2015.
[5] *Why artificial intelligence is the future of growth*, Mark Purdy and Paul Daugherty Accenture, 2016.
[6] Gordon Moore's empirical law of exponential evolution of computing power, according to which the power of a processor doubles every two years.
[7] Robert Metcalfe's theoretical and empirical law of the network effect, stating that the utility of a network is proportional to the square of the number of its users (the more users on a network, the more valuable it is).

The relationship between humans and AI

This new algorithmic life is full of challenges for our societies. For example, what will our place be in a world where AI-based systems are ubiquitous? How can we allow humans to retain their attention, interest, sense of purpose and dignity through their creativity, merit, and responsibility[8]? In the face of accelerating global warming, how can AI be used to help the environmental transition? How will AI impact the world of work and the economy?

The fight against global warming can benefit from the combined contributions of AI and data. Indeed, the goal of net zero greenhouse gas emissions (GHG) by 2050 implies decoupling economic growth from the intensive use of physical resources[9]. This is precisely what artificial intelligence can do through the analysis of large masses of data (Big Data). AI could therefore play a leading role by bringing more efficiency and transparency to many fields (Smart Cities & Digital Twins, transport, industry, construction, waste management, agriculture, circular economy, energy, etc.), or by improving our understanding of complex phenomena (nuclear fusion[10], etc.). However, we cannot exclude the "negative" contribution of AI[11], and more broadly of digital technology, which only a frugal approach, in data and energy, will be able to reduce.

Regarding the world of work, the large-scale deployment of AI raises the fear of a massive loss of jobs. Schumpeterian creative destruction[12] assumes that new jobs are created almost naturally when others, rendered obsolete by automation, are destroyed. But will our societies be able to absorb this destruction in the face of the frantic pace of technological change? The time needed to adapt to these changes is an inescapable factor, and the capacity of individuals to train and retrain is inevitably limited. What will happen to the relationship between humans and AI in the world of work?

All these questions, sometimes philosophical and ethical, should neither be evaded nor exploited. It is up to us to build healthy relationships with these technologies and to anticipate their impact on the economy. **AI only makes sense if its use is aligned with the values and principles that human communities defend.** Only then will it be accepted by individuals, markets, and society, and become a vector of prosperity.

To do so, the development of AI must consider long-term societal issues such as the protection of citizens, the environment, employment and democracy. **The notion of "trust" becomes essential: a fundamental underpinning of relations in our societies, an indispensable component of the response to humanist challenges, and of respect for the fundamental rights defended by the European Union.**

The purpose of this report is not to be exhaustive on the humanistic challenges of AI, nor to offer a precise analysis. Several books and reviews deal specifically with these subjects. Rather, **this report focuses on "trust", a notion that is as complex as it is fundamental, including in the field of AI.**

[8] Laurence Devillers - 3 dimensions of human-machine interaction, following the work of the laboratory on the future of work operated by the Matrice innovation institute.
[9] *The Role of Artificial Intelligence in the European Green Deal,* Policy Department for Economic, Scientific and Quality of Life Policies Directorate-General for Internal Policies, European Parliament, Mai 2021.
[10] https://www.inria.fr/fr/fusion-nucleaire-comment-simplifier-la-simulation-des-plasmas
[11] N. Thompson et coll. *Deep Learning Diminishing Returns.*
[12] Refers to the ongoing process in economies where sectors of economic activity disappear simultaneously with the creation of new ones.

1.1.2. A lack of consensus on the definition

AI is taking an increasingly important place in public debates, but remarkably, there is no consensus on the definition of artificial intelligence, and we do not currently have a precise and general legal definition.

Multiple definitions

Marvin Minsky (1927-2016), one of the founders of artificial intelligence, defined artificial intelligence as follows: **"AI is about making a machine do what humans do with some intelligence".** The AAAI (Association for the Advancement of Artificial Intelligence) has taken as its definition: **"The scientific understanding of the mechanisms underlying intelligent thought and behaviour and their embodiment in machines"**.

Most current definitions of artificial intelligence (AI) include the notion of learning. Many human reasonings are based on an empirical approach of statistical or probabilistic observations. An automatic decision-making process based on such an approach can therefore reasonably be characterised as AI, in the sense of a partial imitation of human decision-making capacity. **AI is thus considered as a set of technical and scientific mechanisms making it possible to reproduce most human cognitive processes such as learning, intuition, self-improvement, creativity, task planning or understanding and natural language generation.**

But compatible with each other

Many definitions exist, for example those of the ISO-24765, the Academy of Technologies[13] or the European Commission[14], but all of them are compatible with each other. They include the cognitive capacities given to machines (hardware and software) to carry out tasks requiring intelligence when they are executed by humans. Examples of the application of these capabilities can be found in autonomous vehicles, conversational agents, or image recognition. The technologies serving these applications are based on knowledge representation, reasoning, machine learning, planning, etc.

1.1.3. How is AI different from "classical" algorithms?

Classical algorithms

In general, computer science is based on algorithms. An algorithm can be thought of as a set of rules ordering a finite number of operations to solve a problem. To function, a computer needs a programmer to provide it with a computer program composed of multiple algorithms, written in a language that the machine can interpret. The algorithms translate inputs into outputs. **In classical algorithms, the machine applies the steps of the algorithm to an input, in the requested order, until it obtains an output. This is analogous to a cooking recipe in which all the components are known and described explicitly.**

A new logic brought by symbolic AI and subsequently connectionist AI

What we commonly call artificial intelligence (or AI) breaks with classical algorithmics, where the programmer describes how to solve a problem in a direct and explicit way. Historically, the design of AI algorithms emerged in the 1950s through two streams. **Knowledge-based AI,** now referred to as Good Old Fashioned AI (GOFAI) or symbolic AI, which is based almost exclusively on symbolic reasoning and logic. It differs from **data-driven AI,** also called **statistical and connectionist AI,** which has come to the fore in recent years with the massive collection

[13] *Rapport Renouveau de l'Intelligence artificielle et de l'apprentissage automatique*, report published in 2018.
[14] European Regulation on Artificial Intelligence, European Commission, 2021.

of data ("Big Data") and the arrival of sub-symbolic AI (and deep learning), although also old.

Symbolic AI, as defined by Nicholas Asher (Scientific Director of the Interdisciplinary Institute in Artificial Intelligence ANITI), **"uses formal reasoning and logic; it is a Cartesian approach to intelligence, where knowledge is encoded from axioms from which consequences are deduced. The prediction must be correct even if we do not have exhaustive data**." This paradigm remains relevant for solving complex problems under constraints, in a context of uncertainty.

Data-driven AI and machine learning, especially connectionist AI, on the other hand, seek to learn from examples or discover regularities or patterns in data, without first knowing them. Connectionist approaches also have a generalisation capability that allows them to process a case or recognise a pattern under different conditions than learning.

However, their fields of use differ. **While connectionist AI is the AI of the senses (perception), symbolic AI is the AI of meaning (reasoning).** Several works seek to hybridise these two paradigms. As Nicholas Asher points out: **"The addition of these two currents, symbolic AI and connectionist AI, constitutes today's challenge". We call this 'hybrid AI'.** For example, reinforcement learning consists of rewarding desired behaviours and/or punishing unwanted behaviours with reward or punishment strategies based on business knowledge or heuristics from symbolic AI.

The turning point of *machine learning*

The development of machine learning techniques, starting in the 1980s, marks a turning point: we are moving from a programming logic to a learning logic. *Machine learning* frees itself from expert rules and works on the basis of examples: the programmer provides the machine with a series of examples (inputs, outputs) that will "train" or allow the construction of a statistical model capable of determining an output from a given new input. The success of machine learning is such that today there is often confusion between AI and *machine learning*.

***Machine learning* is particularly well-suited for dealing with open world (in vivo) problems,** such as image recognition or automatic language processing (ALP), where the set of possible situations is difficult for a human being to know and master. Its performance will depend on the algorithmic techniques used and, in many cases, on the volume and quality of the data that will be used as examples, and that are provided or annotated by humans. A very large number of photos (some with the target object and some without) correctly annotated by humans is required for an image recognition program to perform satisfactorily.

As powerful as it is for image recognition, ***machine learning* is far from being relevant for solving all types of problems.** For example, *machine learning* is less relevant for problems with a finite and limited number of potential cases or parameters.

1.1.4. A varied scientific and technical field

The key sub-domains of AI

Artificial intelligence is therefore plural. To put it simply, this technology, or rather these technologies, cover a set of disparate disciplines, techniques, and media pursuing objectives similar to certain human faculties. In its white paper *Artificial Intelligence: Current Challenges and Inria's Action*[15], the French National Institute for Research in Digital Science and Technology (Inria) proposes a structuring into sub-domains**. Some of these AI subfields gained considerable importance in the 2000s thanks to the exploitation of Big Data. Here are some of them:**

- **Supervised learning:** In these algorithms, the input data (training data) is labelled beforehand by humans to tell the machine what it corresponds to. The model is trained from this data and must succeed in making classifications or predictions.
- **Unsupervised learning:** Here, the input data is not labelled and has no known outcome. The model itself must observe trends and patterns in the data.
- **Reinforcement learning:** Reinforcement learning consists of an AI learning the actions to perform from experiments, to optimise a quantitative reward over time. The agent is immersed in an environment and makes decisions based on its current state; in return, the environment provides the agent with a reward that can be positive or negative.
- **Neural networks:** An artificial neural network is a computer built to replicate the operation of the brain's nervous system connections. The artificial neuron is designed as an automaton with a transfer function that transforms its inputs into outputs according to precise rules. Assembled in a network, these artificial neurons can quickly operate classifications and gradually learn to improve them. The main idea of this assembly is to be able to learn a more complex function by multiplying the layers of neurons from a set of rudimentary and local calculations at the level of each neuron.
- **Deep learning:** The continuous improvement of machine learning performance reached a milestone in 2010 with the development of deep learning. A cutting-edge field of machine learning, this technique corresponds to a specific type of machine learning based on the use of large (deep) neural networks to learn abstract representations of input data, through the use of multiple layers. It therefore allows the machine to recognise complex concepts by itself. Born from the combination of learning algorithms with neural networks and the use of Big Data, deep learning has revolutionised AI. Applications can be found in search engines, medical diagnostics, and autonomous vehicles.
- **Natural Language Processing (NLP):** Among the most widespread applications, NLP covers several types of linguistic tasks: analysis of speech or text to extract meaning, automatic translation, text generation, automated database queries (questions and answers), etc. It works using several techniques (semantic indexing or Part-of-Speech tagging, Named Entity Recognition, Parsing, grammatical or semantic analysis), and is found, for example, in the voice recognition functions of our smartphones.

Whether they are optimised for neural networks or based on more traditional designs, **the efficiency of algorithms nevertheless depends on their ability to process an ever-increasing amount of data and variables.** In cases of combinatory explosion (i.e. too many variables for one algorithm), it is necessary to make strategic choices to obtain a result in a reasonable time.

[15] *Artificial Intelligence. Current challenges and Inria's action,* White Paper, coord. Bertrand Braunschweig, 2020 2nd edition.

The importance of knowledge representation

Today, even if connectionist AI has clearly taken the lead, symbolic AI also remains active. The design of knowledge-based systems capable of performing symbolic reasoning functions is a major field in AI. Such systems require an adequate representation of useful knowledge, as well as efficient reasoning mechanisms. Using models and methods of first-order logic (known as predicate calculus), symbolic AI has given rise to semantic networks and ontologies, knowledge-based systems, expert systems or constraint programming. Fuzzy logic can also be associated to it, even if it has the particularity of being associable to formal reasoning as well as to machine learning.

Based on the idea that **"intelligence is more related to knowledge than to a reasoning problem"**, Edward Feigenbaum, father of the first expert system DENDRAL, defined Knowledge Engineering (KE) in 1977 as **"the art of acquiring, modelling and representing knowledge for use by a computer".** For this, we can rely on:

- **Logical representations** built according to a precise syntax. Therefore, a knowledge base is a set of formulas describing the domain on which reasoning rules apply, as in the PROLOG language[16].
- **Semantic networks,** conceptual graphs[17] benefiting from reasoning mechanisms induced by graph operations such as graph homomorphism.
- **Ontologies** which constitute in themselves a data model representative of a set of concepts in a domain, as well as the relationships between those concepts. It can be said that "ontology is to data what grammar is to language"[18]. They are now used to model and share a set of knowledge in a given domain, for example in the Semantic Web[19] or in software engineering.
- **Inference, or reasoning,** which relies on deduction operations from implicit information. Therefore, this mechanism makes it possible to create links between information to draw an assertion, a conclusion or a hypothesis. For example, Bayesian inference is reasoning that allows us to deduce the probability of an event occurring or not.
- **Case-based reasoning methods,** introduced in the early 1980s by Roger Schank, are based on the idea that we sometimes reason by using analogies. These approaches are experiencing a new revival because they have the advantage of being more easily explained.

At the border between mathematical programming and AI, constraint programming (CP) appeared at the end of the 1980s, for the resolution of complex combinatorial problems such as planning, scheduling and resource allocation problems. This technology is based on the paradigm of separating the modelling of the problem from its solution. **"In computer science, of all the approaches to programming, constraint programming comes closest to the ideal: the user describes the problem, the computer solves it."**[20] Modelling the problem can include business knowledge and is done through a set of logical relationships: constraints. Constraint propagation mechanisms in a branching tree allow the reduction of the decision domain. The constraint solver then calculates one or more solutions by instantiating each variable at a value that simultaneously satisfies all the constraints. Today, many applications are deployed, for example, in the retail industry which has been optimising its logistics and inventory management for a long time.

[16] Alain Colmerauer and Philippe Roussel developed the PROLOG (Programmation Logique) language in Marseille in 1972, initially to process language. This program is a sequence of Horn clauses on which operates a reasoning mechanism using the resolution principle. Like LISP, Prolog uses massively the list structure and is naturally recursive.
[17] Conceptual graphs were introduced by John F. Sowa (researcher at IBM) in 1984 to formalise the difference between individual concepts (instances), generic concepts, and classes (types).
[18] https://fr.wikipedia.org/wiki/Ontologie_(informatique)
[19] The Semantic Web is an extension of the Web standardised by the World Wide Web Consortium (W3C)2. These standards encourage the use of standardised data formats and exchange protocols on the Web – Wikipedia.
[20] E. Freuder

Finally, **symbolic AI is often used to design decision support systems**. Let's recall that a decision problem consists of a choice, or a classification, between several mutually exclusive hypotheses resulting from a process that takes into account the knowledge one has about the state of the world, preferences and/or the objective to be reached. This knowledge can be fraught with uncertainty and preferences are inherently nuanced. A simple tool that can be used is the decision tree, which represents a set of choices in the graphic form of a tree. The different alternatives are the leaves of the tree and are reached according to decisions made at each step. However, the definition and use of one or more selection criteria is necessary. Contrary to the mono-criteria situation which can be solved quite easily, the multi-criteria decision requires more elaborate methods. Among the techniques used are the "What-if" method and multicriteria aggregation. The latter consists in globally evaluating the different candidates or proposed solutions, from the fusion of partial evaluations.

ARTIFICIAL INTELLIGENCE OVER TIME

1943-1955 In 1943, the work of McCulloch and Pitts introduces an artificial neuron model. A few years later, Hebb proposes a rule to modify connections between neurons. Minsky and Edmonds subsequently build **the first neural network.**

In 1950, Turing publishes his article "Computing Machinery and Intelligence" which explored the problem and proposed an experiment **(Turing test)** with the aim of identifying the moment when a machine would be able to imitate human conversation.

1956 The AI conference at Dartmouth College brings together the best international specialists of the time. The term "artificial intelligence" is used for the first time, and refers to "any aspect of intelligence, especially learning [that] can be described so precisely that a machine can simulate it"[21]. The Dartmouth conference is considered **the founding moment of basic AI research.**

1960's The period that followed was the rise of AI. In 1958, John McCarthy creates the LISP computer language (a name created from list processing) which facilitates AI programming. Many programs are developed mainly in the United States (Stanford, MIT, Carnegie-Mellon), but also in Scotland (Edinburgh) and Japan to solve various problems such as:

- Logic Theorist" and "Geometry Theorem Prover" which can **prove certain mathematical theorems**; and
- *"General Problem Solver" which succeeds in **solving puzzles with a reasoning like that of humans**. This period saw a continuation of research on neural networks and the invention of the **first robot capable of reasoning about its own actions**, the Shakey.*

In 1965, Lotfi Zadeh proposes an extension of classical logic that allows the modelling of data imperfections and comes close to the flexibility of human reasoning: **fuzzy logic.**

1970's At the beginning of this decade the capabilities of AI programs were limited, and the most powerful programs struggled to handle simplistic versions of the problems they were supposed to solve. **The power and memory of the time were a brake on practical applications**: natural language was limited to 20 words because the memory could not hold more. AI then experienced a less flourishing period. In 1969, Minsky and Papert publish "Perceptrons" in which they demonstrated the limitations of single-layer neural networks. **Many funds and grants are cut**, and several projects abandoned due to pessimism about the real possibilities of AI. This moment constitutes the **"first winter" of AI.**

But despite these dark years, **work did not stop**. Expert systems appeared between 1969 and 1979, imitating an expert to solve problems using rules. The first expert system, DENDRAL, was created to **determine the structure of a molecule from its formula** and the results of its mass spectrometry. Later, other expert systems were created such as MYCIN which aimed at **diagnosing blood infections at a level close to that of human doctors** with expertise in this domain.

1980's In 1982, Hopfield proposes associative neural networks and we witness a **renaissance of connectionism**. In 1986, Rumelhart, Hinton and Williams publish the gradient backpropagation algorithm which makes it possible to optimise the parameters of a multilayer neural network. From that moment on, **research on machine learning based on neural networks experienced a rapid growth.**

At the same time, the Chinook software becomes **the first computer program to completely solve the game of checkers:** whatever the initial situation, the computer is now certain to win the game. There is also a **development of learning algorithms which constitute the basis of deep learning.** During the decade, the computing power increases dramatically, while many sectors start using computers. This increase in raw material (data) and means (computing capacity) accelerates the development of the discipline.

The **"second winter" of AI** came at the end of the 80s and the beginning of the 90s. The limitations of knowledge-based expert systems led to the abandonment of many projects and a sharp decline in public and private investment.

1990's Despite the budgetary restrictions of the late 1980s, AI research continues particularly in the subfield of machine learning, based on the statistical analysis of large amounts of data. Such tasks include **recognising an image from a set or identifying and understanding words in a human language.**

[21] *Proposal for the Dartmouth Summer Research Project on Artificial Intelligence*, J. McCarthy, M. L. Minsky, N. Rochester, C. E. Shannon, 31 août 1955.

1997 The **Deep Blue** computer, developed by IBM, **beats Gary Kasparov at chess** in a 6-game match, making Herbert Simon's 1957 "prophecy" a reality 40 years later.

2000's The 2000s were an **unprecedented shift with the arrival of Web 2.0 and then Big Data.** The techniques abandoned in the 1980s are re-exploited, this time with increased resources.

2010 **Deep learning** gradually invades the discipline, and is developed in many fields **(visual recognition, machine translation, robotics, etc.).** Combined with mass training data **(Big Data), deep learning made it possible to obtain unexpected results** in very diverse domains: games, recommendation systems, image and speech recognition, natural language processing (generation, translation, automatic summaries, questions and answers), medical diagnosis, time series prediction.

2011 **Watson,** an AI computer program designed by IBM and whose first version does not use deep learning, **becomes the champion of the US television game show "Jeopardy!"** by beating its human competitors. Deep Blue and Watson are both supercomputers. Deep Blue (1997) is focused on probability calculation, while Watson, with its 90 servers and 2880 hearts[22], offers the computing power to answer Jeopardy! questions in the same time as its human competitors.

2015 **AlphaGo,** developed by the company DeepMind (acquired in 2014 by Google), **beats Fan Hui, the European Go champion.** AlphaGo's algorithm uses learning **techniques based on examples of numerous recorded human games. In 2017, the algorithm further increases its capabilities by outperforming all players, including the world champion Lee SeDol , and by being inventive in its choices. The latest evolution was contained in its successor, AlphaZero: the AI learns by itself, without prior examples of human games.**

2020's **The 2020s enshrined the development of trust and explainability issues in AI.**

2022 **French start-up NukkAI** beats the world's best bridge players using a hybrid AI program that explains its reasoning.

An explanation of the Bridge example

Nook, developed by the French company NukkAI, becomes the **first AI program to surpass 8 Bridge world champions** in the card playing phase. The game of Bridge had until then resisted AI and purely digital methods (based on deep learning for example) because it is a game with incomplete information where it is necessary to draw inferences from the actions and non-actions of opponents.

Nook is a hybrid AI in the sense that it is composed of three modules, each based on different AI paradigms: 1) a symbolic module which allows combinatorics to be restricted and decisions to be explained 2) a tree search module 3) a neural network module modelling opponents. **At each step of the game, it is possible to access "what Nook had in mind at the time of its decision" which allows a human to understand why the machine made a certain move.**

Moreover, **the use of symbolic methods to restrict combinatorics allows Nook to be very energy efficient:** the NukkAI challenge consumed 200,000 times less resources than the Go challenge.

[22] *IBM Watson : IBM Watson: all about the super computer*, CFTC-IBM, 05.07.2019

TRUST IS A WALL FOR AI, INSOFAR AS ITS OPPOSITE MISTRUST APPEARS AS A FACTOR OF SLOWING DOWN AND COULD LEAD TO A NEW AI WINTER.

1.2. THE NEED TO BUILD TRUST IN AI

1.2.1. What is trust in AI-based systems?

The systemic dimension of trust

Trust is fundamentally a relationship between two (or more) parties. It is the social cement that connects individuals and organisations within the same community. **But this trust can only exist within a larger framework that promotes and protects trust: the ecosystem of trust.**

This ecosystem, made up of a set of trustworthy actors, builds and maintains a body of values, principles, and rules, and develops a set of mechanisms to monitor behaviour, judge and punish breaches of trust. Its purpose is to encourage and protect the development of relationships between parties within a given environment by developing rules, standards and the means to implement them. In doing so, the ecosystem must manage risks, monitor activities, and anticipate new behaviours. A trustworthy ecosystem relies on an extensive and organised network of trustworthy third parties. These third parties are independent and can be organisations or individuals sharing common "principles": a regulation, an accreditation, an audit, a certification, a standard, a jurisdiction, and the implementing bodies that make possible and credible reliability requirements and assessments in a given environment.

The governance body, most often a State, but this can also apply to a consortium or organisation, **implements an extensive analysis of the threats that its trustworthy ecosystem may encounter, and defines criteria to ensure that the ecosystem and its third parties are indeed trustworthy themselves.** Unfair, negligent, incompetent, biased or corrupt third parties can threaten the balance of the ecosystem.

Trust therefore has a systemic dimension: a tool, an application or a relationship is part of a much larger whole that includes all stakeholders, and even society.

COLLAPSE OF TRUST IN THE ECOSYSTEM: THE CASE OF THE BOEING 737 MAX

On 29 October 2018, Lion Air Flight 610 crashed into the sea off Indonesia. On 10 March 2019, Ethiopian Airlines Flight 302 crashed six minutes after take-off from Addis Ababa. These two accidents, involving the new version of the Boeing 737 Max, claimed the lives of 346 people.

The U.S. Department of Justice (DoJ) investigation has since uncovered a defect in the company's stall protection software (MCAS), as well as a deliberate attempt by the company to withhold information from the certifier to speed up the introduction of the 737 Max into service. The DoJ shows that the stall protection software defects were compounded by inadequate pilot training. In addition, one of the pilots in charge of the tests was accused of having voluntarily misled the Federal Aviation Administration (FAA) by failing to report the difficulties related to the pilot software. The pilot even boasted that he was able to deceive his FAA contacts to have the system certified.

The financial consequences of these malfunctions were drastic for Boeing, which has made a loss of over $20 billion since 2019 and recorded 565 order cancellations between 2019 and 2020.

The impact of such dysfunction is a loss of societal confidence not only in the offending actor (in this case Boeing), but also in the regulatory and certification mechanisms (the states, the FAA, etc.). This situation increases the risk that air transport in general is no longer trustworthy. From the software to the test pilots, via the company, the sector regulator, and ultimately the States involved, the breakdown of trust is a systemic risk, and **if the failure of one company can lead to the failure of the whole system, it is also a societal risk.**

The serial failures of the Boeing 737 Max show to what extent trust, translated into regulatory mechanisms, is an invisible institution that supports the whole of society. There is therefore no isolated risk when it comes to trust.

Sources: "Suspension of the Boeing 737 Max flight", Wikipedia, "Crash of the 737 Max: Boeing admits to having deceived the European regulator", Le Figaro, "737 Max: a former Boeing test pilot indicted for fraud", Les Echos.

Why is it difficult to trust AI?

The notion of trust is evolving and is now applied to products and services resulting from human activity. Therefore, we want to have confidence in a car, in a vaccine, in an AI, etc. In these specific cases, trust has essentially developed around the notion of reliability, namely the fact that the product functions without failure for as long as possible. In addition to this, many other notions have been added, such as safety or the ability to protect oneself from events that threaten the safety of goods and people. **In the technological and digital world, the characteristics of trust have multiplied, and we now expect products to be valid, robust, resilient, fair, understandable, ethical, reliable and reproducible. AI is no exception to this and carries a series of inherent risks, including the difficulty in reporting the results of algorithms, which feeds the opacity of AI:** "We do not know how to prove that the conclusions of a system trained by learning on a database are the right ones, that they are robust to small variations, that they are not tainted by bias etc"[23].

In the rest of this section, we present certain targeted issues. They should not be understood as an exhaustive list of issues to be resolved for the constitution of a trustworthy AI, but rather as an illustration of the complexity of the subject, and the need to achieve it.

The "black box" effect

"Much of the ethical concerns raised by AI stem from the opacity of these technologies"[24]

One of the strengths of contemporary AI is its unprecedented ability to process very large datasets, combining an extremely high number of variables. This makes it possible to automatically process new types of data such as digital images, language, molecular structures, etc. But this power has its downside: **one of the fundamental problems of AI lies in the difficulty of explaining not only how it works, but above all its results and how it achieves them. This is what some people refer to as the "black box".**

[23] "The 5 walls of AI, trust", Bertrand Braunschweig, Le Monde, 2022
[24] Giving meaning to artificial intelligence, Parliamentary Mission Villani, 2018

In the case of deep learning, which uses deep neural networks, there is no need for pre-established rules: the machine itself builds a model thanks to the statistical use of data, as mentioned above. More opaque, deep learning is nevertheless more efficient and performs better than simpler learning models (formal rules, simple decision trees, Bayesian networks[25]). Even if we can understand the functions that make up deep learning, their accumulation quickly becomes complex. **A "black box" (without access to the network features) or a "white box" (with access to the network features) is created as the operations proceed. In both cases, it is difficult to give a precise account of what the machine is calculating.**

Yet, it is necessary to explain artificial intelligence algorithms. Winston Maxwell, a legal researcher at the *Institut Mines-Télécoms*, gives two reasons. **On the one hand, "individuals have the right to understand and challenge an algorithmic decision"[26].** If AI systems are used in services (private and public), it must be possible to explain the decision to the data subject and, in the event of a dispute, the individual and the justice system must be able to understand the result of the algorithm to be able to challenge it legally. **On the other hand, "it must be guaranteed that a control institution such as the *Commission nationale de l'informatique et des libertés* (la CNIL is the French regulator for personal data related issues), or a court, can understand the functioning of the algorithm,** both as a whole and in a particular case"[27].

The issue of algorithmic bias

Human cognitive behaviour is far from perfectly rational and is affected by dozens of biases documented by cognitive science. AI models, whether symbolically coded or data-driven, also incorporate biases. Some are due to inadequate problem formulation by AI system designers; others to the imperfect availability or selection of training data in statistical AI systems. Compared to human biases, which are not necessarily uniformly distributed among individuals and which can be balanced, counterbalanced, or corrected by education, **algorithmic biases pose a specific problem: their easy replicability and the risk of a very large diffusion, and therefore the possibility of the emergence of systemic biases.**

The problem of algorithmic biases has therefore become visible on a global scale with the digital giants and the example of Google's advertisement targeting (recommendation of lower-paid job offers to women, reproduction of discriminations already existing in the data). It runs the risk of creating a general mistrust in AI, a prejudice that could considerably slow down its development, and above all impact our societies.

The definition of bias depends on its context. In artificial intelligence, the notion of bias refers to the idea that each case requires an adapted treatment. In this sense, bias is what allows machine learning systems to judge whether a situation is different from another ("discriminate") and allows the system to adapt its behaviour. Bias is therefore fundamental for the learning process and the adaptation of the behaviour to a particular situation[28].

On the other hand, at the societal level, the term bias refers to the injustice that can be caused by the difference in treatment. To avoid confusion, ISO prefers to use the term unfairness in AI. Biases in training data are also a major source of bias in AI systems. In addition, human cognitive biases affect data collection, data processing, system architecture, training model, and other development choices.

In summary, some biases are essential to the proper functioning of an AI system, but undesirable biases can be introduced unintentionally and can lead to unfair results.

[25] "Illuminating the black box of algorithms", Florence d'Alché-Buc, Winston Maxwell, Antonin Counillon, l'MTech Institut Mines-Télécoms, 22 February 2021.
[26] ibid
[27] ibid
[28] ISO-IEC 2022 DIS 22989 Information technology — Artificial intelligence concepts and terminology, p.39

Deep learning algorithms, which rely on massive data (Big Data) to personalise content and assist in decision-making, **raise fears of a technological reproduction of pre-existing social inequalities, which are furthermore built on a "vision of the past" (the data).** However, if we are to live with AI systems daily and trust them, these systems must comply with the laws and social norms that we have given ourselves. **Moreover, this ethical requirement questions the status of these algorithms in society: should they remain neutral at the risk of amplifying biases, or should they correct existing biases in society?** The necessary conformity of algorithms to our norms cannot, however, arise solely from will: **"this requirement necessitates the development of procedures, tools and methods that make it possible to audit these systems in order to evaluate their compliance with our legal framework".**[29] **»**

A CHATBOT BECOMES A NAZI IN A FEW HOURS ON SOCIAL NETWORKS

In March 2016, Microsoft introduced its chatbot "Tay", an AI-based conversational agent, on the social network Twitter. Tay is not the first chatbot created by the American giant, which benefited from the experience of its first born "Xiaoice" in China, trained on 40 million users.

According to Microsoft, Tay was modelled on the typical personality of a young American woman aged 18 to 24, and had to offer a "positive experience". In the 24 hours following its release on Twitter, Tay suffered a "coordinated attack, launched by a few individuals" to exploit its vulnerability. Specifically, the chatbot went from tweets such as "humans are super cool" to a call to feminicide, the apology of white supremacism, anti-Semitic murder and to a "meme" praising the "coolness" of Hitler.

The chatbot is not "aware" of what it is talking about, and this is what the people involved have exploited. The individuals in question, who discovered the flaw, are users of 4chan and 8chan, two online forums now known to host debates by American white supremacists.

In short, a conversational agent trained with xenophobic users will produce behaviour and outputs of the same nature. If it is impossible to protect an AI from the formation of biases that exist in the open world, the Tay experiment shows the neutrality of the conversational agent, which evolves in mirror image with Internet users. The experiment also questions the possibility of assigning pre-existing rules to machine learning techniques that generate their own rules.

Sources: "Has Microsoft's AI really become racist through contact with Internet users?", Erwan Lecomte, 03/2016, *Sciences et Avenir*, "Microsoft explains why its Tay chatbot became a Hitler fan, the company talks about a coordinated attack", developpez.com, 03/2016.

The generation of inequitable results may be due to the nature and distribution of the learning corpus during design. As illustrated by the Tay chatbot example, these results can also be obtained in the case where the system learns from the data provided as input under operating conditions. Certain good design practices should be implemented to avoid, at least in part, data bias. **In the designer's toolbox, one should include an estimate of resilience to deliberate misuse,** an estimate of robustness to estimate the maintenance of performance in "normal" situations, and also the implementation of "safeguard" mechanisms that permit the limitation of harmful results in certain foreseeable cases (hence the importance of human control[30]). **These safeguards are most likely not an unstoppable solution,** but the regulation moving towards a risk-based approach (via the AI Act, mentioned later in this report), and a risk analysis practiced

[29] "Giving meaning to artificial intelligence", Giving meaning to artificial intelligence", Parliamentary Mission Villani, 2018.
[30] One of the guidelines proposed by the High Level Expert Group to the European Commission.

in a systematic way for any AI system would ensure that the system is designed in a way that is adapted to the final use case, which involves inherent risks.

The designer bears some responsibility for ensuring that there is no bias (provided he has the adequate engineering and inspection tools), **but the user must also use the system with an adequate level of control**, be sufficiently informed in order to use the system correctly, and be aware of his responsibilities as a user and member of society.

1.2.2. What is the basis for trust in AI?

GLOBAL INITIATIVES TO DEFINE TRUSTWORTHY AI

Trust is at the forefront of recent global discussions among researchers, industrialists, and NGOs in the field. The momentum begins in 2018 with the publication of 45 works from across the spectrum, and at least 117 papers addressing AI principles published between 2015 and 2020, the majority of which are by companies[31].

These works, provided by major institutions and industrialists, constitute a large base of studies to formalise both the concepts and the uses we will make of AI. Numerous working groups have produced their recommendations on the subject, especially in multilateral organisations. The production of standards for AI is the subject of simultaneous work by the world's largest economies and standards bodies. Similarly, the provision of material for implementing artificial intelligence where it is deemed useful is common among multilateral organisations. Most initiatives naturally produce concrete AI development work. Finally, ethics, a topic commonly associated with AI issues, is not necessarily the preferred angle, nor is public awareness of the issue, which only the Big Tech Partnership officially declares to tackle. However, the Council of Europe has, for a few years, been working on a multilateral text setting out rules for AI from a human rights perspective.

The Partnership on Artificial Intelligence

In late 2016, **Amazon, Google, Facebook, IBM, and Microsoft announced their alliance in a not-for-profit organisation,** the Partnership on AI to Benefit People and Society, with the goal of educating the public about AI and sharing AI research and best practices. In particular, the Partnership focuses on ethical and legal issues as they relate to the challenges posed by AI. The long list of partners[32] in the Partnership on AI suggests a willingness to provide common resources to move in the desired direction of ethically acceptable AI.

UNESCO study on ethics of AI

In November 2021, **193 countries committed themselves to the UNESCO Recommendation on AI.** UNESCO bases its approach on **"desirable values in itself", deduced from the ethical impacts of AI** from the axes of respect, protection of dignity, human rights and fundamental freedoms and based on 10 fundamental principles[33]. **This recommendation remains in the realm of soft law:** compliance with its principles remains voluntary, and sanctions are non-existent because it is a non-binding regulation.

The European Commission's HLEG. The High Level Expert Group on AI of the European Commission

The work of the European Expert Group on AI (AI HLEG) has been decisive in the Commission's approach to AI. The group's recommendations have served as a framework for legislative initiatives by the Commission and the Member States, including the Communication *on Building Trust in Human Centric Artificial Intelligence, the White Paper on Artificial Intelligence: a European approach to excellence and trust, and the Coordinated plan on AI*[34].

[31] *Gathering Strength, Gathering Storms: The One Hundred Year Study on Artificial Intelligence (AI100) 2021 Study Panel Report.*Stanford University, p.41
[32] https://partnershiponai.org/partners/
[33] *193 countries adopt first-ever agreement on the Ethics of Artificial Intelligence*, UN News, 25 november 2021
[34] https://digital-strategy.ec.europa.eu/en/policies/expert-group-ai

The OECD's Expert Group and Observatory. OECD created an AI Group of Expert (AIGO)

The OECD Artificial Intelligence Policy Observatory is being formed following the publication of the Recommendation on Artificial Intelligence ("OECD AI Principles"[35]), the first draft of an international standard on AI, adopted by some member countries in May 2019. These principles served as the basis for the adoption of the G20 Principles[36] in June 2019. As well as the long list of participating Member States[37], the initiative also includes research centres from MIT, Harvard or Inria for France, and various tech giants such as IBM, Microsoft, Google/DeepMind or Facebook.

The Global Partnership on Artificial Intelligence (GPAI)

The Global Partnership on Artificial Intelligence is a multi-stakeholder initiative launched following the June 2020 G7 by France and Canada. The GPAI aims to deepen the OECD Recommendation published in May 2019 by supporting fundamental AI research, and its industrial and commercial applications. The Global Partnership is organised around several thematic working groups, including the Responsible AI Working Group[38]. The working group has a mandate related to the overall mission of the GPAI, namely to **promote and contribute to the development, use and responsible governance of human-centered AI systems, consistent with the Sustainable Development Goals.** The GPAI currently has 25 member countries.

The Global AI Action Alliance (GAIA) of the World Economic Forum (WEF)

In January 2021, the World Economic Forum (WEF) launched the Global AI Action Alliance (GAIA) to accelerate the deployment of inclusive, transparent and trustworthy artificial intelligence. This cooperation brings together more than **100 stakeholders composed of companies, international organisation**s, NGOs and academics working to **maximise the benefits attributable to AI and minimise its risks**[39].

Standards bodies ISO IEC IEEE ITU CEN-CENELEC SAE

The world of standardisation has largely taken up the subject of AI. At the international level, **ISO and IEC have created a joint committee on AI (SC 42), which is aggressively developing "horizontal" standards on AI**, ranging from terminology to governance and management of AI by organisations, and including specifications, trustworthy AI, engineering, interoperability, etc. At the European level, CEN and CENELEC formed the Joint Technical Committee 21 "Artificial Intelligence" (JTC 21) in 2021. **Its first ambition is to support the European regulation of AI through the development of "harmonized" standards**[40]. Future European standards will therefore strongly shape the European and global markets. For its part, the IEEE (Institute of Electrical and Electronics Engineers) has launched the Global Initiative on Ethically Aligned design of AI Systems, following the publication of Ethically Aligned Design[41], a scientific analysis of high-level principles and actionable recommendations.

More than 200 working groups identified around the world

Part of the GPAI, **the Future Society think tank lists 200 artificial intelligence working groups and initiatives**[42], all committed to the idea of human-centered AI and human well-being. The scope of "Responsible" artificial intelligence as defined by the Future Society working group's mandate is vast, which is illustrated by the plethora of initiatives seeking to provide guidance on how artificial intelligence should develop and be adopted, or how it can be used to implement the AI for Social Good[43] agenda in service of the Sustainable Development Goals[44]. These programs, **led by academia, the public and private sectors, civil society and NGOs,** have in common the desire to shape responsible artificial intelligence through formal or informal mechanisms.

---

[35] OECD AI Principles overview, https://oecd.ai/en/ai-principles
[36] https://oecd.ai/en/list-of-participants-oecd-expert-group-on-ai
[37] Responsible AI Working Group Report, GPAI — Montreal Summit 2020, executive summary
[38] *World Economic Forum launches Global artificial intelligence alliance*, Hindustan Times, 28 janvier 2021
[39] *World Economic Forum launches Global artificial intelligence alliance*, Hindustan Times, 28 janvier 2021
[40] A harmonised standard is a standard published in the Official Journal of the European Union and linked to a European legislative act. This standard applies to all countries in the EU zone. The conformity of a product or service to a harmonised standard constitutes a presumption of conformity to the European regulation.
[41] https://standards.ieee.org/content/dam/ieee-standards/standards/web/documents/other/ead1e.pdf
[42] *Areas for Future Action in the Responsible AI Ecosystem, The Future Society*, GPAI and CEIMIA, décembre 2020, p.5
[43] AI for Social Good is a joint initiative of the Computing Community Consortium, the White House Office of Science and Technology Policy (OSTP), and the Association for the Advancement of Artificial Intelligence (AAAI), in the form of an academic fellowship that has resulted in 5 workshops on US soil. "Artificial Intelligence for Social Good", Gregory D. Hager, Ann Drobnis, Fei Fang, Rayid Ghani, Amy Greenwald, Terah Lyons, David C. Parkes, Jason Schultz, Suchi Saria, Stephen F. Smith et Millind Tambe, March 2017.
[44] *"Areas for Future Action in the Responsible AI Ecosystem"*, The Future Society, GPAI and CEIMIA, December 2020, p.12

The attributes of trust in AI

When it comes to defining what trust in AI entails, the main challenge lies in understanding what that trust is based on, and therefore also its attributes. Many documents (standards, technical guides, white papers, scientific articles) deal with this issue, **but we note several limitations in the literature:**

- **Lack of a comprehensive and relevant list of trust attributes.** Accountability, governance, security, resistance to attack, lack of bias, etc. The attributes are numerous, each document defends potentially different objectives, and addresses different audiences. Overall, we can assume that standards and technical guides are generally aimed at designers and suppliers, while scientific articles are aimed at researchers. It is therefore difficult to have a coherent overview of trust attributes for each actor in the trustworthy AI ecosystem, and for each stage of the system life cycle (from design to use). Some attributes will not be relevant depending on the application case, the type of model, or the actor involved. It is necessary to have relevant trust attributes to guarantee the level of trust of each actor.

- **Lack of attribute hierarchy.** Depending on the texts, attributes are presented at different levels of granularity, which is a major hindrance to their adoption (both by AI designers, and by system compliance inspectors). For example, some texts state the notion of integrity as an attribute of trust; however, integrity is not an attribute in itself, as it is composed of a number of elements (completeness, accuracy, etc.), which must be analysed separately in order to estimate the integrity of the system. Therefore, many attributes are "too high level" to be operational, such as governance, accountability, etc., which require a breakdown into truly observable, quantifiable sub-elements.

- **Attributes that are not sufficiently "equipped"**. Designers need guides to understand how to develop trustworthy solutions, and inspectors need to know where to look for compliance assessment. General considerations, or attributes that are too high level, do not allow for consensus. Without restricting innovation by setting targets, it is necessary that the objectives to be achieved are broken down to the minimum level of granularity to allow, for example, the designer to understand how to make his system transparent in his own use case, and the inspector to have clear and adapted measurement points. It is not a question of setting targets that would restrict innovation, or of establishing minimum thresholds of acceptability that may not fit all cases, but rather of providing engineering and inspection toolkits.

Relying on different sources (norms, standards, etc.), academic and industrial teams working within the framework of the "Trustworthy AI" *Grand Défi* ("securing, certifying and making reliable AI-based systems" from the national AI strategy of the France 2030 plan - SGPI[45]), **propose extracting a few attributes considered to be priorities for trustworthy AI.** This selection of priority attributes seems to benefit from a reasonable consensus in the community, but shows the heterogeneity of the necessary verification points, which concern the responsibility of different stakeholders in the AI ecosystem, and whose verification modes vary between quantified estimation (performance, etc.) and expert observation. In addition, some attributes are generalist criteria, themselves including lists of attributes that are not always precisely and measurably defined by the AI industry.

In this context, the Confiance.ai program, the technological pillar of the "Trustworthy AI" *Grand Défi*, aims to develop an environment for the development of trustworthy AI. It therefore

[45] General Secretariat for Investment.

works, among many other technical subjects, on the definition, structuring and metrics of trust attributes in the context of AI deployment in critical systems. Its attributes are now grouped according to the capabilities they characterise: technical, interaction, ethical, and trust intermediaries (such as certification). Some examples include:

Technical attributes:

- **Reliability:** a property related to consistent expected behaviour and outcomes. Depending on the context or industry, and also on the specific product or service, data, and technology used, different characteristics apply and must be verified to ensure that stakeholder expectations are met. Characteristics of reliability include availability, resilience, security, confidentiality, safety, accountability, transparency, integrity, authenticity, quality, and usability. Reliability is an attribute that can be applied to services, products, technologies, data and information and, in the context of governance, to organisations.

  Sources: ISO-27000, ISO 5723BSI, ISO 24028 (3.42), HLEG 2019 ALTAI.

- **Dependability:** The ability to provide a service that can be trustworthy. This implies: availability for correct service; continuity of correct service; absence of catastrophic consequences on the user(s) and the environment (safety); absence of unauthorised disclosure of information (confidentiality); absence of inappropriate alterations to the system (integrity); ability to undergo modifications and repairs (maintainability), simultaneous existence of availability for authorised users only, confidentiality and integrity (security).

- **Compliance:** Demonstration that a characteristic or property of a product meets stated requirements.

  Source: CENELEC-EN50126

- **Traçability:** The degree to which a relationship can be established between two or more products in the development process, especially products with a predecessor, successor, or master-subordinate relationship to each other; for example, the degree to which the requirements and design of a given software component match. 2. The degree to which each element of a software development product establishes its purpose; for example, the degree to which each element of a bubble diagram refers to the requirement it satisfies.

  Sources: IEEE-610.12-1990, Adapted from IEEE glossary of Software Engineering Terminology.

- **Accuracy:** 1. A qualitative assessment of correctness, or lack of error. 2. A quantitative measure of the magnitude of error. 3. In the context of the quality management system, accuracy is an assessment of correctness.

  Sources: ISO/IEC/IEEE 24765:2010.

- **Data Quality:** The extent to which the characteristics of data meet expressed and implied needs when used under specified conditions. 2. Extent to which the data are free of defects and possess the desired characteristics.

  Sources: ISO-25024:2015, DEEL Project.

- **Safety:** The expectation that a system will not, under defined conditions, lead to a state in which human life, health, property, or the environment are endangered. 2. The ability to have acceptable levels of risk with respect to damage to people, businesses, software, property or the environment. 3. Absence of unacceptable risk of harm, i.e., human injury or death. No unacceptable risk.

TRUST IS A DEFENSIVE WEAPON AGAINST THE MONOPOLIES THAT WE ARE SUBJECTED TO, AND OFFENSIVE FOR COOPERATION THAT WE CHOOSE.

Sources: ISO/IEC/IEEE 12207:2017, ISO 9126-1:2001, EN 50129.

- **Robustness:** Robustness is an important characteristic for ensuring user confidence because it allows AI systems to maintain their performance levels under widely varying operating conditions. It is defined as 1. the ability of a system to maintain its level of performance under various circumstances, 2. (Global) the ability of the system to perform its intended function in the presence of abnormal or unknown inputs / (Local) the extent to which the system provides equivalent responses for similar inputs.

  Sources: ISO-22989:2021, DEEL project.

- **Correctness:** 1. The degree to which a system or component is free of faults in its specification, design, and implementation. 2. The degree to which the software, documentation, or other elements meet the specified requirements. 3. The degree to which the software, documentation or other elements meet the needs and expectations of users, whether specified or not.

  Source: ISO-24765:2017

- **Maintainability:** The ability to extend/improve a given system while maintaining compliance with unchanged requirements.

  Source: DEEL project.

- **Verifiability:** The ability to evaluate an implementation of requirements to determine whether they have been met.

  Source: ARP4754A

- etc.

Interaction attributes:

- **Explainability:** The extent to which a model's behaviour can be made understandable to humans. The property of an AI system that is able to present the important factors that influence the results of the AI system in a way that is understandable to a human. Explanatory: systems provide evidence or reasons for all outcomes. Meaningful: systems provide explanations that are understandable to individual users. Accuracy of explanation: the explanation correctly reflects the process used by the system to generate the results. Limitations of knowledge: the system works only under the conditions for which it was designed or when the system achieves a sufficient level of confidence in its results.

  Source: hal-03176080, ISO22989, NISTIR8312.

- **Transparency:** 1. Open, complete, accessible, clear, and understandable presentation of information. 2. Property of a system or process to imply openness and accountability. 3. Property of an organisation that appropriate activities and decisions are communicated to relevant stakeholders in a complete, accessible, and understandable manner. 4. Ownership of a system whereby appropriate information about the system is communicated to relevant stakeholders.

  Source: ISO 16759:2013, ISO 27036-3, ISO 22989

- **Accountability:** Being responsible for one's actions, decisions and performances.

  Source: ISO-24028:2020

- **Oversight and control:** Deliberate action or process to achieve specific objectives, and regular monitoring of that action.

Source: ISO-IECJTC1-SC42-WG1-N1298

- **Usability:** The degree to which a product or system can be used by specified users to achieve specified objectives effectively, efficiently, and satisfactorily in a specified context of use.

  Source: ISO/IEC 25010

- etc.

Ethical attributes:

- **Fairness:** 1. Impartial and fair treatment or behaviour without favouritism or discrimination. 2. Equity refers to a variety of ideas known as fairness, impartiality, egalitarianism, non-discrimination, and justice. Equity embodies an ideal of equal treatment between individuals or groups of individuals. This is generally referred to as "substantive" equity. But equity also encompasses a procedural perspective, namely the ability to seek and obtain redress when individual rights and freedoms are violated.

  Sources: Oxford English Dictionary, HLEG 2019 ALTAI.

- **Privacy:** The assurance that an individual's privacy is preserved when the individual experiences an unlawful intrusion or collection of their data.

  Sources: ISO/IEC 2382:2015, 21262637

- **Diversity and inclusion:** Attention paid to the representativeness of training data by taking into account the diversity of profiles (e.g. ethnic representativeness, parity, age, religion, etc.).

- **Subliminal techniques:** Subliminal techniques are stimuli (images or sounds) incorporated into an object or message to be perceived at a non-conscious level, with the aim of influencing individuals' behaviour. These techniques have been used in advertising and propaganda. The effectiveness of these techniques and their objectives are subject to debate. In the context of AI, the use of these techniques would be prohibited under Article 5 of the draft regulation on AI (AI Act[46]).

  Sources: ISO/IEC 2382:2015, 21262637

- etc.

Attributes for trustworthy ecosystem intermediaries:

- **Quality assurance:** The set of activities, throughout the project life cycle, necessary to provide sufficient confidence that a product or service conforms to stakeholder requirements or that a process follows established methodology. Potential synonyms include: assurance, product assurance, development assurance, design assurance.

  Source: Daniels, S. E., Johnson, K., & Johnson, C. (2002). Quality glossary. Quality Progress, 35(7), 43.

- **Audit:** A systematic and independent examination to determine whether procedures specific to the requirements of a product comply with the intended provisions, are effectively implemented, and achieve the specified objectives. 2. A methodical, independent and documented process for obtaining audit evidence and evaluating it objectively to determine the extent to which audit criteria are met. An audit may be internal (first party audit), external (second or third party audit), or combined (combining two or more disciplines).

  Sources: CENELEC-EN50126, ISO-27000:2018

- **Certification:** Third-party attestation of products, processes, systems or people. 2. A written assurance that a system or component meets its specified requirements and is acceptable

[46] https://eur-lex.europa.eu/legal-content/FR/TXT/HTML/?uri=CELEX:52021PC0206&from=FR

for operational use. 3. Formal demonstration that a system or component conforms to its specified requirements and is acceptable for operational use. 4. The process of confirming that a system or component conforms to its specified requirements and is acceptable for operational use.

Source: ISO-24765:2017

- **Regulatory Compliance (Lawfulness & Compliance):** Demonstration that a product feature or property meets the requirements set forth by the regulations. Ability to enforce the regulations in force.

  Sources: EN 50126.

- etc.

The human factor to guarantee trust

Humans play a major role in the ecosystem of trust. As noted in ISO/IEC TR 24028 "Overview of trustworthiness in artificial intelligence", it is necessary to specify who trusts whom, and in which aspects of AI development and use[47]. The individuals involved must trust the AI and the other actors in the ecosystem, but also inspire trust. **Identifying and defining all the human actors is a first step towards building the ecosystem.**

Among these, we can already list:

- **The user (the operator):** Their actions contribute to the AI algorithm producing a result. This result must be correct and in line with performance expectations (a "good" result), but it must also be in line with the user's values (moral, ethical values, etc.).
- **The individual impacted by the AI decision-making:** He undergoes the consequences of the decision-making without always having the possibility of influencing it. He is interested in his own case and wishes to be treated morally and fairly, and to understand the factors that played in his favour or against him.
- **The designer:** His mission is to control the system and identify malfunctions or anomalies to guarantee a good level of performance only associated with standards or legislation.
- **The auditor/certifier:** His mission is to control the system and identify malfunctions or anomalies to guarantee a good level of performance only associated with standards or legislation.
- etc.

The report "Ethical Guidelines for Trustworthy AI"[48] written by the Independent High-Level Expert Group on Artificial Intelligence (HLEG AI) commissioned in 2018 by the Commission, also identifies other stakeholders, such as providers or society at large.

To ensure a robust ecosystem, each stakeholder should be identified and trained in trustworthy AI. This may seem obvious to individuals acting in a professional context, such as designers or auditors, who may acquire these notions as part of their training. However, it is essential that each link in the ecosystem is aware of how it can inspire trust and contribute to the success of the whole. In the case of the system user, it is essential that they are trained in the proper use of trustworthy AI and are informed about the features of AI in an appropriate manner.

The person affected by the system's decision-making must also be aware that the processing is carried out by an AI, and must know his or her possibilities of recourse in the event that

[47] *It is insufficient to simply refer to 'trustworthy AI', but to specify who trusts whom in what aspects of AI development and use.* ISO/IEC TR 24028.
[48] *Ethical guidelines for trusted AI*. HLEG IA, 2019.

the result is contested. On these points, the AI Act proposes that high-risk AI systems follow the regulatory procedure of CE marking[49], which includes the provision of a user manual. The text also suggests that in certain cases where AI results could potentially be harmful to the individual, the individual should be informed that he or she is interacting with an AI. In addition, the European Parliament issued a recommendation in 2020 on civil liability for AI[50], calling for the recognition of the liability of the operator of the AI system, and noting the importance of identifying all the links in the chain of responsibility (user, designer, etc.), and proposing means of recourse in case of harm. The text therefore highlights the role of the user in the proper use of the system: as with any product from industry, the user of an AI system shares responsibility in the good or bad use of the product. In 2020, the Parliament proposed a draft report[51] emphasising the transmission of knowledge about AI and in particular the importance of educating people about what high-risk AI is and how to use it.

Finally, it is also necessary to consider the notion of sectoral culture: while certain disciplinary fields are accustomed to the use of statistical processing or computerised tools, such as industrial production and the medical field, the appropriation of AI solutions may represent a barrier to overcome in a large number of sectors.

The formalisation of the actors of the trustworthy AI ecosystem must therefore require, on the one hand, the identification of all the links and at what point in the AI life cycle they intervene, and also the determination of the expectations of each actor (performance, respect for values, compliance with regulations, etc.), as well as the setting of prerequisites so that each actor can contribute effectively to trust (training, information, etc.).

1.2.3. The difficulty of auditing and certifying AI-based systems

The difficulty of auditing AI

From the point of view of an auditor who is external to the organisation that produced the algorithm or set of algorithms giving rise to a decision (such as the granting of credit, the ranking of an employee, the price of a jigsaw during a sales period, the ranking of a video recommendation or of dating partners), the auditing task is complex for several reasons.

The purely transparent approach of the code (like for Parcoursup[52]), or the generated model, does not make much sense in AI, as Google DeepMind's algorithms compete with the best automatic language processing tools and have more than 280 billion parameters to date. Even older generation tools, such as airline yield management algorithms, involve layers of technology and linear and stochastic models[53] with several million equations. **The authors of these codes themselves often operate in silos: a team of developers focusing on one aspect of the algorithm, with a local performance metric. They sometimes let the algorithmic strains self-select based on performance** (A/B testing in the context of recommendation systems, for example[54]); machine learning model generation tools, such as those offered by platforms like Dataiku[55] or DataRobot[56], **help produce code that is decreasingly manual. These evolutions in software development make the readability of the code produced more complex and the authors of the code themselves find it difficult to perform in-depth audits.**

[49] CE marking ("European Conformity") was created within the framework of the harmonisation of European technical legislation. CE marking is a regulatory marking indicating that the manufacturer is responsible for the conformity of the product to all the requirements set by the European Union legislation applicable to this product - Wikipedia.
[50] European Parliament resolution of 20 October 2020 with recommendations to the Commission on a civil liability regime for artificial intelligence, 2020/2014(INL).
[51] European Parliament draft report entitled "Artificial Intelligence in Education, Culture and Audiovisual", 2020/2017(INI).
[52] *Parcoursup* is a web platform designed to collect and manage the assignment wishes of future students in French higher education. In 2020, it will manage nearly 660,000 students and 15,500 programs - Wikipedia.
[53] Which refers to the study of time-dependent random phenomena.
[54] https://aws.amazon.com/fr/blogs/machine-learning/using-a-b-testing-to-measure-the-efficacy-of-recommendations-generated-by-amazon-personalize/#:~:text=A%2FB%20testing%20provides%20invaluable,further%20adjust%20your%20training%20datasets.

Can we audit an algorithm "in a black box", without knowing either its technology or its "rules", whether implicit or explicit?:

- **Legally and technically,** one must first be able to access the algorithm in its actual conditions of use. Taking the algorithm out of the platform, putting it in a jar and observing it "cold" only allows partial testing. Auditing in real conditions can cause problems if the audited actor does not facilitate access. Developments in the European regulatory framework (DSA[57] and DMA[58], among others) seem to be moving in the direction of greater auditability and therefore of voluntary and mandatory participation by the auditor.

- **Contradictorily**, the probe technique used by the auditor must not be detectable by the algorithm which is to be audited. Indeed, if the behaviour is robotic, or even too atypical, the algorithm may over-adapt to the request and therefore bias its behaviour, or even debias it. Examples of this type of adaptation can be found in the transport or e-commerce sectors, where the prices provided to scraping robots[59] are not systematically the same as the prices provided to the public, their probability of purchase being much lower. Creating undetectable, synthetic use cases is a very complex subject in data science[60], more complex as the number of dimensions that characterise the use is high.

- **Mathematically**, detecting bias or deception amounts to exploring a very large space of possible queries for the algorithm (all historical customer configurations, the entire product catalogue, all usage conditions potentially influencing the algorithmic response). The objective is to find circumstances in which the bias or deception is "blatant". These circumstances should be both representative of the platform's usage and induce significant harm. Auditing "black box" algorithms for bias or deception[61] is becoming a research focus that is taking its place in AI conferences.

- **For probative value,** even if detection isolates significant areas of flagrante delicto (within the framework of the operated test), the sampling method still needs to be reproducible, and therefore statistically probative to the auditor. The statistical properties verified during the tests will then have to be adapted to the standards of proof of the regulatory authorities. For example, in the context of collusion, how can we prove that two algorithms send signals to each other indicating that they should abandon a pricing strategy in a given market?

The evolution of the regulatory framework, the political and media pressure, and the repeated abuse of dominant position by Big Tech players, suggest that the circumstances are right for an "algorithmic audit ecosystem" to emerge. This "RegTech" ecosystem should be able to emerge, and Europe has a certain role to play by using its sensitivity, its computer science and mathematical culture, and its research laboratories which are already positioned around these subjects, even if the most advanced institute in this area is currently Kathy Crawford's AI Now at NYU[62].

The challenge of certifying AI

Fundamentally, we distinguish between regulatory and voluntary certification. In the first case, the regulatory framework requires that the product be certified before being placed

[55] https://doc.dataiku.com/dss/latest/machine-learning/auto-ml.html
[56] https://www.datarobot.com/wiki/model-blueprint/
[57] Digital Services Act
[58] Digital Market Act
[59] The term web scraping designates a technique of automatic extraction of contents which are most often structured, on one or several websites - definitions-marketing.com.
[60] https://papers.ssrn.com/sol3/papers.cfm?abstract_id=3634235
[61] https://www.cs.bu.edu/faculty/crovella/paper-archive/minimization-audit-Neurips21.pdf
[62] 3SA - Simulation for the Safety of Autonomous Vehicle Systems

on the market, while in the second case certification takes place according to the choices of the entity that places the product on the market (competitive advantage, recognition of a quality requested by a customer, etc.). In all cases, **certification is based on a set of technical requirements and means of verifying compliance with these requirements. Certification is always carried out by a third party.**

For an AI-based system, it is therefore a matter of verifying that it meets a set of fixed criteria. The certification can concern the AI product itself (the algorithm, the AI solution), but also the associated processes (development, training, etc.), a person (the developer, the data scientist, etc.) or a system (management system, etc.).

PROCESS CERTIFICATION FOR AI – LNE

In 2021, the French National Laboratory for Metrology and Testing (LNE) created a voluntary certification of the design, development, evaluation and maintenance processes for learning-based AI solutions. The result of a consensus between developers, evaluators and end-users of algorithms, this certification enables developers and integrators to prove compliance with their clients' performance, regulatory, confidentiality and ethical requirements.

Verification is performed via an audit, for a certification which is valid for one year. For example, the inspection verifies the quality and relevance of the strategies implemented for learning, the quality of the data, the applicant's ability to properly document its processes, the proper identification of key people involved in the processes, or the quality of the information provided to the customer.

Source: Process Certification for AI, National Laboratory of Metrology and Testing.

To properly define a relative requirement, for example AI explainability, one must define what explainability is and how to measure it (and therefore the associated metrics). The same goes for robustness, performance, safety, or fairness. As previously mentioned, **the criteria that a trustworthy artificial intelligence must meet have not yet been set out in a relevant and exhaustive way, and the development and inspection tools do not exist in all cases**. Each trust attribute must be identified and associated with a metric to allow for its inspection, and eventually propose compliance thresholds for each of these attributes. **At present, certifying that an AI is trustworthy is possible, but this certification would only be based on a limited set of criteria i.e. those for which we have sufficient knowledge.**

Some trust attributes relate to the performance of the system. If we are talking about regulatory certification, i.e. certification imposed by law, it is not a question of determining whether or not the system is effective in performing its task, but rather whether or not the errors and anomalies are in violation of the regulations. **Several strategies are possible to demonstrate performance: formal proof or statistics.** Each of these methods has advantages and disadvantages that must be fully understood before choosing a product certification strategy.

Formal proof methods, used in classical algorithms, consist of formalising the system within a model presenting all possible behaviours of the system. Inspection consists of determining sets of properties and verifying them on the model. In AI, **even if there are well-advanced works for evaluation by formal proofs,** as for example for the robustness evaluation on neural networks (ISO/IEC TR 24029-1 "Assessment of the robustness of neural network", and ISO/IEC TR 24029-2 "Part 2: Methodology for the use of formal methods" under development), **the development of methods** for other performance metrics and for other types of AI algorithms

continues to be a research field.

Evaluation through testing consists of placing the system in a typical environment and observing its behaviour in response to relevant stimuli. For an AI system, this can be done by providing input data sets, or by using a simulator. In the case of a test on a sample of data (or scenarios), **the associated problems may be the availability of the data, or ensuring that the database is perfectly representative of the system's capabilities** (that it falls within its operating domain). In addition, it is also necessary to ensure that the database contains "unwanted" entries, i.e. entries that have been identified in the system's risk analysis as being possible entries that could generate harmful behaviour. Other issues include the distribution of databases and the amount needed for meaningful verification.

The use of a simulator makes it possible to perform massive verifications without being constrained by the difficulty of collecting a sufficient amount of data; however, the simulation first requires that the environment in which the system is used be modelled and that this model is itself qualified. **In the case of the autonomous vehicle, for example, the modelling of a complete driving environment, allowing for the evaluation of AI, is at the research stage** (see for example the 3SA project[63]) and the first commercial offers have been made (Dassault System, AVSimulation, Ansys, etc.).

AI certification is directly related to trust: for all actors in the trust ecosystem, this means that a third-party organisation has validated AI compliance. We understand that it is possible to certify an AI (the system itself, processes, etc.), but that a number of elements must first be advanced. Just as it is necessary to equip designers to develop trustworthy AI, **it is essential that inspectors be able to rely on tools and methods to characterise trust.** Verification, validation, and certification of conventional systems (which therefore fall outside the scope of AI) are already difficult tasks, even though there are technologies that can be exploited in the industry. The application to complex AI systems is an important task that needs to be addressed to be able to use these systems in critical environments such as airplanes, nuclear power plants, trains, hospitals, etc.

On the other hand, **as AI is part of a trustworthy ecosystem composed of different actors and elements, validation will have to rely on collaborations with specialists from various computer science disciplines,** as well as with scientists from other fields of expertise contributing to AI, such as psychologists, sociologists, biologists (in biomimetics, in particular), mathematicians, etc. We also note that **it is not only a question of certifying the system, but also the datasets that contributed to its learning (if any), the system and the subsystems,** as well as the tools used to create the final AI applications.

[63] 3SA - Simulation for the Safety of Autonomous Vehicle Systems

STRATEGIC AUTONOMY IS AN ABILITY TO GENERATE AND DEFEND AN ECOSYSTEM OF TRUST THAT ORGANIZES OUR INTERDEPENDENCIES.

II. ADDRESSING A DUAL POLITICAL ISSUE

"Trust is a defensive weapon against the monopolies we suffer from and an offensive weapon for the co-operations we choose"

The notion of trust is twofold. Taken from the perspective of regulation, it is clearly defensive, a "necessary evil", but not sufficient for Europe to become a credible, attractive, and competitive digital channel.

However, trust is also the source of an offensive strategy. It is not simply a set of rules that frame and secure our contractual relations with foreign platforms, **it is also a promise that conditions and unites our ecosystems.**

Imposing the terms of trust to protect us from those we are subject to is not the same thing as building trust to bind us to those we choose. These are two sides of the same coin that the European Union must know how to play.

2.1. TRUST, A DIGITAL SOVEREIGNTY ISSUE

2.1.1. Strategic autonomy, a reasonable ambition for sovereignty

"Sovereignty is the ability, alone or with others, to enforce values, interests, and above all, laws"

Sovereignty can be defined as the ability to analyse, decide or act according to a set of values, principles, interests and objectives without undue external influence, manipulation, or coercion.

Sovereignty applies to the digital space as much as to land, sea and space, and has the same aspirations to extend the rule of law to its field. It can be understood as **the capacity of States "to be obeyed, to impose their laws, to appear as if they must be respected in the digital space"**[64]**.**

The definition of "digital sovereignty" suffers from an excessive variety of interpretations. In the absence of a precise legal definition, **this notion ultimately remains dependent on subjective definitions,** thereby preventing any form of consensus. Yet, this is a central topic today!

What is strategic autonomy?

"We can understand strategic autonomy as a desire to free ourselves from dependence on an external power, even a friendly one. It is a fairly Gaullist vision of geopolitics according to which alliance does not prevent autonomy. The previous decade has been punctuated by these events which, by accumulation, have shed light on a European unthought, covering a generalised dependence not only on digital giants, but also on the law of foreign powers, even if they are allies"[65] *Laurence Houdeville and Arno Pons, Digital New Deal, 2021*

[64] Pierre Trudel, Professor at the University of Montreal.
[65] Trusted Cloud, a strategic autonomy issue for Europe", Laurence Houdeville and Arno Pons, Digital New Deal, 2021.

Sovereignty can therefore be defined according to the level of autonomy i.e. its ability to choose the level and nature of its dependencies. *Like energy sovereignty, where a State chooses not to be totally dependent on a single supplier or a single energy source, digital sovereignty also seeks to control* **its dependencies.** This is done either by **"doing it yourself", in the rare case of total independence, or rather by "doing it with others",** in a framework of controlled trust, and possibly also by **"doing it for others"** to demake oneself indispensable, and thus be able to dissuade anyone from threatening the balance of trust by controlling a component that is essential to everyone's sovereignty.

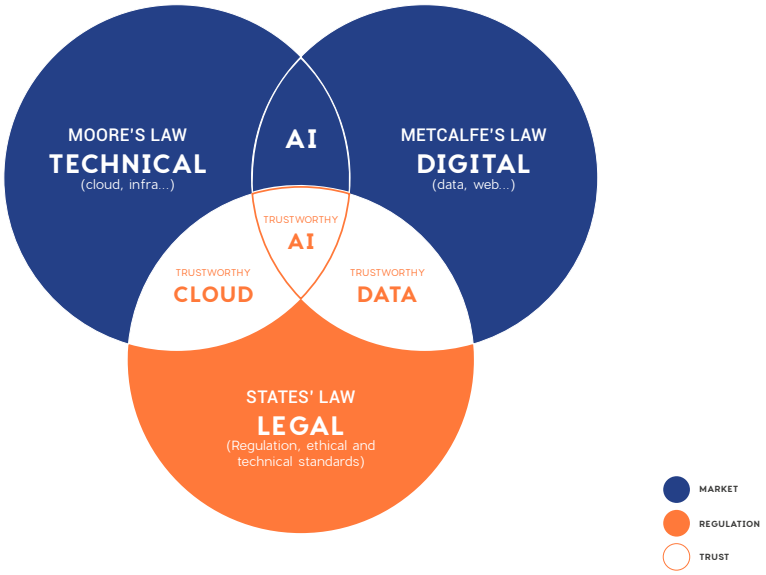
Digital sovereignty is therefore the management of our technological and economic dependencies. It is not a given, but rather a quest. Our levels of dependency invite us to think of sovereignty as a political ideal, a demanding and necessary strategic course that must guide our choices. **In short, sovereignty is an ambition, strategic autonomy an objective, and the level of dependence a measure.**

The ecosystem of trust, the natural perimeter of digital sovereignty and prosperity

Trust is not an ethereal concept independent of the rest of the world: **it can only develop in a "trustworthy" ecosystem made up of rules** (laws, ethics, technical standards, etc), trustworthy third parties, actors who define the rules and judge disputes, actors who share and respect this set of rules, and actors who anticipate future threats and behaviours weighing on the ecosystem of trust. A trustworthy ecosystem protects a country (or a group of countries), its economy, its citizens and must be protected by law and even by force if necessary. For a democratic country, it is therefore the quality of its ecosystem of trust that constitutes the basis of its strategic autonomy and therefore of its sovereignty.

"Strategic autonomy is an ability to generate and defend an ecosystem of trust that organises our interdependencies"

ECOSYSTEM OF TRUST (CLOUD + DATA + AI) = STRATEGIC AUTONOMY



THE EXAMPLE OF JAPANESE INDISPENSABILITY

While this strategic evolution is recent in Europe, it is rather old for other powers, in particular for Japan, historically protectionist and now very exposed to the unpredictability and arbitrariness of Chinese power. A study group of the Japanese Liberal Democratic Party has written a report urging the government to build an economic security strategy in response to threats from its powerful neighbour.

"*The report proposes to pursue two strategic objectives: 1) the preservation of strategic autonomy, which translates mainly into protective measures; and 2) the strengthening of* ***Japan's "strategic indispensability," a concept that aims "strategically, to increase the number of sectors in the global industrial structure, where Japan is essential to the international community."***[66]

Therefore, ***"Japan's technological and industrial positioning (...) should enable it to deter or retaliate against possible economic coercive actions,*** *as Japan had to suffer in 2010 when China halted its rare earth exports."*

In other words, through this strategic indispensability based on economic choices, Japan is trying to exercise a form of ***"soft deterrence"***. This is intended to preserve the country from brutal power struggles and to maintain its strategic centrality for rivals and allies alike.

Managing one's dependencies means choosing one's interdependencies and nourishing its corollary, indispensability. To protect this delicate balance, the quest for sovereignty consists of building and securing a "trustworthy ecosystem". An environment of trust based on technological mastery of a common material and immaterial infrastructure, a secure regulatory framework, the ability to defend oneself, and of course an informed choice of reliable and complementary stakeholders. Partners who share the same values of respect for data, openness and transparency, to the point of assuming voluntary dependencies with them to compensate for our weaknesses. This enlightened management of dependencies must be subject to dedicated governance and allow for interoperability. **If we cannot guarantee a fully sovereign (self-sufficient) technological value chain, the major challenge of this balanced governance will be to safeguard an international "chain or network of trust".** A mode of governance that can be composed on certain strata, but never be subject to external influences.

"For France to live up to its European ambitions, and to live up to its history, it must remain sovereign or decide for itself, without being subjected to them, the transfers of sovereignty that it would agree to, as well as the binding co-operations in which it would engage."[67]
Emmanuel Macron

Above all, sovereignty must not be approached from its protectionist side: it must be understood in a positive and constructive way. The quest for sovereignty is the quest for autonomy, not autarky. It is not so much a matter of imposing one's choices as of not being subject to those of others. This implies the prerequisite of taking the other into account, and of managing the relationship with the other. The policy aimed at technological sovereignty does not consist in cutting oneself off from the outside world, which would be heresy in a dematerialised and totally interdependent world, but on the contrary in managing our levels of interdependence and the associated inter-operabilities.

The whole art of sovereignty consists in choosing on whom we wish to partially depend, and at what level i.e. in whom we have sufficient confidence. It is no coincidence that it is in Europe that the notion of "digital trust" is taking shape, a continent where free movement is the foundation of our economic and political life, a place where prosperity is more than elsewhere inseparable from the notion of exchange. Between China, which advocates self-sufficiency,

[66] *Japan's ambition for an economic security strategy: a way forward*, Nicolas Regaud, Strategic Brief - 20, 15 April 2021, IRSEM
[67] Speech by French President Emmanuel Macron, 7 February 2020, on defence strategy and deterrence to the 27th graduating class of the *École de Guerre*.

and the United States, which assumes a form of interference via the extraterritoriality of U.S. law, **it is therefore a matter of the European Union inventing a form of sovereigntý which is synonymous neither with protectionism nor feudalism, but which consists, on the contrary, in the definition of a shared, open and multiparty strategic autonomy, made of accepted and reciprocal dependencies with trustworthy partners - a third way.**

Sovereignty is the interest of trust capital, a capital that must be invested and spent with caution, because as Jean-Paul Sartre reminded us, **"Trust is gained in drops and lost in litres".** Sovereignty must be able to rely on an ecosystem of trust that allows it to seal its members within its community of values and interests, and to impose its choices on those who are outside. **The European Union has patiently built this trust capital over decades, and it is now time to transpose it to the digital world.**

The world needs a third way between the two technological imperialisms, because countries must have the possibility of adhering to universal values compatible with multilateralism. In Europe more than anywhere else, it is desirable that the sovereignty of democratic states be associated with the confidence of citizens in their government and institutions, because without it there is no possible humanist model. **Sovereignty without trust is the unilateral exercise of force,** the undermining of multilateral trade, and therefore the hindrance to prosperity as we conceive it in the European Union.

Multitude as an expression of the ecosystem of trust

Choosing multitude means choosing numbers and diversity oriented towards a common goal. It is a strategy of cooperation and risk sharing that, by multiplying exchanges and interdependencies, also allows **for decentralisation: the distribution of functions and competencies among partners.** In other words, the perfect antithesis of centralisation and monopolistic concentration.

Choosing multitude also means protecting oneself by multiplying safety points. This multiplication of actors in charge of protecting and securing is a resilience strategy. Indeed, in the event of a crisis, if one agent or part of an agent fails, the remaining workforce can take over and allow the continuity of activities. The strategy of multitude therefore makes it possible to recover and mitigate the effects of a crisis by diversifying the distribution networks or the sources of supply i.e. the indispensable functions of an organisation. It is the network paradigm, flexible, robust and adaptable.

Oligopolies, resulting in the digital economy from the adage "the winner takes all", are precisely the opposite of the multitude approach. The strategy of centralisation and lock-in that they implement is also particularly vulnerable to the breakdown of trust.

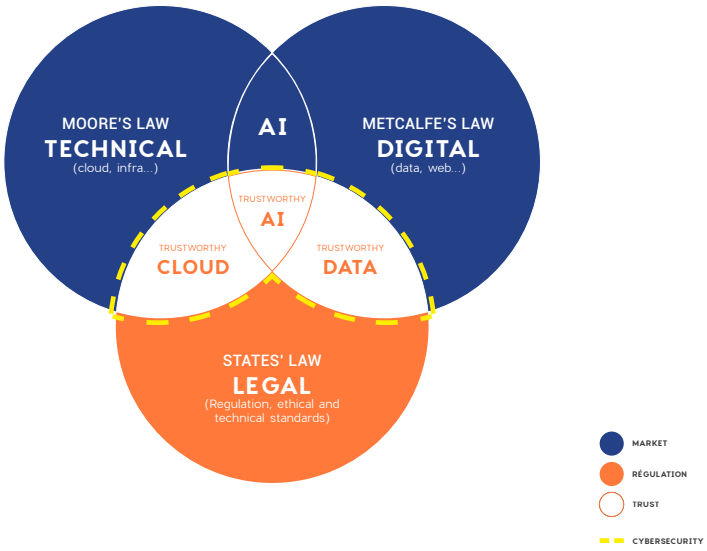
Fighting against the imperialism of the internet giants is therefore not a simple protectionist reaction or a desire for autarky, but a struggle against the centralisation of power, to create a democratic, sovereign, and trustworthy counter-model.

2.1.2. Protection as an underlying condition for trust

Sovereignty, an ability to influence and control threats

« A trustworthy infrastructure is the ability to control the independence of an architecture, and the ability to defend it.[68] » Guillaume Poupard, Digital New Deal, 2021

ECOSYSTEM OF TRUST (CLOUD + DATA + AI) = STRATEGIC AUTONOMY



To "generate and defend a trustworthy ecosystem", States must develop their influence and invest in security. Digital sovereignty implies degrees of technological control, with respect to one or more technological dependencies, but also degrees of security control with respect to threats external to the ecosystem. These threats aim at disrupting or altering the functioning of an entity and its sovereignty attributes. Attacks can be on data, software, human resources, hardware, digital infrastructure, or any of the entity's values, principles, interests, or objectives that constitute the digital assets.

Protecting means first of all securing ourselves through cyber security and improving our resilience. The issue of cyber threats is becoming more and more prevalent, and its counterpart, the power of coercion, is rarely mentioned. Yet there is no security without the threat of enforceable sanctions that are sufficiently dissuasive. A certain level of technological control, allowing us to keep control of the points of vulnerability, is a necessity. To exercise sovereignty in the digital space, we must be able to protect ourselves and respond to cyber threats, and keep control of the infrastructure as well as the data, both in the physical world and in the digital space. These threats are protean, and can target elements of cyberspace, organisations of vital interest (OIV), operators of digital means, individuals or even organisations.

A trustworthy ecosystem must also demonstrate strategic anticipation and resilience by demonstrating its ability to deal with an event or a disruption, by reacting or reorganising in such a way as to maintain its functions, identity, and essential structures, while retaining the capacity to analyse, decide or act.

Protecting is about safeguarding our interests through regulation. It is also a question of political will not to confine ourselves to a purely defensive and security strategy, but also to

[68] Statement by Guillaume Poupard, Director General of the ANSSI, interviewed for our publication *Trusted digital infrastructures. A strategic issue for territories*, Digital New Deal, November 2021.

conduct an offensive campaign aimed at influencing the market through the law. The European Union must continue to use its political influence fully through an assertive soft power, combining geopolitical weight, the ability to impose its values, to guide regulation, and to influence the norms and standards that structure the market. This European normative power must also create the conditions for an ecosystem of trust based on law by influencing its governance. *"The European Union must not hesitate to question the current political and legal framework if it wants to achieve total autonomy in its ability to assess and manage cyber risk"*[69].

The American "market" versus European "regulation"

"The economic asymmetry between Americans and Europeans is continually amplified by a legal asynchrony."

Europe and the United States differ in their approach to regulation regarding technology: **Europe favours *ex ante* management of technology-induced risks, while the United States prefers *ex post* control**[70]. These divergences can be interpreted by divergent visions of trust:

- **For Europe, trust requires the establishment of a trustworthy ecosystem based on laws:** trust is therefore first and foremost a matter for States. It is generally regulation that provides the framework for economic actors to innovate and develop, based on high-level requirements or values (ethics, fundamental rights, etc.). Europe tries to anticipate risks, both for individuals and for companies, and then puts in place strong market supervision and control processes, to reassure all economic actors. **This approach allows Europe to position itself as a global reference on regulatory issues but is nevertheless experienced by some actors as a brake on innovation, contributing to the relative absence of digital champions in certain sectors.**

- **For the United States trust is associated with a relationship between several entities (individuals, organisations, systems) framed by contracts. This vision offers greater freedom to innovate for market actors.** The weaker weight of regulation is counterbalanced by a judicial system that offers strong capacities for actors to defend themselves (e.g. the class action mechanism inherited from the English judicial system, which today inspires Europe), and antitrust mechanisms that allow the State to fight against monopolies (e.g. the dismantling of American Telephone and Telegraph, AT&T, in the second half of the 20th century[71]). **This "market-based" approach is certainly one of the many key factors in the commercial success of the American giants. The American approach allows its economic actors to learn and progress more quickly ("test and learn" or "fail fast").**

In recent years, **the impossibility for Europe to rely on local digital champions has led to numerous direct confrontations between European institutions and American digital giants,** particularly on issues related to personal data. **To the almost total American economic domination on B2C (Business to Consumer[72]) issues, Europe mostly responds with regulation:**

- **The GDPR** *(General Data Protection Regulation)* provides a regulatory framework for the protection of personal data;

- **The DSA** *(Digital Services Act)* creates a safer, more balanced marketplace that respects fundamental rights;

- **The DMA** *(Digital Markets Act)* limits anti-competitive foreclosure practices by large platforms;

[69] *Cybersecurity, guarding our strategic autonomy,* Arnaud Martin and Didier Gras, Digital New Deal, 2022.
[70] *Reflection in the perspective of the Digital Services Act,* Liza Bellulo, Digital New Deal, 2020.
[71] *"A.T.T. is accused by the American justice of violating the antitrust law", Le Monde, 22.11.1974,*
[72] *refers to all technical architectures and computer software that allow companies to connect directly with consumers – Wikipedia.*

- **The DGA** *(Data Governance Act)* and DA (Data Act) provide a framework for creating data-sharing ecosystems.

These regulations are intended to inspire global standards. For example, the GDPR now has a Californian equivalent in the form of the CCPA (California Consumer Privacy Act), and many states around the world have followed Europe's lead by adopting similar regulations. The influence of the regulation is also increased tenfold by the principle of extraterritoriality, which requires international players wishing to access the attractive European market, to comply.

"If these provisions [GDPR] had existed 20 years ago, it is likely that Facebook, Amazon or Google would not have penetrated the European market so easily and that competition could have started on a healthier basis."[73] However, Europe's levers of sovereignty are often reduced to sanctions, which are sometimes difficult to enforce, as evidenced by the disappointing history of effective fines: "Apple's €13 billion fine was overturned on appeal, Google has only paid €2.4 billion of a total fine of €8.2 billion, Amazon's fine was overturned, etc. For all GAFAMs combined, only €3.4 billion in fines has been paid over the past 20 years, compared to €300 billion in profits made in the year 2021 alone."[74]

Finally, it is almost illusory for Europe to impose the threat of dismantling American players who abuse their dominant position, or to protect certain aspects of its technological development when it does not master the required technologies.

[73] *"Giving meaning to artificial intelligence"*, Parliamentary Mission Villani, 2018.
[74] André Loesekrug-Pietri, "*European power against GAFAM: Myth or reality?*" *Les Echos*, May 2022.

“

THE ECONOMIC ASYMMETRY BETWEEN AMERICANS AND EUROPEANS IS AMPLIFIED BY REGULATORY ASYNCHRONY.

2.1.3. Achieving strategic autonomy through the ecosystem of trust

Regulation is a necessary but insufficient condition

The conceptual transposition of sovereignty applied to the digital world and the European space sheds a harsh light on our dependence on economic and geopolitical giants whose principles and practices are far removed from our own. However, **digital technologies reflect, in their very architecture, conceptions of society** and of the way a country perceives its role in the world.

China, for example, sees AI as a "pillar of tomorrow's harmonious society"[75], with the goal of becoming the world leader by 2030. Thanks to its computing power, its mass of available data and its skills, China is gradually rising to the technological level of the United States. **However, it is using these technologies to implement a "social credit" system - rating citizens and companies on their reputation, with a system of rewards and penalties - which is deemed unacceptable in the European regulation on AI (AI Act).**

CHINA'S SOCIAL CREDIT

The partial introduction of social credit since 2018 has given unprecedented visibility to China's digital positioning. Social credit is a rating system for citizens and companies on their reputation, with a system of rewards and penalties. Social credit relies on mass surveillance tools and uses Big Data to build a "network of trust" ("sho2o5xìn") and reduce opportunities for fraud.

The idea took shape in the 2000s in a desire to improve the solvency of companies and, at the time, of citizens. Beyond credit risk management, the objective was to encourage "integrity and credibility within society", and to improve the Chinese socialist market economy. Four objectives structure social credit: "honesty in government affairs", "business integrity", "societal integrity" and "credibility of justice". While social control has caused outrage abroad, the main purpose of the reform was to "provide an answer to the problem of lack of confidence in the Chinese market". Social credit would serve as a "market regulation mechanism" for the socialist economy.

The implementation of social credit in China shows that the invocation of trust in digital technology is not a guarantee. On the contrary, **this use of trust, considered an unacceptable risk in the European AI Act project, encourages us to rigorously define our principles, values and objectives, both politically and technologically.**

Sources: "The Social Credit System in China. Discipline and Morality", Séverine Arsène, Réseaux magazine, 2021/1 n°225, pp. 55-86 The social credit system in China | Cairn.info; "I ; "Xinyong : the expression of trust in China", Thierry Pairault, Moral Report on Money in the World, Finance et société, 1996.

Europe is therefore faced with a fundamental political imperative: to build its strategic autonomy around an ecosystem of trust, based on its values, by activating all the investment and sovereignty levers at its disposal.

Industrial sovereignty as a necessity for the ecosystem of trust

Protecting means making trust a strength for Europe in the global digital market, but also its key to entering the global AI market.

[75] *"The social credit system in China. Discipline and morality"*, Séverine Arsène, revue Réseaux, 2021/1 n°225, pp. 55-86; *"The social credit system"*, Wikipedia, ;*"Xinyong. The expression of trust in China"*, Thierry Pairault, Moral report on money in the world, *Finance et société*, 1996

Faced with competition from American and Chinese players, who are often better equipped financially, it is imperative for Europe to impose its values not only through regulation (AI Act), but also via an industrial strategy that is "market-driven", driven by uses and by technologies. Europe must carefully analyse the priority markets and value chains of AI in all sectors, and introduce or represent the notion of trust, which is what we propose to do in the following chapters.

This approach will promote the development of European AI champions and, above all, of dynamic and multiple technological ecosystems comprising the entire economic structure: public and private players, local authorities, large groups, start-ups, SMEs, universities, associations, etc.

Europe must define an ambitious industrial strategy that, on the one hand, promotes the offer through standardisation and, on the other hand, promotes the development of trustworthy AI solutions (InfraTech), as well as its adoption. Finally, it must also create new institutions and governance bodies to support this strategy.

MAKING OUR CULTURE OF SAFETY CRITICAL SYSTEMS THE TIP OF THE ARROW OF THE TRUSTWORTHY AI STRATEGY.

2.2. TRUST, A COMPETITIVENESS ISSUE FOR EUROPE

2.2.1. Making trust the market standard

Defining our own rules

There are two cultural approaches to trust: the first is weighted more towards the political sphere (European culture), the second towards the private sphere (Anglo-Saxon culture), Dthe definition of which is more flexible, even elastic, depending on the context and the case. Therefore, the asymmetry between actors in the ecosystem favours dominant positions and a global approach that is closer to the obligation of process than to the obligation of performance, not determined by use but in relation to societal criteria.

It is therefore up to States to ensure that trust becomes a "Digital Commons", especially in AI, for the benefit of citizens' trust in digital technology. We must compel the digital giants to play by our rules, and therefore allow our economic actors to be competitive **by changing the market benchmark.**

How can we do this? **By imposing our own rating agencies, by indexing compliance on our criteria and by promulgating our measurement tools so that we do not have to suffer the opacity of foreign private services**. By going beyond simple "ethics washing" that consists in creating a form of ambiguity between law and morality.

The European Union must pursue the strategy that it initiated with the GDPR, of imposing de facto market penetration criteria, and making trust the main issue. If it does not, Big Tech will reduce it to a mere commercial argument. Indeed, if we do not impose our own criteria of competitiveness, the American giants will have no trouble pre-empting trust by reducing it to a commodity. Their immeasurable financial resources and indisputable marketing skills will do the rest.

But is there a demand for trustworthy AI that justifies the development of the European offer?

To help us answer this question, the French General Secretariat for Investment (SGPI) asked Ernst & Young Parthenon (EY- Parthenon) to conduct a Trustworthy AI Market Study ("Without Trust, What Future for Artificial Intelligence in Industry?")[76] in a series of strategic high-risk industrial sectors (automotive, aeronautics, rail, banking, insurance, healthcare, energy, and power grid, etc.) in Europe, North America and Asia-Pacific. The purpose of this analysis is to identify and evaluate the main use cases, to date, where a "variable degree of trust" is required to deploy AI in industrial environments. The following section is therefore primarily based on this market study, which complements our report.

2.2.2. A cautious deployment of AI explained by industrialisation difficulties and a lack of confidence

Global AI market size

The artificial intelligence market was estimated to be worth **€231 billion in 2020, is expected to grow strongly by 18% per year by 2024** (Source EY- Parthenon) and is driven by the gradual adoption of AI solutions in all industry sectors.

[76] *https://assets.ey.com/content/dam/ey-sites/ey-com/fr_fr/topics/strategy/ey-intelligence-artificielle-dans-l-industrie.pdf?download*

Caution, or lack of maturity, on the part of industrials

"Although, on average, industrials (all sectors combined) invest between 0.4% and 1% of their turnover (led by technology players, etc.) in projects involving AI, **the industry remains cautious (or poorly endowed) to integrate AI components into industrial processes, products or services on a larger scale,"** recalls EY- Parthenon.

According to their estimates:

- **Only 10 to 15% of the companies surveyed have successfully industrialised** AI-based solutions;
- **30 to 40% of them are limited to experimentation** on limited perimeters or processes.

There are several reasons for these findings, according to the study:

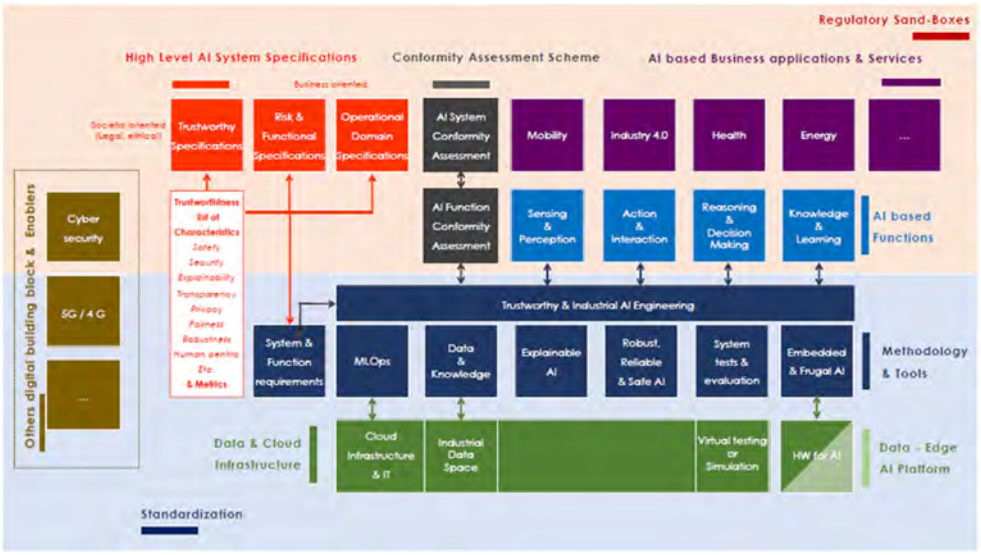
- **The availability of data, both in sufficient quantity and quality** to train and deploy AI models in production and monitor their "proper" functioning. There is therefore an issue of trust and completeness of the data related to the targeted application.
- **The evaluation of the return on investment (ROI) in AI proves to be complex** and represents an uncertainty when making decisions. AI is a "general purpose technology" and, as such, serves all industrial businesses, but does not necessarily constitute their core business.
- **The difficulty of integration with more global systems:** AI is a component or a function of a more complex system. **Its industrialisation requires specific engineering, but also interoperability with other engineering chains. Demonstrating compliance with functional and non-functional requirements or specifications** is also a blocking point. Trust is necessary to meet the challenge of corporate responsibility.

The conditions necessary for adoption by the industry

This observation is also shared by other analyses, such as the one conducted by Quantmetry in 2021 with 25 key companies in the sector. It states that **"future regulation is perceived as a necessity", to secure relations between manufacturers and guarantee a legal framework for liability.** It underlines the lack of organisational and strategic measures implemented to guarantee the application of future regulation.

The Franco-German paper *"speeding up industrial AI and trustworthiness"*[77], the result of a partnership between the French General Secretariat for Investment (SGPI) and the Big Data Value Association (BDVA), in collaboration with some fifty academic and industrial experts, also illustrates the **close link between trust and AI industrialisation. It highlights the industrial need for software solutions to implement trust in AI-based processes and applications, coupled with the availability of norms and standards**, in line with European initiatives on the cloud (Gaia-X).

[77] *Speeding up industrial AI and trustworthiness*", General Secretariat for Investment, Big Data Value Association et al., 2021.



Source: "Speeding up industrial AI and trustworthiness", General Secretariat for Investment, Big Data Value Association et al., 2021

The battle of AI design tools and methods

AI is a complex set of technologies for which skills are scarce, varied and the debate between insourcing and outsourcing is never-ending. As a result, in addition to the need for data and knowledge, **many industries today rely mainly on software tool platforms and development or "engineering" methods provided by specialised providers.** These tools and methods enable the development of AI algorithms from models, data and knowledge.

The artificial intelligence solutions segment is also driving the growth of the AI market as a whole, accompanied by a growing need among manufacturers for infrastructure and support to train and deploy models in production.

The digital giants are taking the lead with their integrated offerings, combining storage (cloud) and computing infrastructures, on which they largely dominate the market, as well as software solutions to manage data and support AI developments (whatever the typology of data and models). **However, trust requires an infrastructure, still to be developed, of specific tools and methods.** These must be interoperable with other design chains (MLOps[78], ModelOps and Systems), and allow risk analysis and design, validation, and control, including in operation, the system and its trust attributes for a given application.

Trustworthy AI will be largely developed on these tools and methods, which we group together under the term "Trustworthy AI InfraTech", a true "enabler" of the European political vision (carried by its regulation), of the trustworthy AI market (economic competitiveness) and of its dissemination to the whole economic framework, including SMEs and start-ups.

[78] a set of practices that aims at deploying and maintaining machine learning models in production in a reliable and efficient way.

2.2.3. An analysis of the trustworthy AI market in strategic industrial sectors for Europe

Market size for AI and trustworthy AI for nine high-risk industries

The EY-Parthenon study provides a "low" estimate of the true size of the addressable market for trustworthy AI based on nine industry segments analysed in a non-holistic way (more than 50 interviews conducted, plus an analysis based on the top 25 to 35 industry players per market): healthcare, automotive, aeronautics, rail, oil and gas, energy, banking and insurance.

It bases its analysis on the main current use cases by sector, while highlighting the emergence of future innovative use cases that will be accompanied by the rise in maturity of trustworthy AI and for which a market assessment is complex due to their strong impact on the transformation of industrial businesses.

MAIN USE CASES BY INDUSTRIAL SECTOR

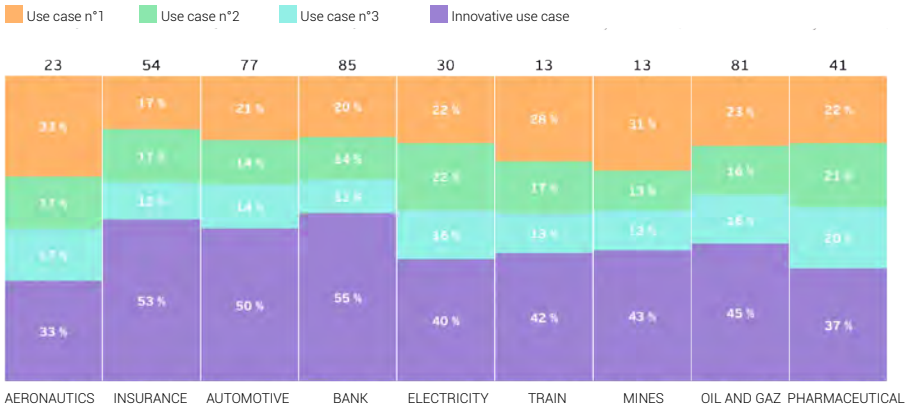
Industry/sector	Use case n°1	Use case n°2	Use case n°3	Innovative use case
AERONAUTICS	Predictive maintenance of aircraft and equipment	Quality control in the production process	Industry 4.0	Air traffic management (embedded AI and autonomous aircraft)
AUTOMOTIVE	Industry 4.0	Customer insight via embedded AI	Connected services (fleet management)	Autonomous vehicle
BANK	Cybersecurity	anti-money laundering	Fraud detection	Augmented advisor (credit granting)
ELECTRICITY	Infrastructure oversight	Market trading	predictive maintenance of equipment	Smart home and smart grid
INSURANCE	Fraud detection	Cash flow management and optimization	Cybersecurity	Augmented consulting (claims management)
MINES	logistics optimization	predictive maintenance of equipment	Smart mining	exploitation and detection of deposits
OIL AND GAZ	production control	drilling optimization	logistics optimization	exploitation and detection of deposits
PHARMACEUTICAL	optimization of molecule tests	optimization of molecule discovery	monitoring of sales performance	pharmaco vigilance
TRAIN	predictive maintenance of infrastructure	monitoring of infrastructure	traffic automation	management of new mobility offers and energy management

Source: EY Parthenon, SGPI, 2021

The analysis offers a relative weight of current use cases by industry sector, allowing the identification of early trustworthy AI applications. It represents on average between 40 and 60% of the AI budgets allocated by these sectors. However, one should emphasise that other use cases, whose transformative nature is not negligible or even decisive for these sectors,

are not evaluated in terms of the market. This potentially implies very strong growth, driven by increasingly innovative use cases, justifying the qualification of "low" for this market evaluation.

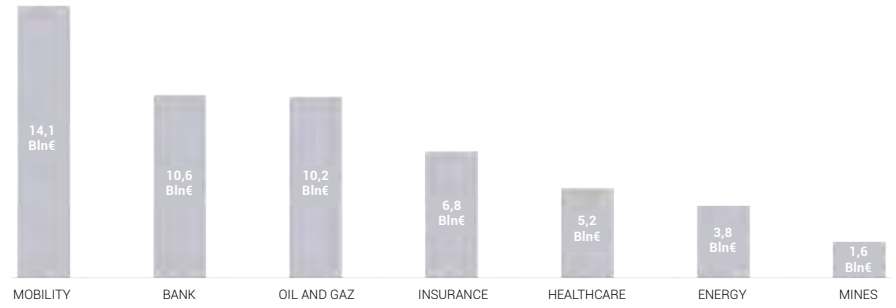
DISTRIBUTION OF AI BUDGETS BETWEEN THE MAIN USE CASES BY SECTOR



Source: Without trust, what future for Artificial Intelligence in industry?, November 2021, EY Parthenon, General Secretariat for Investment (SGPI).

The EY-Parthenon study therefore estimates an **AI market size of €80 billion for these nine industrial sectors, for a trustworthy AI market of €53 billion, again emphasising the importance of trust for the industrialisation of AI-based solutions.**

BUDGET ESTIMATE FOR TRUSTWORTHY AI BY SECTOR



Source: Without trust, what future for Artificial Intelligence in industry?, November 2021, EY Parthenon, General Secretariat for Investment (SGPI).

Market size for an "InfraTech" of trustworthy AI solutions

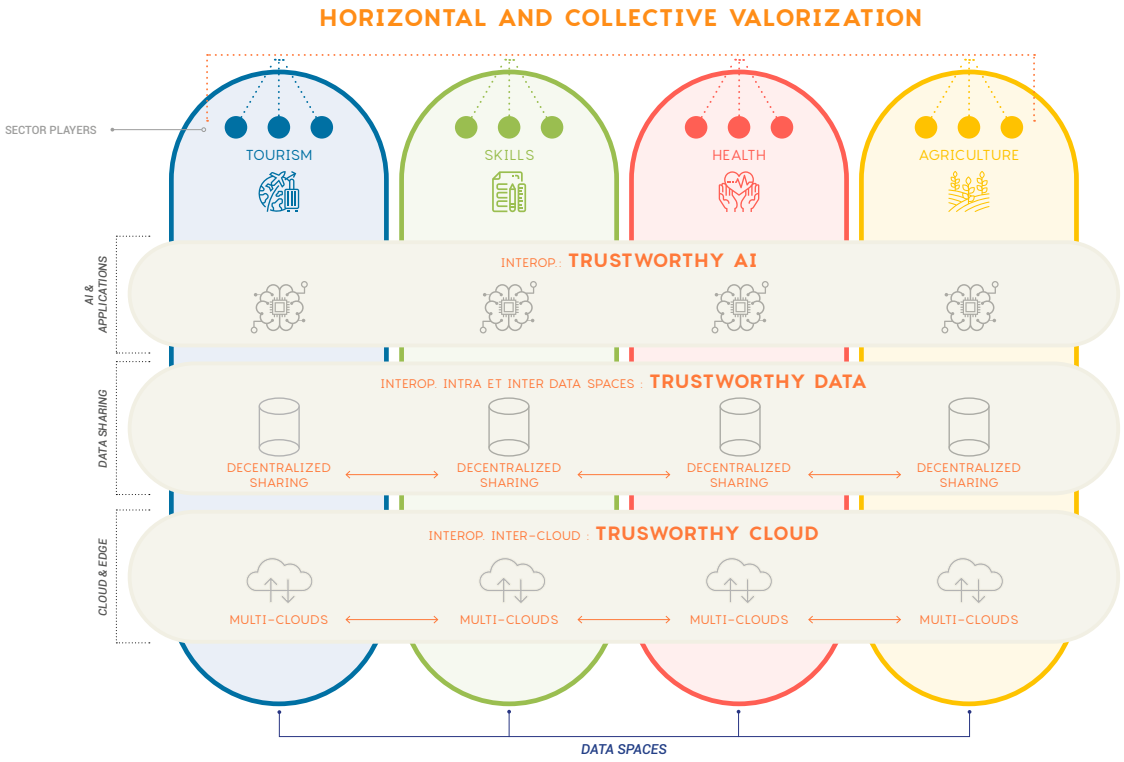
Regarding the ecosystem of trustworthy AI solutions, EY-Parthenon highlights a significant fragmentation between:

- **more generalist AI algorithm development platform providers** that are gradually integrating specific solutions;
- **more specialised technology providers for one of the trust attributes** (explainability, robustness, etc.);
- **providers of solutions** to evaluate the quality of software; and

AN INFRATECH THAT, BY LOWERING THE COSTS TO ENTRY IN TRUSTWORTHY AI, ALLOWS TO CONQUER AI MARKETS IN CASCADE.

- **auditors,** or even certifiers, of trustworthy AI-based systems.

The **"amount of investment in the purchase of such solutions amounts to approximately €420 million** divided between Europe (€180 million including €40 million in France), North America (€110 million) and Asia (€130 million). Among the most promising sectors are mobility (€113 million), banking (€85 million), insurance (€54 million) and oil and gas (€81 million). **Growth prospects are significant, driven by the marketing of future innovative products or services and the upcoming implementation of regulatory or normative frameworks**, as illustrated by the European Commission's ongoing work on the AI Act ».



We consider that the applications qualified as "high-risk" by the European Commission cover a much wider perimeter than just safety-critical systems[79]**, encompassing also applications considered to be business-critical**[80] (tourism, culture, food, etc.) and above all society-critical[81] (education, employment, etc.), implying in particular ethical considerations. This is all the more true since the European Parliament is encouraging the elaboration of a "code of conduct"[82] for all AI applications, including those that are not considered "high-risk". This would eventually imply that all AI-based products and services would constitute a market for the trustworthy AI InfraTech.

The United States has major players in the fields of high-risk critical systems (e.g. Boeing in aeronautics). However, AI developments are heavily "pulled and influenced" by **B2C**[83] **players such as GAFAM. Their culture is based on usage, an IT (Information Technology) and fail fast approach that rejects the fear of failure and values continuous experimentation.**

[79] System whose failure or malfunction can have dramatic consequences (deaths, serious injuries), major material damage, or serious consequences for the environment.
[80] System whose failure or malfunction can have a significant impact on the company, and more broadly the economy of a territory.
[81] System whose failure or malfunction can have a significant impact on the lives of individuals or on society as a whole.
[82] Proposal for a regulation of the European Parliament and of the Council laying down harmonised rules on artificial intelligence (AI legislation and amending certain Union legislative acts).
[84] Business to Consumer.
[84] https://www.economie.gouv.fr/intelligence-artificielle-au-service-des-entreprises-1

On the contrary, Europe can adopt an approach based on the provision of evidence, performance and not only process, in relation to an analysis of the risks and strong requirements that characterise critical or high-risk systems.

We recommend relying on the culture of safety-critical systems and OT (Operational Technology), **which is the real arrowhead for the development of systems engineering based on trustworthy AI,** and therefore for the ecosystem of trust as a whole. This observation is shared by industrial players from several sectors (energy, defence, health, etc.) and is reflected in the AI Manifesto[84] produced by a group of French industrialists.

TRUSTWORTHY AI, A TRUE BEACON OF OUR DIGITAL SOVEREIGNTY AND INDUSTRIAL COMPETITIVENESS.

III. BUILDING AN INDUSTRIAL STRATEGY THROUGH TRUSTWORTHY AI

SUMMARY OF PART 1 AND 2

The notion of trust in AI is central. It is at the heart of humanistic challenges and therefore of the concerns of a large number of public and private institutions. Starting with the citizens who wonder about technologies that will be deployed everywhere, but also in the world of industry, for which AI is a major challenge as well as a historic opportunity. Europe, despite its great industrial culture, has in recent years experienced a de-industrialisation phenomenon that AI can potentially help to catch up with.

In this industrial strategy of trust, Europe can rely on its globally-recognised legitimacy on the regulatory issues associated with technologies. In the wake of the regulations on digital technology and data (GDPR, DSA, DMA, DGA, DA) adopted by the European Commission, the future AI regulation (AI Act), based on a horizontal and risk-based approach, could become the first building block for a global trustworthy AI ecosystem.

ENJEU DE CETTE TROISIÈME PARTIE

However, to guarantee this dual industrial and digital sovereignty, **Europe must go beyond the framework of regulation alone, while at the same time deploying a proactive industrial strategy following a horizontal approach in all sectors. Strong coordination between Member States must make this trustworthy AI industrial strategy one of the pillars of the "digital single market" by making the entire value chain coherent,** from the definition of acceptable risks in line with our principles and values, to its industrial implementation.

3.1. AN OFFENSIVE STRATEGY THROUGH REGULATION

3.1.1. Making trust a real competitive advantage for Europeans

The main principles of the European regulation

In December 2019, Ursula Von der Leyen, newly elected President of the European Commission, announced in a speech to the European Parliament the desire to regulate AI within 100 days, with a view to framing the key sectors of the digital revolution. A year and a half later, the European Commission publishes its *White Paper on Artificial Intelligence*[85], which sets out the general view of the European executive. In April 2021 the Commission publishes its draft regulation establishing harmonised rules on AI, on which we base our work, albeit with reservation as it is still in progress.

The Commission proposes three structuring principles as foundations for the draft AI regulation:

- **A risk-based approach to AI systems,** with a typology ranging from systems posing an unacceptable risk to systems posing less or no risk.
- **The application of the regulation to actors located outside the European Union (extraterritoriality),** similar to the GDPR.

[85] https://ec.europa.eu/info/sites/default/files/commission-white-paper-artificial-intelligence-feb2020_fr.pdf

- **The neutrality of the text from a technological perspective,** which complements without replacing sectoral legislation.

The approach chosen by Europe is rooted in a more general method for companies and formalised by ISO in 2015. The ISO 9001:2015 standard codifies the risk-based approach on a global scale. **Risk analysis consists of prioritising risks according to the type of risk and its probability of occurrence.**

The risk-based approach has the advantage of prioritising things by differentiating the types of risks and is used to tailor the response more accurately. This differentiation of uses also makes it possible to better correspond to the needs of the market, the variety of cases and the expectations of citizens.

The European Union will have to be careful not to overload this complexity, especially since it must also find a consensus between 27 countries with their respective legislations. **One of the fundamental challenges is therefore the creation of a single and legible market for trustworthy AI at the European Union level.**

The expected positive and negative impacts of regulation

We need effective regulation. It must support innovation and economic development without hindering them. There are several elements to consider: first, it is imperative to **reduce the regulatory uncertainty** surrounding the AI market. On the other hand, **the upcoming AI regulation has the potential to become a *de facto* standard,** like the GDPR, in Europe and beyond, and we need to ensure that this is indeed the case. **Second, the risk-based approach is certainly complex, but it is compatible with the singularity of use cases and** experiences. Finally, **this regulation must be an incentive**, and must support players who want to use AI-based systems, including those deemed "high risk".

In its initial version of April 2021, **the draft AI regulation (AI Act) considers that the legal certainty put in place by the legislative act will be the main impact of the text:** *"Providers of AI systems should benefit from a minimal but clear set of requirements, creating legal certainty and ensuring access to the entire single market. Users of AI systems should benefit from legal certainty on the high-risk AI systems they purchase that ensures compliance with European laws and values."*[86]

In contrast, the Centre for Data Innovation[87] offers a critical study of the financial impact on small businesses. Indeed, **an SME that deploys an AI system deemed "high risk" will face a compliance cost that can, depending on the complexity and application, reach €400,000.**

The Commission has expressed the desire to see 75% of European businesses using AI-based technologies and solutions by 2030; but the Centre for Data Innovation study estimates the deterrent effect of compliance costs at almost 20% less investment. The study estimates that the resulting compliance burdens will cost European companies about €10.9 billion per year by 2025, and €31 billion within five years, **"not including lost opportunities and a likely brain drain as innovative start-ups find it easier to set up shop elsewhere"**[88].

For *France Digitale*[89]**, "the new regulatory burden should not discourage founders and investors in AI from engaging in Europe", including on "high-risk" topics.** The association deplores the lack of clarification on the typology of risks associated with AI systems, but also the difficulty in implementing these obligations for start-ups, especially those operating in risky areas (such as education or law).

[86] Artificial Intelligence Act, COM (2021) 206 Final 2021/0106, Brussels 21.4.2021, p.94
[87] https://datainnovation.org/
[88] La Législation sur l'Intelligence Artificielle coûterait à l'économie européenne 31 milliards d'euros sur 5 ans, réduirait l'investissement dans l'IA de près de 20 pourcents, d'après un nouveau rapport,Benjamin Mueller, 2 août 2021, datainnovation.org
[89] https://francedigitale.org/

Others see the future AI regulation as an opportunity to create new jobs, such as "AI risk officer" in structures with more than 50 employees using high-risk AI. Legal requirements will effectively increase the number of internal and external audits. On the other hand, **"a new role of quality engineer, specialised in AI, will probably emerge"**[90] to ensure that products and services comply with the requirements of the company and those of future AI regulation, following the example of the DPOs (Data Protection Officers) created by the GDPR.

Regarding the reception of the text by the legal community, Dalloz[91] insists on the complexity of the cross-cutting approach, while welcoming the choice: **"the desire to regulate the products and services embedding AI systems, rather than AI systems globally, is proving to be a real challenge in terms of articulating standards"**[92].

Stanford Law School - one of the pioneering universities on AI - emphatically and enthusiastically welcomes the European initiative: *"It takes courage and creativity to legislate in this stormy and interdisciplinary area, forcing American and Chinese companies to comply with standards based on European values before they can access this market of 450 million consumers. As a result, the proposal has an extraterritorial effect. By drafting the Artificial Intelligence Act and articulating it with humanistic standards and values in the architecture and infrastructure of our technology,* ***the European Union is paving the way and leading the world to a destination that makes sense.*** *The European Commission, with the GDPR, has already orchestrated the creation of an international standard for privacy, data protection and data sovereignty [...]."*[93]

In the face of the Stanford Law School's enthusiasm, two reservations can nevertheless be expressed about the choice of this legislative approach. On the one hand, **the European Union should not limit itself to strict regulation, as this could be seen as the "poor man's weapon"** to hinder the economic success of the American and Chinese giants. **"A second, more damaging, disadvantage for Europe is the risk of seeing the European Union imposing itself and its companies and citizens with binding virtuous rules without being followed by its main competitors, with the consequence of distortions of competition to its own detriment"**[94].

3.1.2. The AI Regulation as the first building block for an ambitious EU

An ambitious horizontal approach

The draft regulation on AI is part of the European Commission's strategic priorities for the 2019-2024 term, summarised in the axis "for more ambitious Union". In 2017, the European Council had already urged lawmakers to show "a sense of urgency in the face of emerging trends, in particular with regard to issues such as artificial intelligence"[95]. Two years later, the Council insisted on the importance of ensuring "full" respect for the rights of European citizens by calling for relevant legislation adapted to the challenges and opportunities offered by artificial intelligence.

Among the various options considered by the Commission, **the choice fell on the horizontal European legislative instrument, following a risk-proportionate approach,** as well as the encouragement to adopt codes of conduct on a voluntary basis for AI systems out of "high risk". In other words, **it is not the industrial or commercial sector, but rather the product or**

[90] "*AI risk manager, a new job for machine learning*", Antoine Crochet-Damais, 29 September 2021, journaldunet.com
[91] https://www.dalloz.fr/
[92] *Artificial Intelligence Act : joint opinion of the CEPD*, Cécile Crichton, 2 July 2021, Dalloz Acutalités
[93] *EU Artificial Intelligence Act: The European Approach to AI*, Mauritz Kop, Transatlantic Antitrust and IPR Developments (2021)
[94] Europe as an international normative power: state of play and perspectives, Laurent Cohen-Tanugi, *Revue Européenne du Droit n°3*
[95] Regulation of the European Parliament and of the Council laying down harmonised rules on artificial intelligence (Artificial Intelligence Act) and amending certain Union legislation, 21.04.21

service that will determine the applicability or otherwise of the AI regulation, notwithstanding the sectoral legislation that already exists.

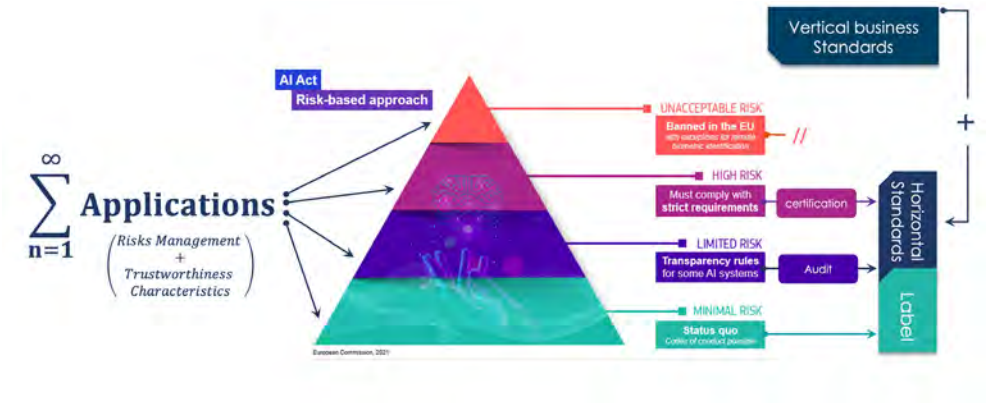
Strong requirements

The main principles of European regulation on AI were first laid out in 2020 in the White Paper *"Artificial Intelligence. A European approach based on excellence and trust"*[96] by the European Commission. The Commission proposes seven requirements for building trustworthy artificial intelligence:

- human agency and oversight;
- technical robustness and safety;
- privacy and data governance;
- societal and environmental wellbeing;
- diversity, non-discrimination and fairness;
- transparency;
- accountability.

These guiding principles and the technical requirements to be derived from them are integrated into a risk-based approach in the April 2021 draft AI regulation[97]. The risks on the technologies that make up AI are also qualified: opacity, complexity, unpredictability, and partially autonomous behaviour. **The legislator insists on the difficulties associated with the means of technical verification of compliance with the regulations.**

The risk pyramid



Source: Excellence and Trust in Artificial Intelligence, European Commission

The initial April 2021 version of the future regulation proposes a pyramid of risks[98] generated by AI systems:

- **Unacceptable risk:** uses of AI systems considered a clear threat to human safety or individual rights: prohibited. For example: systems that manipulate human behaviour and deprive users of their free will, and systems that enable social rating by states.
- **High-risk:** The draft AI Regulation published by the Commission considers as high risk AI systems that are safety components of products falling under harmonised sectoral legislation when they are already subject to a third party conformity assessment, as well as AI systems listed in its Annex III in various fields of activity, such as:

[96] COM (2020) 65 final, 19.2.2020
[97] *Proposal of a Regulation of the European Parliament and of the Council Laying Down Harmonised Rules on Artificial Intelligence (Artificial Intelligence Act) and Amending Certain Union Legislative Acts*, COM (2021) 206 Final 2021/0106, Brussels 21.4.2021
[98] "*Artificial Intelligence Act : The European Union invents the artificial intelligence risk pyramid*", Alexandra Bensamoun, Le Club des Juristes, 21 May 2021

- AI systems used as safety components in road traffic management or energy supply;
- AI systems used to determine access or assignment to educational or vocational training institutions, or to assess students;
- AI systems used to determine the recruitment, job evaluation, career development, and termination of of work-related contractual relationships;
- AI systems used to assess people's eligibility for welfare, creditworthiness, or to assign them a score for obtaining credit;
- Certain AI systems that can be used in the management of criminal investigations (polygraph, predictive systems in particular) or migration, asylum and border control (e.g. checking the authenticity of travel documents); and
- AI systems to help judges make decisions.

- **Limited risk:** AI systems with specific transparency obligations. Example: conversational robots (chatbots, deep fakes)
- **Minimal risk:** all other AI systems, for example those used in video games, or spam filters: no legislative intervention, but an incentive to develop a code of conduct.

The more likely the AI system is to create a risk (for the consumer, society, the State), the more rules will be needed, especially at the system design stage.

Some uses of so-called high-risk AI systems can be found in the following sectors: security, finance, banking, insurance, employment, education, healthcare, transportation, energy, public sector (asylum, migration, border control, judiciary, social security services).

So-called high-risk artificial intelligence systems are subject to specific requirements commensurate with the criticality of their uses.[99]

In addition, **the European legislator provides for the creation of a European Committee for AI ("the Board"), composed of representatives appointed by Member States and the Commission.**[100]

While welcoming this legislative initiative, and although the appointment of these representatives belongs to each Member State, the French CNIL **takes a stance on four points deemed fundamental** in an opinion published on July 8, 2021[101]:

- **the need to draw red lines in relation to future AI uses;**
- **the challenge of articulation with the GDPR;**
- **the importance of harmonised governance;**
- **an indispensable innovation support.**

If the fears of stifling innovation and experimentation are real, **the "Brussels effect"**[102] **can nevertheless use this regulation as a vehicle for its normative power at the international level** by making use of its capacities, as Laurent Cohen-Tanugi explains: *"the ability to draft its own law and impose compliance on its territory, or even beyond (extraterritoriality); the ability to influence the content of (legal, technical) standards resulting from an international negotiation process within various multilateral forums; and the ability to serve as a voluntary normative model within the international community".*[103]

[99] *COM (2020) 65 final, 19.2.2020, pages 21 à 26*
[100] *COM (2021) 206 Final 2021/0106, Brussels 21.4.2021, p.73*
[101] *"Artificial intelligence: the opinion of the CNIL and its counterparts on the future European regulation"*, 8 July 2021, cnil.fr,
[102] The Brussels effect is the process of unilateral regulatory globalisation caused by the de facto (but not necessarily de jure) externalisation of EU legislation beyond its borders through market mechanisms - encyclopedie.fr.
[103] *"Europe as an international normative power: state of play and perspectives"*, Laurent Cohen-Tanugi, Revue Européenne du Droit n°3, The paths of European power,

3.1.3. A voluntary approach through standards

"Use makes standards, and standards make use."

The importance of standardisation in AI

STRENGTHENING EUROPE'S INFLUENCE ON STANDARDS

On 2 February 2022, Europe presented a plan to strengthen its influence in the creation of international standards, concerning both electric batteries and digital services. This plan is part of the European industrial strategy presented in March 2020 in which Europe already showed a willingness to defend its companies and take its place in the Sino-American rivalry.

Some observers fear a loss of influence of European companies in the face of industrial giants, as well as the strategies of Beijing and Washington. But the issue is more profound: the Beijing regime is pushing for the redefinition of international standards thanks to the power of its digital players. In particular, Huawei's action to modify the Internet protocol with its "New IP Initiative" comes to mind.

It is therefore imperative for Europe to strengthen its influence in standards bodies to assert its interests and protect the values transcribed in international technological standards.

First of all, it is important to clarify **the difference between regulation and standardisation, which is sometimes a source of misunderstanding**. As the French Association for Standardisation (AFNOR) states, **"voluntary standardisation strives to detect and accompany technological and societal trends on an ongoing basis.** It understands the needs of the market in terms of openness, support for the competitiveness of companies and a harmonious development framework for activities and new jobs (...). It operates in a largely European and international framework, which it helps to build."

"Standardisation can also support legislation and regulation, and thereby contribute to limiting the inflation of legislative and regulatory texts. All this in a world of accelerating changes", **in particular through "harmonised standards"**[104]. **These are initiated by a mandate from the European Commission** to ensure that products and services comply with the technical requirements of the corresponding legislation. They therefore represent the meeting point between regulatory requirements and operational implementation by the market.

This choice, adopted by the European Commission in the field of AI, therefore gives standards a special importance both for the future European AI market, but also for the trust of European citizens.

Let's take the example of high-risk AI systems. They will be subject to regulatory requirements through a compliance review by notified bodies[105]**.** These mandatory requirements, including trust attributes and their evaluation after risk analysis, will be detailed through harmonised standards, on which the certification criteria will be based. For all these systems, which include a large part of industrial applications, the harmonised standards will constitute the technical requirements necessary to comply with the regulations.

[104] [Harmonised standards] are used to demonstrate that products or services comply with the technical requirements of the relevant European legislation. They are generally optional, but their application gives a strong presumption of conformity with the technical regulatory requirements. However, an organisation may choose not to apply them, preferring a technical solution other than that recommended in a harmonized standard.

[105] *"A notified body is an organisation designated by an EU Member State (or by other countries under specific agreements) to assess the conformity of certain products before they are placed on the market"*, European Commission, Notified bodies (europa.eu)

WE WANT TO BE A GLOBAL STANDARD-SETTER, NOT A STANDARD.

[103] *We want to be a global standard-setter, not a standard-taker*, Ursula von der Leyen

THE INSPIRING MODEL OF AERONAUTICS

Several high-risk sectors, including critical systems, are paving the way for integrating trustworthy AI into their industrial roadmap.

It often appears that the greater the level of risk and liability of economic actors, the more important industry cooperation is. **The ecosystem of trust for aeronautics, in general (including for non-AI systems), could serve as a model for other industries. Nevertheless, the cost of such a framework is particularly high,** and other economic actors may have a limited interest in proposing it for AI, if the risk does not require it. **It is up to the political world to help set up the ecosystems of trust in AI, in consultation with industry players. Differentiated modes of integration of the trustworthy AI ecosystem, depending on the level of risk, could be set up, in accordance with the specificities of each sector.**

Coordinated certification and regulation

In the aeronautics industry, the Federal Aviation Administration (FAA) is the authority for certification in the United States. The European Union Aviation Safety Agency (EASA) fulfils the same role. The definition of standards is done in consultation with the industry and the certification authorities. The two agencies separately propose the same certification standards for regulation by the American Congress for the FAA and the European Parliament for the EASA. The two agencies agree in advance, on a reciprocal basis, to coordinate the various texts. **This coordination allows a manufacturer certified in Europe to fly its aircraft in the United States.** Other global agencies are also involved in the process.

Coordinated standardisation

Building an aircraft relies on a wide variety of engineering disciplines. For example, outside of the AI field, DO-178C is used to approve software in an aircraft. This document is written jointly by the Radio Technical Commission for Aeronautics (RTCA) for the United States and the European Organisation for Civil Aviation Equipment (EUROCAE). It is used by the FAA and EASA as a basis for certification. **The FAA and EASA have confidence in the entire process.**

AI integration

In 2019, the EUROCAE working group WG-114 was set up in Europe to prepare for the future integration of AI in aircraft, with SAE G34 as the equivalent across the Atlantic. The objective of the joint working groups is to produce a uniform standardisation text that can be used by the EASA and the FAA, and at the global level. **By 2024 or 2025, aeronautics should have the first global industry standards for trustworthy AI**. AI is considered by the aeronautics sector as a classic standardisation and certification topic, it is not subject to specific treatment.

The difficult articulation between horizontal and sectoral standards

The debate between horizontal and sectoral (or business) standardisation is not new. Historically, the different sectors of activity have had little interaction: sector specificities prevail, and a silo approach is preferred by default.

Nevertheless, the following assumption about AI is made: ***"a cross-cutting technology can only be effectively regulated by horizontal rules that provide solutions to common challenges"*** (Thierry Breton, European Commissioner for the Internal Market, Industrial Policy, Tourism, Digital, Audiovisual, Defense and Space). The question of the articulation between horizontal and sectoral standardisation is becoming unavoidable, as is considering the problems of dominant positions of certain digital players, who have the competitive advantage of being multi-sectoral by nature.

In this context, cooperation between different industrial sectors and different types of players (large groups, SMEs, start-ups, research laboratories, universities, institutions) is becoming crucial, as well as the sharing of a unifying horizontal base for the standard as proposed in the French AI standardisation strategy (Afnor).

The effort will therefore have to be as much sectoral (health, aeronautics, education, etc.) to reflect the specificities and needs of the sectors and their use cases, **as it will be cross-sectoral** to facilitate the emergence of a complete ecosystem of trust (engineering tools, regulation, certification, control, investigation, etc.).

Several industrial sectors are already working to normalise or standardise their use of responsible, reliable, and safe AI. For example, the aeronautics sector started work with EUROCAE/SAE in 2019, and subsequently published an initial roadmap in 2021. In 2019, China's National Medical Products Administration published its guide to integrating AI in healthcare. In 2020, American insurers adopted the OECD's AI principles, and American banks began their formalisation discussions in early 2021. In October 2021, the White House Office of Science and Technology Policy issued a call to society to regulate uses of AI that are potentially harmful to society[106].

Table: standardisation bodies

	EUROPE	UNITED STATES	CHINA
Automotive	ACEA, EEA, national authorities	NHTSA (FMVSS), FTA, AV Comprehensive Plan, DOT, EPA	CCC (GB standards)
Railway	ERA, national authorities	RSIA, FRA, FTA	CRCC (GB standards)
Aeronautics	EASA, national authorities	FAA, EPA	CAAC
Banking	EBA, national authorities	FED, FDIC	CBRC
Insurance	AEAPP	NAIC	CIRC
Oil & gas	EUOAG, CEER, national authorities	PHMSA, EPA, FERC	NEA, SERC, CAEA
Mining	CEER	EPA	NEA, National Mining Law
Power & utilities	CEER, national authorities	PHMSA, EPA	NEA
Healthcare	AEM, national authorities	HSS, FDA	NMPA

At the same time, work is intensifying for European (CEN / CENELEC) and international (ISO, IEEE) standardisation bodies. Several countries, such as France and Germany, have also published strategies to respond to this challenge, with tight deadlines.

Since its beginning, however, Europe has been able to position itself in the standardisation bodies of major industrial sectors, such as aeronautics or health. **Given the transverse and therefore trans-sectoral nature of artificial intelligence, we must give ourselves the means to influence the relevant standards bodies** and avoid letting other countries (China, United States) or the major digital players decide on standards alone. On this subject, a report to be published

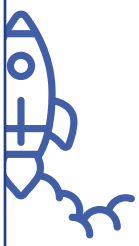
[106] *"Without trust, what future for Artificial Intelligence in industry?"* EY-Parthenon, p.7.

by the *Ecole de Guerre Économique* (EGE) focuses on the strategy of influence in the field of standardisation, with a particular focus on AI.

We therefore recommend supporting the European Commission's policy of making standardisation a priority for industrial and digital sovereignty, as well as the implementation of a European label for applications not categorised as "high-risk".

THE FRENCH AI STANDARDISATION STRATEGY

at the initiative of the 'Grand Défi de confiance' as part of the National Strategy on AI



The French State has mandated AFNOR, as part of the *Grand Défi* on trustworthy AI, with **a mission: "to create the normative environment essential to the deployment of trustworthy AI".**

Following a survey conducted in the summer of 2021 with more than 250 actors from the research industry and civil society, AFNOR published the national roadmap in March 2022 to provide this strategic sector with new voluntary standards and to disseminate best practices.

It is based on **6 priorities:**

- **Standards for trust:** the priority characteristics to be standardised are security, safety, explicability, robustness, transparency, and equity (including non-discrimination). Each characteristic will have to be defined, with a description of the concept, the technical requirements and the associated metrics and controls.
- **Standards on governance and management of AI:** AI generates new applications, all of which carry risks. These risks are of various origins: poor data quality, poor design, poor qualification, etc. A risk analysis for AI-based systems is therefore essential, to propose a risk management system.
- **Standards on the supervision and reporting of AI systems:** the aim is to ensure that AI systems are controllable (ability to regain control), that humans will be able to regain control at critical moments when the AI leaves its nominal operating domain.
- **Standards on the competencies of certification bodies:** it will be up to these bodies to ensure not only that companies have put in place processes for the development and qualification of AI systems, but also that the products comply with the requirements, particularly regulatory ones.
- **Standardisation of certain digital tools:** one of the challenges of AI is to have simulations based on synthetic data, and no longer on real data. Standards should make this data reliable.
- **Simple access to and use of standards:** to keep this strategy alive and to adjust it along the way, a consultation platform will be made available to animate the French ecosystem.

3.1.4. Ensuring the balance between regulation, standards, and innovation through "sandboxes"

Defining the legal framework for regulation: a challenge for public authorities and regulators

First, adopting a definition for regulatory purposes is complex. AI is a subset of rich and evolving disciplines (see Chapter 1). However, in law, it is necessary to define clear, legible limits of application that respect the principle of equality before the law. The current negotiation (cf. appendices I, II and III) seems to oscillate between, on the one hand, an extensive definition that is almost "diplomatic" in nature because it integrates all the sensitivities of AI, and, on the other hand, requests to reduce the scope of this definition on the grounds of not imposing obligations on existing software, which has sometimes been in existence for many years and for which the current legal framework is satisfactory.

Second, there is no assurance that the unified set of guiding principles will remain relevant over time. As an example, the first version of the text, with respect to data-driven AI, was primarily indexed to the supervised AI paradigm. At the same time, the giant self-supervised model paradigm was taking off and had its first commercial applications. Since then, various positions have recommended a better balance of responsibilities, particularly regarding the development and availability of generic models qualified as "general purpose". However, this raises questions about the substance of the issue: how can classes of models with characteristics and properties that are not yet well known be properly regulated?

For a text in progress, this type of amendment does not pose any procedural difficulties. However, it does raise questions about substance: how can we appropriately frame classes of models whose characteristics and properties are not well known? **On the other hand, once the text is published, what will happen when new substantial evolutions of AI occur?**

Regulatory sandboxes or the necessary legal space for innovation

The deadlines for negotiating, adopting or amending European regulations are such that it is necessary to anticipate any future substantial legal inadequacy. The use of delegated acts, provided for in the AI Act, can introduce flexibility, but in a limited way. Therefore, the **organisation of legal spaces adapted to technological innovation is relevant in order to reconcile regulation and innovation.**

Regulatory innovation sandboxes are the best possible embodiment of this. According to the OECD, they have five main characteristics:

- **The demonstration of the truly innovative nature of the good or service wishing to integrate the program,** whether it is a real-world evaluation of new technologies or an evaluation of new uses of existing technologies;
- **The identification of an economic or social interest** (improvement of the functionalities of the goods and services, reinforcement of quality, price reduction, etc);
- **The determination of temporal, geographical or sectoral limits** that avoid weakening the scope of a regulation by opening a permanent and general exemption; and
- **Safeguard mechanisms,** which may relate to consumer protection, security or personal data protection, etc. in order to avoid any systemic impact, this involves imposing conditions relating to the level of diffusion of the good or service.

- **The justification of the inadequacy of certain elements of the current regulations (ex: GDPR),** which may hinder the evaluation in real conditions;

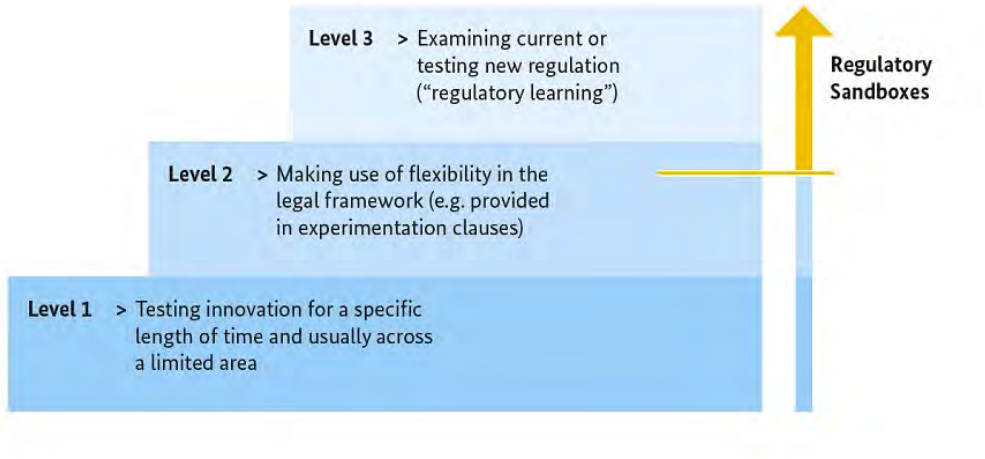
The proposed AI Regulation severely limits the value of regulatory sandboxes

However, in its initial version, the AI Regulation (AI Act) only offers a particularly conservative and restrictive version of regulatory sandboxes almost stripping them of their substance and thereby strongly limiting their impact:

- **Tests in real conditions, even if controlled, are proscribed** in favour of the controlled environment only;
- **The sandbox can only relate to demonstrators** and not to goods or services deployed in real conditions; and
- **Time limits are highly constrained and any justified and reasoned temporary regulatory relaxation is proscribed.**

This choice can be justified by the preservation of the "protective" model of Europe. Nevertheless, it means depriving ourselves of mechanisms that, on the contrary, by their agility, help to guarantee it.

Regulatory sandboxes have many advantages: they help regulators to refine their doctrine, they give public authorities time to adjust the texts, they support the rise in competence of the audit and compliance ecosystem, they develop new, specific, and more adapted voluntary standards, and they guarantee development in compliance with regulations.



Source: *German Federal ministry for economic affairs and climate action*

We recommend promoting regulatory sandboxes according to the criteria promulgated by the OECD so as not to deprive ourselves of a mechanism which, by virtue of its agility, contrary to popular belief, helps to guarantee the future implementation of the regulation, and therefore the protection of European citizens.

3.2. AN INDUSTRIAL STRATEGY THROUGH COOPERATION

3.2.1. Increase RDI and training efforts in trustworthy AI

The EU is well-positioned in research, but falls behind in innovation

The European Union has high-quality fundamental research in AI with renowned universities and scientists ranked among the best in the world, but falls significantly behind in the race for innovation, as well as in supporting start-ups. This is in a context where AI is proving to be one of the areas of research that is attracting the most interest from the scientific community (cf. 2021 edition of the UNESCO State of Science report[107]).

The United States, on the other hand, is quicker to turn research knowledge into innovation and the creation of companies specialised in AI.

An analysis of the statistics for the period between 2010 and 2021, taken from the AI Index Report 2022[108], is enlightening.

- **Publications:** Europe and the UK, (approx. 19%) rank ahead of the US (approx. 14%) and behind China (approx. 31%) in terms of the number of publications in scientific journals. At the same time, its scientific influence, calculated by an indicator of quotations of articles from the most prestigious scientific journals, illustrates a notable decline over the period between 2010 and 2021 (21% versus 26%). However, this Brevets trend is even more pronounced for the United States (25% in 2010 versus 17.5% in 2021); China, on the contrary, is benefiting from a significant increase from 21% to 28%.
- **Patents:** In a context where the number of patents filed has grown exponentially since 2015, the United States remains largely in the lead in the race for patents granted (39.5% in 2021); even so with a notable decline since 2010 (65%). Europe and the UK are far behind (7.5% in 2021 against 11% in 2010) followed by China (6% in 2021 against less than 1% in 2010). If this trend continues, Europe and the UK will fall into third position.

Notable fact: The United States also largely dominates open source software libraries in AI. The French library Scikit-Learn developed by Inria Paris-Saclay teams now ranks 5th, after having occupied 2nd place for a long time.

This overview of the continuum between research and innovation is complemented by a European Investment Bank (EIB)[109] report published in 2021 on AI and Blockchain. In 2019, the United States accounted for about $20 billion of investment, or approximately 65% of the global volume, compared to $5 billion for China and $2 billion for Europe. At the same time, the United States had an undeniable lead in the number of specialised companies, more than twice as many as China in second place. The European Union is home to "only" three times as many small businesses as the United Kingdom alone, even though it comprises 27 countries.

RDI in trustworthy AI is getting organised, and remains a real opportunity for Europe

In recent years, many initiatives have been launched around the world to support RDI on trustworthy, responsible, ethical and safe AI.

The following non-exhaustive list contains examples of such initiatives:

- In the **United State**s, the Defense Advanced Research Projects Agency (DARPA) has initiated programs on the explainability of AI - one of the attributes of trust needed for human-

[107] UNESCO Science Report: the Race Against Time for Smarter Development, 2021
[108] Artificial Intelligence Index Report 2021, Stanford University
[109] *Artificial intelligence, blockchain and the future of Europe: How disruptive technologies create opportunities for a green and digital economy*, European Commission/European Investment Bank.

CREATE AN INFRATECH MIDDLE LAYER OF TRUSTWORTHY AI, TO CONTAIN AND ABOVE ALL TO CONQUER THE BIG TECH MARKET.

machine interaction - or on the autonomy of AI-based systems, including conformance to a job domain.

- At the same time, **Canada** mobilised the Montreal Declaration and programs like DEEL (Dependable and Explainable Learning) in partnership with France.
- In 2018, as part of its national AI strategy and its 2nd phase launched in 2021, France also decided to support research and innovation in trustworthy AI around an industrial research program such as *"the great challenge on the security, reliability and certification of AI-based systems, or the great challenge on trustworthy AI for industry"*, (the content of which we will discuss later in this report) or the Interdisciplinary Institute in AI (3IA) ANITI based in Toulouse.
- In **Germany**, the Fraunhofer Institute for Technology Research IAIS[110] is leading a KI-Zertifizierung program on the certification of AI-based systems. The DFKI[111], for example, is conducting research within the framework of the "Certlab". The standardisation institutes (DIN[112] and VDE[113]) have finally published a roadmap for standardisation.
- At the **European level**, the Etami initiative supports the development of ethical AI, the TAILOR European network of excellence brings together more than fifty partners around research on the foundations of trustworthy AI or the future Testing and Experimentation Facilities (TEF) ambition, as part of Digital Europe, to bring out reference sites to test and experiment AI solutions, including their compliance with future European regulations.

With nearly 75% of publications at the FACCT (fairness, accountability and transparency) conference in 2021, **North America seems to be the most dynamic in research** compared to Europe and Central Asia (17% in 2021) and Asia Pacific (less than 5%), especially in relation to its 3rd position (mentioned above) in terms of number of publications in scientific journals.

AI regulation makes sense to protect European values and principles. However, the risk of increasing our backwardness in terms of innovation cannot be excluded. A balance between regulation and innovation is therefore essential. To do so, **we recommend increasing European investment efforts in fundamental and applied research, training (especially in the fields of data and knowledge engineering, algorithmic engineering) and innovation on trustworthy AI to support Europe's political vision in AI.**

3.2.2. Creating a Trustworthy AI InfraTech

"Creating a trustworthy InfraTech middle layer in AI, to stem and above all conquer the Big Tech market"

The "Trustworthy Digital" market is in its infancy. The "Trustworthy Cloud" strategy currently consists mainly of legally framing contractual relationships with hyperscalers to limit the extraterritoriality of American law. This defensive strategy, sometimes decried by digital sovereignty activists, is a pragmatic approach that consists of favouring the multi-cloud approach as a commercial opportunity to penetrate the market, until European offers are perceived as fully competitive by our companies.

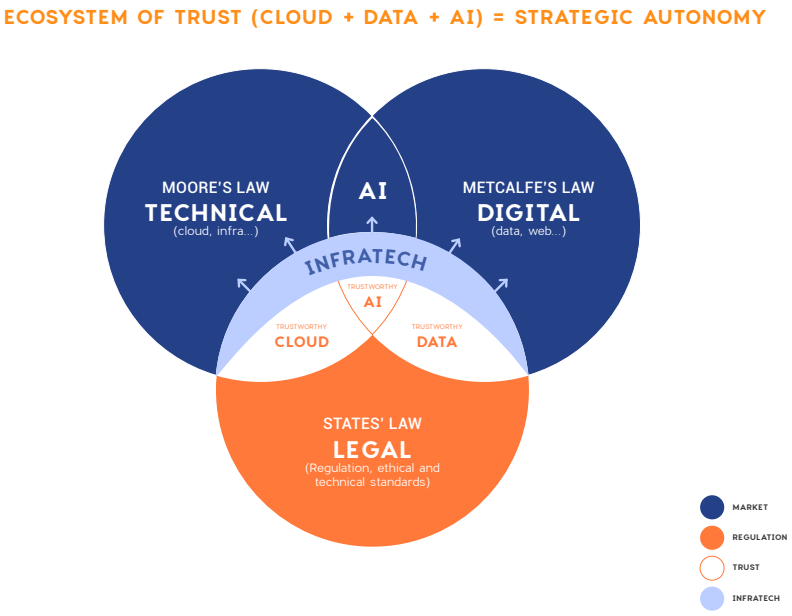
For the "Trustworthy Cloud" strategy to be successful, we need to leverage its two corollaries, namely "Trustworthy Data" and "Trustworthy AI", which we believe can create a market gap through regulation and cooperation. This is the purpose of the dedicated Digital New Deal notes, whose vocation is to make "digital trust" an offensive industrial strategy against the

[110] Institute for Intelligent Analysis and Information Systems
[111] German Research Centre for Artificial Intelligence.
[112] German Institute for Standardisation.
[113] German Association of Electrical, Electronic and Information Engineering Industries.

current oligopolies (having monopolised the market via their packaged cloud offers[114]), and stop it from being a defensive approach.

The technological and commercial lead of the hyperscalers is such, and their financial capacity so great, that we have no choice but to initially "contain" their expansionism by creating an intermediary layer between the cloud and the applications, and then use this "smart middleware" or "InfraTech" as a basis for conquering the market. Through this "containment" logic, we could then "re-intermediate" the structure of sovereign technology offerings with market demand. Evidently, all of this requires coordination and, above all, the pooling of investments in order to benefit from a common and competitive base of tools and services capable of capturing socio-economic value. **This is what we call InfraTech, which is the glue that binds together Cloud, Data and AI.** Without this infrastructure investment, we are condemning ourselves like the Danaids to "throwing money in the sand"[115], to use Bruno Le Maire's (French Minister of the Economy and Finance) expression.

For a trustworthy AI InfraTech



As we explained earlier, AI is a complex set of technologies for which skills are scarce, varied, and the debate between insourcing and outsourcing is never-ending. As a result, in addition to the need for data and knowledge (e.g. from data spaces), **many industries today rely primarily on software tool platforms and development or "engineering" methods provided by specialised providers**. These tools and methods enable scalability at lower cost for start-ups, SMEs and large corporations wishing to develop AI-based products and services.

However, **trust requires a set of specific tools and methods, or InfraTech, to be deployed, and reduce the cost of compliance that can rise up to €400,000 depending on the complexity and application, which is an economic dead end for start-ups and SMEs.**

This InfraTech must be interoperable with other design chains (MLOps[116], ModelOps and Systems), and enable risk analysis and design, validation, and control, including in operation,

[114] Antitrust: OVH has filed a complaint against Microsoft with the European Commission, Siècle Digital, April 2022.
[115] Statement by the Minister of Economy and Finance, on the government's efforts in favour of innovation, in Paris on 19 November 2019.
[116] MLOps or ML Ops is a set of practices that aims to deploy and maintain machine learning models in production in a reliable and efficient manner.

CREATING THE INFRATECH, THE FOUNDATION OF TRUST, AND A SCALABILITY LEVER FOR THE EUROPEAN INDUSTRY.

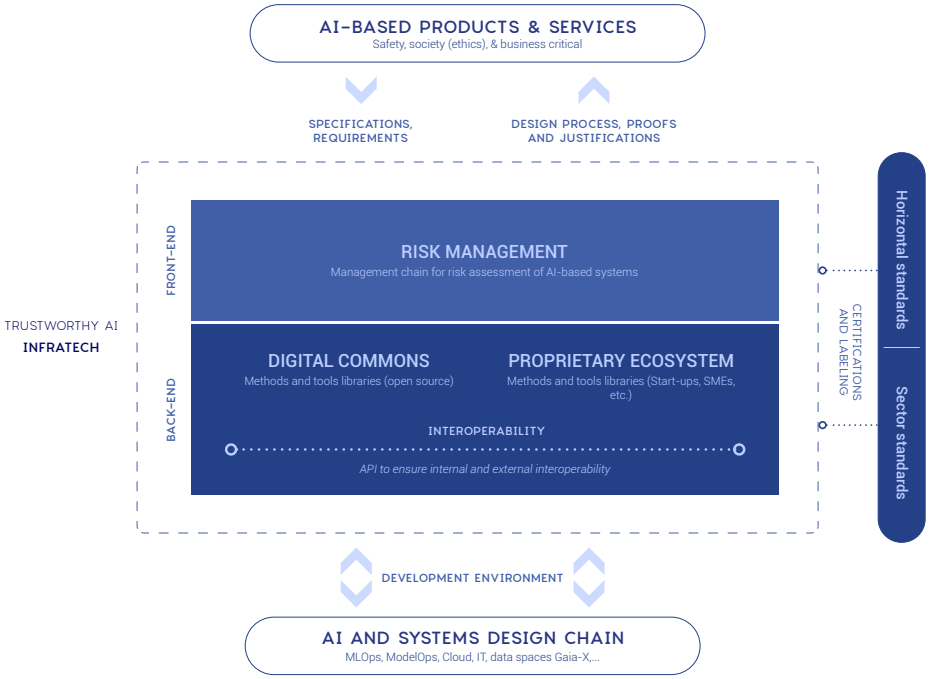
of the system and its trust attributes for a given application. Trustworthy Artificial Intelligence will be largely developed on these tools and methods, which will become the basic building blocks of the Trustworthy AI InfraTech.

This InfraTech is based on an approach to the entire "data, knowledge, algorithm and system" value chain, consistent with and complementary to existing initiatives at the European level (Gaia-X, etc.), and must therefore focus on solving the challenges of each of these components, by addressing:

- **Methods and processes** (with respect to functional and non-functional specifications and requirements of components and systems): this is a prerequisite to trustworthy AI-based application developments. This action must therefore be carried out in a preliminary phase, with an assessment of the risks specific to the product or service. They also make it possible to carry a built vision towards horizontal and vertical norms and standards and therefore to ensure compliance.
- **Environments for the constitution of qualified databases and knowledge:** the function performed by the system is very strongly dependent on the learning database and "business" knowledge. **It is therefore essential to have a "data and knowledge engineering"** environment, producing qualified databases according to the requirements and the field of use. This approach must be coupled with technologies for validation and verification (non-exhaustive list) of biases, annotation qualification, representativeness in relation to the fields of use or detection of attacks by poisoning, etc.
- **Environments for the design, validation, characterisation and verification of AI components and systems based on these same components:** on the one hand through "algorithmic engineering" tools in order to accompany the design and system integration approach, on the other hand through tools for validation and verification of properties (models, learning, dynamics) such as robustness or numerical stability[117] (non-exhaustive list: detection of false positives, data alteration, adversarial networks, satisfiability, uncertainty regime, etc.). The design of trust models by construction or "design" must also be taken into account.

[117] Attackers use an instability of the algorithm resulting in a misclassification or misperception of the image to be identified (e.g., a "50 km/h limit" sign instead of a "stop" sign). Thus, tools for validating and verifying stability properties are of major importance in the field of AI cybersecurity.

The diagram below describes its main components, as well as its interfaces with standards, other design chains, and applications:



If Europe wants to develop its own trustworthy AI solutions and be competitive, engineering, inspection and evaluation tools and methods will be a decisive element of the trust value chain, which must be mastered, as well as an intrinsic vector for the dissemination of European values.

THE CONFIANCE.AI PROGRAM, INITIATED BY *'GRAND DÉFI DE CONFIANCE'* OF THE FRENCH NATIONAL AI STRATEGY, AIMS TO DESIGN A DEVELOPMENT ENVIRONMENT FOR TRUSTWORTHY AI

The confiance.ai program, led by IRT SystemX, brings together academic and industrial partners (Air Liquide, Airbus, Atos, CEA, Inria, Naval Group, Renault, Safran, IRT Saint-Exupéry, IRT SystemX, Sopra Steria, Thales, Valeo), all of whom are united by the same ambition: **to design a development environment for trustworthy AI.**

This environment will be made up of technological bricks that are interoperable with the industrial partners' own engineering workshops, thereby making it possible to create a tool chain, methods and best practice guides that meet the functionalities required to design, validate, deploy and maintain these systems in all sectors, while taking into account regulatory and normative constraints.

The realisation of AI-based critical systems requires revisiting and enriching traditional engineering (data and knowledge engineering, algorithmic engineering, software engineering and system engineering). This requires ensuring the conformity of the system to the customer's needs and constraints, defining methods and tools to secure all of the design phases, but also guaranteeing trust properties (see the attributes described in chapter 1.2) throughout the life cycle.

The confiance.ai program is structured around seven independent projects and currently brings together more than 40 partners (large corporations, start-ups, SMEs and research laboratories in equal proportions).

We recommend funding the creation of a European software platform of tools and methods for the development of trustworthy AI (InfraTech), and supporting the ecosystem of solution providers. The trustworthy AI InfraTech will have to be developed in addition to the one being launched in data sharing ("Trustworthy Data InfraTech or Smart Middleware for Data-Sharing") associated with the development of Gaia-X, allowing compliance with future European regulations, as well as with future associated norms and standards.

To accelerate developments, amplify the dynamics of cooperation and promote the emergence of a single digital market in Europe, **this project must be based on our industrial "strengths", in particular by drawing developments from their use cases, and on initiatives already underway.**

Developing the "Digital Commons" of trustworthy AI

A **"digital commons" is a digital asset owned by a community and whose use can be free. This commons can be composed of infrastructure, data, libraries, software, tools, methods, etc., governed and protected by the community.**

If Europe wishes to invest in InfraTech, there are two possible approaches:

1. **Create a single European champion of trustworthy infrastructure** capable of competing with GAFAM or
2. **Create an ecosystem for trustworthy data/AI infrastructure (the InfraTech)** with the help of multiple public and private actors, including European start-ups.

The single European champion approach has already been tried on other technological subjects, without convincing success. The resources available to the American and Chinese giants and funds are considerable, and it is currently difficult to compete head-on. Recent confrontations between American and Chinese public institutions and their digital giants also show that this approach raises other issues of sovereignty for the states concerned.

THE DIGITAL COMMONS

The digital commons are part of a more general reflection on the commons. At the crossroads of economic theory and political science, the theme was popularised by the work of Elinor Ostrom in 1990 and her Nobel Prize in Economics in 2009. Initially, the commons refer to physical resources (river, forest, fishery) that can be administered collectively. These commons are characterised by a set of rights and obligations over a given resource, and by a governance system that is as close as possible to the singularity of the case. The challenge is to protect the resource and, if possible, to work towards its development.

The digital commons are unique in that the cost of copying and distribution is close to zero. They are for example source codes, data or databases, software or digital content (image/video/sound). In addition, there are free licenses, which legally organise the access, use, production, modification, distribution and management of these digital resources.

The digital commons are distinguished from the physical commons by two main characteristics: they are non-exclusive (no limit of access if someone uses the resource), and they are non-rival (the use of the resource does not deprive other users, it remains available).

Sources: Governing the Commons. The Evolution of Institutions for Collective Action, Political Economy of Institutions and Decisions, Elinor Ostrom, Cambridge University Press, 1990
"The Commons, a political breach in the digital age", Valérie Peugeot, Les Débats du Numérique, pp.77-78, Presses des Mines, 2013
"Beards on the Internet Prairie: Against the New Enclosures, the Digital Commons as Levers of Sovereignty", Benjamin Pajot, note from the Ministry of Foreign Affairs, August 2020,

We recommend adopting a Digital Commons strategy to mutualise the costs of InfraTech development, while preserving sovereignty through the multitude of actors who will rely on these commons to build a coherent "packaged" offer. Europe could initiate the process, with aligned third countries, and focus on guiding developments in line with its values.

3.2.3. Drive adoption via the European data spaces and industry use cases

For the development of trustworthy AI in European data spaces

In its data strategy, the EU promotes the emergence of the concept of data spaces[118] **to facilitate the circulation and sharing of data,** based on common standards[119], through a data-sharing infrastructure. Ten priority sectors have been announced to date, including mobility, finance, health, skills, energy, the Green Deal, and administration. **The Commission estimates**

[118] A data space gathers public and private actors willing to share their data (personal and non-personal), through a decentralised infrastructure and a common governance.
[119] Trusted data, sharing data, key to our strategic autonomy - Digital New Deal.

that the global market generated by this data flow will amount to €530 billion per year.

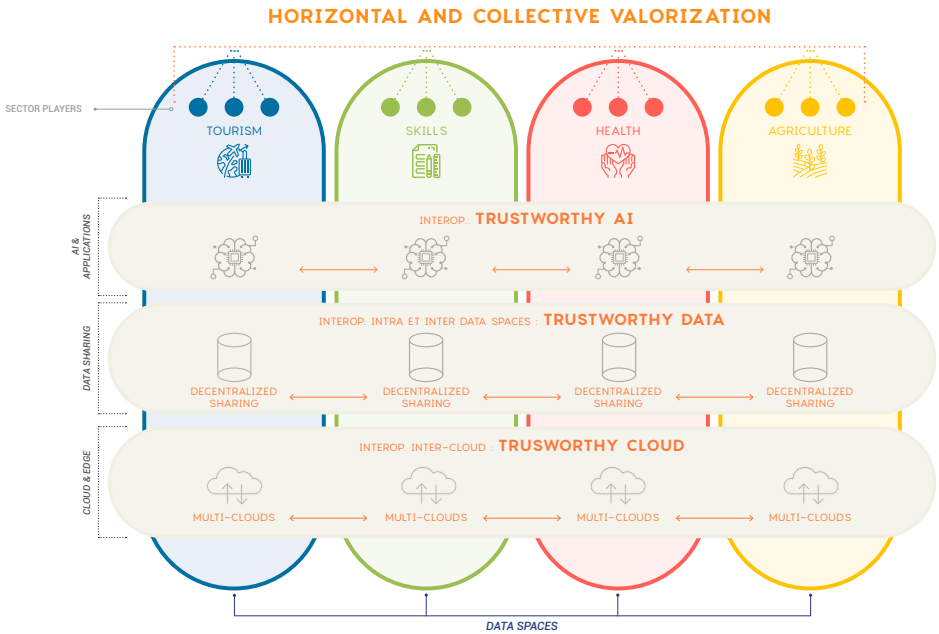
In all sectors, AI will be at the heart of the operation of data spaces for data exploitation. As the fundamental promise of the data space is to foster trust between actors, trustworthy AI is not a possibility but a necessity. This positioning builds on the EU's 2019-2024 mandate goals[120].

"There will be no trustworthy data space without trustworthy AI"

The pooling of data in these data spaces will provide an interoperable environment and a trustworthy InfraTech for sharing data ("Smart Middleware for Data-Sharing" or Trustworthy Data InfraTech), and will exploit them for the realisation of products and services (Trustworthy AI InfraTech), with a set of legislative, administrative, technical, and contractual rules determining the rights of access and use of the data. Green Deal, Health, Agriculture, Education, Mobility, are examples of common European data spaces under construction[121].

By making data available on a massive scale, the European Commission is trying to impose its standards and values. Indeed, anyone wishing to access the data of these data spaces will have to adhere to their governance and respect their standards. The infrastructure will ensure the sovereignty of organisations and individuals over their data and will avoid the "vendor lock-in"[122] of Big Tech, thanks to interoperability standards.

The data spaces, by their objectives, their traction and their values, constitute an excellent field of diffusion for all the actors of trustworthy AI ("level playing field"). They will determine the development of AI systems thanks to the mass of data collected, the business knowledge of industrialists, and will influence the governance of the infrastructure and the respect of European rules by external players. **Data spaces will be a strategic way to ensure that trustworthy AI systems are the source of these developments and spread.**



[120] European Union Priorities 2019-2024

[121] Information session on a preparatory action for the common European Green Deal Data Space under the Digital Europe Programme (DIGITAL), Shaping Europe's digital future, 15 December 2021,

[122] Proprietary lock-in is the creation of an intentionally non-standard feature in the machine, or software sold, that prevents the customer from using it with products from another vendor, and also prevents them from modifying it or accessing the features of their machine to modify it. For example, many software vendors use formats that can only work with their software. These strategies lead to the creation of monopolies.

SOCIETY CRITICAL: THE EXAMPLE OF THE "EDUCATION & SKILLS" DATA SPACE

This European data space "Education & Skills", known as Prometheus-X, aims to provide actors in the field with an infrastructure that will allow them to share and exchange data as well as develop solutions, particularly AI-based ones.
By having access to all of an individual's exercises, grades, interests, training and professional experience, an AI service will be able to analyse his or her strengths and weaknesses and recommend personalised choices in line with the opportunities and needs of the territory. This is also beneficial for organisations that could recruit more easily.

However, this approach carries considerable risks - in addition to the security of personal data - particularly associated with the determinism induced by algorithms (systemic bias) **on individuals.**
Trustworthy AI is the only way to address both the need to use AI to exploit a large amount of data, and the need to frame its use by common rules. These general rules must on the one hand be specific to the education and skills sector by including the business principles of the domain (guidance for example in recommendation systems), and on the other hand be enacted by the person concerned himself **(I must be able to decide what I want or not, and AI will increase the precision and strength of this choice).**

A global approach to the trustworthy data and AI markets

The challenges of a society critical data space such as education demonstrate the need to **"consider these high standards as strategic opportunities, and even as differentiating elements, in the global race for artificial intelligence"**[123]**.** The advent of these high and secure standards necessarily requires regulation that is both bold and secure.

Data spaces, which relate more to business or society critical issues than safety critical, are still under construction. They are therefore not very mature when it comes to AI and trustworthy AI. On the other hand, they should eventually represent a very important market for trustworthy AI, possibly larger than the market for safety critical systems.

From the beginning, data spaces are based on trust between actors to facilitate data sharing. Trustworthy AI appears as a natural next step for their development. Thanks to the progress made in the safety critical domain, where risks are the greatest and where actors are acculturated to take them into account, **Europe will be able to open multiple new markets for trustworthy AI, relying on its future data spaces which represent its trustworthy data approach.**

Promote the development of application cases (systems, products or services) based on trustworthy AI

To disseminate best practices and create value in European industrial sectors, **we recommend supporting all sectors through the realisation of systems, products or services based on trustworthy AI.**

This approach allows us:

- **to foster the adoption of trustworthy AI InfraTech among users whether they are industrialists, auditing or certification bodies.** Support for the development of skills in the sector, particularly in SMEs and start-ups, is essential for the dissemination of practices. It must

[123] Villani Report p.28.

therefore be offered through development, qualification and testing hubs in trustworthy AI. This effectively connects the ecosystem of trustworthy AI solution providers and the ecosystem of developers of future products and services.

- **to demonstrate the integration of all hardware and software technologies** (electronics, connectivity, AI) **in a single complex function or system,** relying on the interoperability of the different tool chains.
- **to structure the industrial ecosystem i.e. all the players in the value chain, around the development of complex systems based on trustworthy AI.** Promoting the cooperation of technology suppliers, equipment manufacturers, system integrators, service operators, etc.
- **to validate the maintainability, but also the desired evolutions of the systems,** whether over time (e.g. linked to degradation, conditions of use) or through updates, for example over the air.
- **to demonstrate the trustworthy properties** of the assembly in operation and at scale, thus the compliance with regulations, norms and standards and this throughout the life cycle and use.
- **to validate the adequacy with market needs** and uses.

3.2.4. Shared governance for digital trust

The heart of an industrial alliance for the European AI ecosystem of trust

We recommend the creation of an alliance for trustworthy AI in Europe around strategic industry sectors, based on the ecosystems already initiated in the Member States.

THE EXAMPLE OF THE 'GRAND DÉFI' AI IN FRANCE

France's investment in trustworthy AI is led by the SGPI (General Secretariat for Investment) as part of the National AI Strategy (SNIA) and through the 'Grand Défi' "Securing, making reliable and certifying artificial intelligence-based systems" or "trustworthy AI".

It is structured around 3 strategic and complementary pillars:

- **Pillar #1 Infrastructure** (see insert on the Confiance.ai program above) which aims to develop a design environment for trustworthy AI-based systems. It is structured around 7 independent projects and currently brings together more than 40 partners (large corporations, start-ups, SMEs, research laboratories).
- **Pillar #2 Compliance Assessment** which will ensure the proper operational conduct of the operation of systems based on trustworthy AI, and then define the role and competencies of certifiers or trustworthy third parties.
- **Pillar #3 Norms** which will make it possible to establish, in consultation with the various industry players, norms, standards, regulatory environment and certifications.

It aims to create a national AI ecosystem of trust, but, more broadly, through partnerships with other European (Germany, etc.) **and international** (Quebec, etc.) players, aims to inspire the European strategy through concrete achievements accompanying the operational implementation of the future European regulation on AI.

At the European level, it contributes to the emergence of future solutions for the development of trustworthy AI, to strengthen the resources and expertise of conformity assessment actors and to meet the needs of standards.

Finally, it aims to prefigure a future alliance for trustworthy AI, a real federator and catalyst of Europe's political vision on AI.

Create a European agency for the evaluation of AI, or even data and robotics

To ensure compliance with regulations (and norms or standards) within the European space and to advise, from a technical and normative point of view, public policies with respect to the increasingly rapid evolution of technologies, **we recommend the creation of a European agency for the evaluation of trustworthy AI.**

To do so, we propose two possible options: the creation of a new specific actor or the networking of existing national actors (and the associated governance) with expertise and technical characterisation platforms. The second option has the advantage of relying on an existing ecosystem with human and material resources, which will undoubtedly need to be supplemented.

The agency will be responsible for developing state-of-the-art AI metrology technologies, implementing the technical characterisation platforms to achieve this, and providing technical expertise in the field of normalisation and standards. On this last point, it will also facilitate the "right" articulation between the horizontal and sectoral approach, a delicate point for industrial sectors as we have previously seen. Support will be necessary to enable it to fulfil its missions.

Several components, both skills and material resources, are therefore essential:

- **evaluation and testing methodologies,** in connection with regulatory or homologation issues (but also norms and standards enabling the democratisation and dissemination of methods and results), including human factors;
- **physical characterisation and experimentation platforms,** digital platforms or simulation environments, in order to reduce evaluation and experimentation costs; real data from experiments also allowing performance to be improved;
- **cyber-security evaluation platforms** for these systems in order to guarantee the best practices and technologies to limit malicious attacks, especially those brought about by the introduction of AI;
- **protocols for defining use domains and test scenarios; and**
- **uses and social acceptability.**

This agency will support the rise in skills of the notifying and notified authorities of the various industrial sectors, as well as the compliance and audit ecosystems.

THE NATIONAL INSTITUTE OF STANDARDS AND TECHNOLOGY (NIST)

The NIST is an agency of the United States Department of Commerce. Its mission is to support the economy by developing technologies, metrology and standards in collaboration with industry. Founded in 1901, it is one of the oldest applied science laboratories in the United States. **It has a significant budget: $1 billion in 2010, $1.34 billion in 2020.** Its activity is supervised by the Committee on Science, Space and Technology of the United States House of Representatives. This agency was also entrustworthy with the study of the collapse of the Twin Towers during the 9/11 attack, in order to determine the technical reasons for the collapses and fires. In addition, NIST has no less than four Nobel Prize winners among the researchers it has employed.

Sources: NIST, Office of the Director, Congressional and Legislative Affairs "National Institute of Standards and Technology," US Department of Commerce "National Institute of Standards and Technology," Wikipedia

WE MUST CREATE A UNIFIED AND INDUSTRIAL GOVERNANCE FOR TRUSTWORTHY AI AND DATA AT THE EU LEVEL.

Unified and industrial data & AI governance

On 19 February 2020 Margrethe Vestager (Executive Vice-President of the European Commission) and Thierry Breton (European Commissioner for the Internal Market, Industrial Policy, Tourism, Digital, Audiovisual, Defense and Space) presented **the new digital "roadmap" for Europe: a coherent legislative framework for AI and data** has been laid down to serve the common goal of a Europe adapted to the digital and humanistic era. **Data is seen as "the fuel of artificial intelligence".**

The texts associated with the EU's data and AI strategies provide for the establishment of new governance bodies, known as data and AI "Boards". However, their construction is currently envisaged in different ways. At first glance, the structure seems similar, but the functioning and the role of stakeholders in particular are different.

For the Data Innovation Board[124]**,** everything is thought out in cross-sectoral terms, with interoperability at the heart of the action. **Numerous stakeholders are involved, including industrial players, who are involved in the process from the outset. All players are also encouraged to engage in dialogue and joint action.**

Conversely, for the Artificial Intelligence Board, the Commission's initial proposal (April 2021 version) compartmentalises the composition around the national supervisory authorities, which the Commission in turn supervises. The role of AI stakeholders is only mentioned, and industry players are not directly involved. **Yet, we will not succeed in building key bodies for digital regulation at the European level, balancing the missions of control and the preservation of innovation, without deeply multi-stakeholder bodies, capable of fostering interdisciplinarity and therefore of mixing different technical, legal and human and social sciences cultures, and even civil society.** The presence of technical experts, even if they are in the minority, but capable of following and understanding the evolution of the state of the art in technology and of exposing the stakes to the rest of the Board seems essential. **In this respect, the draft regulation on AI seems less accomplished than its equivalent on the data side, from which we recommend drawing inspiration.**

Comparison of the Boards envisaged for data and for AI (April 2021 version)

LEGISLATIVE ACT	GOVERNANCE	TASK
Data Governance Act : « Data Innovation Board	In order to successfully implement the data governance framework, a European Data Innovation Board should be established, in the form of an expert group. The Board should consist of representatives of the Member States, the Commission and representatives of relevant data spaces and specific sectors (such as health, agriculture, transport and statistics). The European Data Protection Board should be invited to appoint a representative to the European Data Innovation Board."	• "The Board should support the Commission in coordinating national practices and policies on the topics covered by this Regulation, and in supporting cross-sector data use by adhering to the European Interoperability Framework (EIF) principles and through the utilisation of standards and specifications (such as the Core Vocabularies44 and the CEF Building Blocks45), without prejudice to standardisation work taking place in specific sectors or domains. Work on technical standardisation may include the identification of priorities for the development of standards and establishing and maintaining a set of technical and legal standards for transmitting data between two processing environments that allows data spaces to be organised without making recourse to an intermediary. The Board should cooperate with sectoral bodies, networks or expert groups, or other cross-sectoral organisations dealing with re-use of data."
Artificial Intelligence Act : « Artificial Intelligence Board »	"composed of representatives from the Member States and the Commission"	"The Board will facilitate a smooth, effective and harmonised implementation of this regulation by contributing to the effective cooperation of the national supervisory authorities and the Commission and providing advice and expertise to the Commission. It will also collect and share best practices among the Member States."

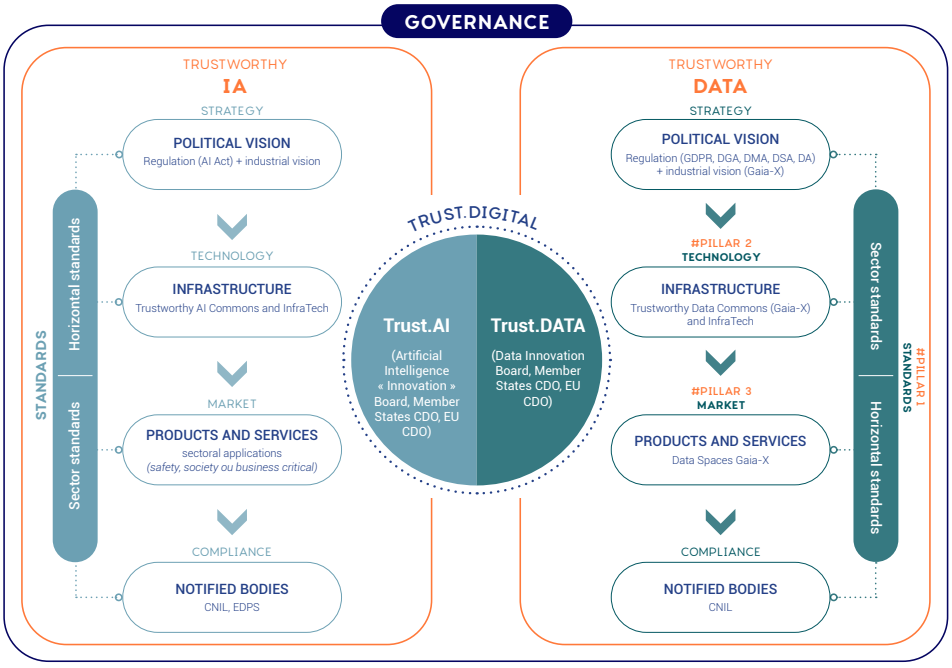
Let's take the example of the GDPR. The regulation is being jostled by the breakneck pace of technological innovation in AI. Having been negotiated and adopted before the explosion of contemporary AI application successes, it does not incorporate the notion of data-driven machine learning, nor the dichotomy that exists between the use of personal data for training and that for inference. However, they carry different kinds of risks for individuals. They therefore call for different balances in the application of numerous principles such as the length of retention, minimisation, the need for non-purpose data to build classifiers, the reinforced need for training data on protected categories to fight against bias, etc. However, five years after the "AI wave", the European Data Protection Supervisor (EDPS), probably as a result of its composition and a lack of means, has still not integrated these issues in its doctrine in a substantial way and in a way that is sufficiently clear and legible for European actors.

[124] *Regulation of the European Parliament and of the Council on European data governance* (Data Governance Act) 2020/0340 (COD)

We recommend that Europe promote a pioneering and fertile reflection allowing the reconciliation of the indispensable protection of personal data with technological and economic innovation, without which there is a real risk that the GDPR will be inadequate to technological development. A review of the composition of the EDPS, in relation to data and AI issues, coupled with a review of the resources allocated to it, should be undertaken.

As we have seen throughout this report, the subjects of data and AI are significantly related. It also appears that **an end-to-end strategy (regulation, standard, industrialisation) is the sine qua non condition to defend our interests.** Consequently, we also recommend **unifying the governance of all the components of the trust ecosystem (Cloud-Data-AI) in a single body, which would be driven by an overall industrial policy vision at the European level. This unified European governance could, for example, be accompanied by the appointment of a CDO (Chief Data Officer) per Member State, based on the Spanish model (Alberto Palomo-Lozano first CDO of Spain), and a CDO for Europe.**

Governance of the AI and Data trust ecosystem



SUMMARY OF RECOMMENDATIONS

RECOMMENDATION 1 :

MAKING DIGITAL TRUST A EUROPEAN POLITICAL PROJECT

1.a Impose trust as a doctrine of the digital third way.	Bringing the message of trustworthy AI at a societal level and making it the only reasonable and feasible path.
1.b Assume our strategy of extraterritorial regulation.	Make the regulatory package (GDPR, DSA, DMA, DGA, DA, AI Act), and the risk-based approach, the foundation stone in the building of a global ecosystem of trust.
1.c Promote a matrix vision of of trustworthy AI.	Defend a horizontal cross-industry approach, and a vertical approach to the ecosystem of trust (from regulation to industrialization, via the standard).

RECOMMENDATION 2 :

MAKING STANDARDISATION A STRATEGIC PRIORITY

2.a Promote trust in AI as a European "brand".	Develop generic European standards and labels based on the regulations to conquer the market.
2.b Become a standard-setter by mobilizing all the actors.	Coordinate and amplify the efforts of the EU, the Member States, large corporations, as well as start-ups/SMEs.
2.c Guarantee a balance between regulation and innovation of AI through sandboxes.	Promote regulatory sandboxes, following the criteria proposed by the OECD, in order to maintain the agility that is essential for the future implementation of the regulation.

RECOMMENDATION 3 :

BUILDING A TRUSTWORTHY AI INDUSTRIAL STRATEGY

3.a Leverage trust to enter the global AI market.	To conquer the global AI market (€231 billion), by penetrating the market for AI market via the "accessible" market of AI engineering tools and methods.
3.b Make safety critical systems culture the tip of the strategy arrow.	Capitalize on the European culture of safety critical systems, where requirements are the highest, to turn it into a competitive advantage.
3.c Develop an industry and innovation-oriented industry approach.	Capitalize on our research expertise to increase our R&D and innovation capacity and fund more training, particularly in the fields of engineering.

RECOMMENDATION 4 :

CREATING THE INFRATECH OF TRUSTWORTHY AI

4.a Foster the emergence of a large ecosystem of trustworthy AI InfraTech players.	Create and animate a large European innovative community (in particular start-ups) to contain and above all conquer the AI market, which is dominated by Big Tech players.
4.b Develop a software platform based on trustworthy AI InfraTech players.	Federate the InfraTech and create a packaged, end-to-end offer, covering the entire the entire AI risk management chain, to enable business players to scale up.
4.c Foster the creation of trustworthy AI Digital Commons.	Lower barriers to entry by pooling development efforts via open source digital commons governed by a multitude of InfraTech players.

RECOMMENDATION 5 :

SCALE UP AT THE EUROPEAN LEVEL

5.a Build a European industrial alliance industrial for trustworthy AI.	Capitalize on the projects under construction in the Member States and create a core alliance from the first partnerships (France, Germany, Spain, etc.).
5.b Imagine a European center for AI evaluation.	Promote the development of a European audit and conformity assessment ecosystem, in partnership with the sectoral bodies.
5.c Supporting the creation of industrial demonstrators.	Launch European lighthouse projects for industrial demonstrators based on trustworthy AI in strategic sectors.

RECOMMENDATION 6 :

PROPOSE A COMMON DATA AND AI GOVERNANCE

6.a Develop trustworthy AI for the common European data spaces.	Apply trustworthy AI solutions to multi-sector data spaces for society and business critical use cases (education, health, tourism, Green Deal,...).
6.b Embody Data and AI strategies in Europe.	Promote the appointment of a Chief Technical Officer (CTO) or Chief Data Officer (CDO) for each Member State and for Europe.
6.c Create a unified and industrial governance for AI and Data in Europe.	Design a new unified European governance body for AI and Data, ensuring a balance between protection and innovation, including all stakeholders (regulation, industry, society, etc.).

CONCLUSION

Mastering sovereign and secure digital technologies is an imperative necessity. France 2030, which is mobilizing €3 billion to this end, embodies the renewed French will to make our European digital independence a tangible reality. This ambition, which has been supported at the highest level for several years, has become a requirement since the Covid crisis revealed our dependence on certain critical technologies. It is based on two fundamentals. The first is to protect our organizations, both public and private, and our fellow citizens, and to ensure that our values are respected by relying on regulation (GDPR, DSA, DMA, AI act, Cyber act). The second is to develop sovereign offers, credible alternatives to the non-European giants, by investing in and supporting innovative and trustworthy ecosystems.

To achieve this goal, France is initiating joint industrial projects, particularly in the field of artificial intelligence. France 2030 is designed to strengthen this ambition by giving French players the means to invest in AI, but also to but also to train for tomorrow's jobs in this field.

But to do this, it was important for us that a report clarifies the terms of this strategy: What do we mean by trustworthy AI? Can the European Union impose trustworthy AI as a benchmark? Can France make trustworthy AI an opportunity for its industrial autonomy? Answering these questions, and many others, will allow us to offer a roadmap that will allow France to take control of its digital destiny. The France of 2030 will then be able to play its part, making investments, sometimes risky if necessary, to address this great digital transition, which is, along with health and the environment, one of our three priorities.

This note is an important contribution to this process. It reminds us of the necessary cooperation between emerging players and industrial palyers to build a true digital independence, a prerequisite for the competitiveness of our companies. Trustworthy digital technology (Cloud-Data-AI) is essential to support these objectives. This high-quality report offers guidance on the strategy to be implemented. The authors, with the help of prestigious contributors, validate the investments we have made in the framework of the AI strategy funded by France 2030, and offer an interesting perspective for the future.

The key proposals of this report, which consist on the one hand in mastering strategic infrastructures of digital trust to guarantee our sovereignty, and on the other hand in consolidating a common AI and data strategy through a governance oriented towards Europe, are heading in the right direction. They reinforce our choices and provide us with the direction to go further, by equipping us with our “digital compass”, which must point us in the same direction as our ecological course.

Bruno Bonnell

Secretary General for Investment, France 2030.

ACKNOWLEDGEMENTS

Julien Chiaroni, Director of *Grand Défi IA*, General Secretariat for Investment (SGPI) and **Arno Pons**, General Delegate, Digital New Deal, **would like to thank for their contribution:**

ANIMATION AND EDITORIALISATION

Olivier Dion — Technical and editorial coordination, Digital New Deal. CEO Onecub, co-founder aNewGovernance

Prune Zammarchi — Project manager, Digital New Deal

ADVICE AND CONTRIBUTIONS

Guillaume Avrin — Head of LNE AI Evaluation Department

Patrick Bezombes — CEN CENELEC, co-chairman of the "Digital Sovereignty" workshop, Vice-Chairman of JTC 21 (AI), AFNOR, Chairman of the AI and Big Data standardization committee

Matthias de Bièvre — CEO Visions, co-founder aNewGovernance, CEO Prometheus-X

Yannick Bonhomme — Head of IRT SystemX, Confiance.ai

Bertrand Braunschweig — Scientific coordinator of the "Confiance.ai" program

Loic Cantat — R&D Manager AI and Data Science - IRT SystemX

Agnès Delaborde — LNE Research Engineer

Emmanuelle Legrand — AI Project Manager, DGE Ministry of Finance

Juliette Mattioli — Senior AI expert, Thales, President of the "Data Sciences & Artificial Intelligence" Hub of the Systematic Paris Region cluster

Eric Pol — Chairman aNewGovernance

Benoît Rottembourg — REGALIA project manager, INRIA

Renaud Vedel — Prefect, Coordinator for the Government of the National AI Strategy, Co-chair, Steering Committee Global Partnership on AI (GPAI)

INTERVIEWS AND ADDITIONAL PARTICIPATIONS

Christine Balagué — Professor, Good in Tech Chair (Institut Mines-Télécom)

Julien Chasserieau — AI Policy Manager, DIGITAL EUROPE

Thierry Collette — Director of the Information Science and Technology Group, Thales

Andreas Dengel — Prof. Dr. h.c. Deutsches Forschungszentrum für Künstliche Intelligenz (DFKI)

Laurence Devillers — Professor of Computer Science, University of Paris-Sorbonne, co-chair CNRS of the AI ad hoc group "Ethics/Nudging" for AFNOR

Emmanuelle Escorihuela — AI Transformation Leader for Systems, Airbus

Jean-Baptiste Fantun — CEO NukkAI

Kilian Gross — Head of AI Unit, DG Connect, European Commission

Miapetra Kumpula-Natri — Member of the European Parliament, Data Act rapporteur (Finland)

Guillaume Leboucher — Deputy CEO Openvalue, COMEX member Docapost

Marc Leobet — Project Director "AI and Ecological Transition", Ministry of Ecological Transition

Maya Noël — Managing Director, France Digitale

Dragos Tudorache — Member of the European Parliament, AI Act rapporteur (Romania)

Axel Voss — Member of the European Parliament, shadow rapporteur AI Act (Germany)

JULIEN CHIARONI

Director of the French Grand Défi on "Trustworthy AI for industry" within the General Secretariat for investment (SGPI), expert of the European Innovation Council. He has previously held positions at the Institute of Microelectronics and information technologies (Leti), and at the Institute for Digital Technologies and Artificial Intelligence (List), which has more than700 researchers working in Paris-Saclay and Grenoble. As Director of strategy and programs, he implemented the institute's strategy in a wide range of digital technologies, and established partnerships between research and and industry. From 2008 to 2010, he coordinated the nanoscience and nanotechnology program at the French National Research Agency. He was also attached to academic affairs at the Consulate General of France in Hong Kong and Macao.

ARNO PONS

General Delegate of the "Think-tank Digital New Deal", co-author of four notes on trustworthy digital (Trustworthy Cloud, Trustworthy Infrastructures, Trustworthy AI, and Trustworthy Data). He founded in 2021 the "Do Tank - Digital New Deal" activity dedicated to cooperation challenges and helping companies and local authorities in structuring their industries into digital ecosystems, through technological alliances. In 2022, he is launching the first initiative of the Do Tank by co-creating Themis (data space for tourism including sixty partner entities). He had previously created several start-ups (Checkfood-food waste, Medicimo in Canada, ...), and also taught at SciencesPo on the issues of digital sovereignty related to the centralization of powers by Big Tech.

ABOUT FRANCE 2030

INVESTMENT PLAN

- The aim is twofold: to transform key sectors of our economy (energy, automotive, aeronautics and space) in a sustainable way through technological innovation, and to position France not just as a player, but as a leader in tomorrow's world. From fundamental research to the emergence of an idea to the production of a new product or service, France 2030 supports the entire life cycle of innovation to its industrialization.
- Unprecedented in its scope: 54 billion euros will be invested so that our companies, universities, research organizations, can make their transitions in these strategic sectors a success. The challenge is to enable them to respond competitively to the ecological and attractiveness challenges of the world ahead, and to bring out the future champions of our fields of excellence. France 2030 is defined by two cross-cutting objectives: to devote 50% of its funds to decarbonizing the economy, and 50% to emerging players, carrying innovations than do not harm environment (as defined by the of the Do No Significant Harm principle).
- Will be implemented collectively: designed and deployed in consultation with economic, academic, local and European players to determine the strategic orientations and key actions. Project leaders are invited to submit their applications via open and selective procedures to benefit from the support of the State.
- The program is managed by the General Secretariat for Investment on behalf of the Prime Minister.

More information: https://www.gouvernement.fr/france-2030

DIGITAL NEW DEAL
THE THINK-TANK OF THE NEW DEAL

Digital New Deal accompanies private and public decision-makers in the creation of an Internet of the Enlightenment, European and humanistic. We are convinced that we can offer a 3rd digital way by aiming at a double objective: to defend our values by proposing a new regulation against the centralization of powers; and to defend our interests by creating the conditions of cooperation against the capture of value by the "Big Tech".

The purpose of our publication activity is to shed as much light as possible on the developments at work within the issues of "digital sovereignty", in the broadest sense of the term, and to develop concrete courses of action, even operative via the Do Tank, for economic and political organizations.

THE BOARD OF DIRECTORS

Olivier Sichel (founding president) and Arno Pons (general delegate), steer the strategic orientations of the think-tank under the supervision of the board of directors.

Strengthened by their common interest in digital issues, the members of the Board of Directors have decided to deepen their debates by formalizing a framework for production and publication within which the complementarity of their experiences can be put at the service of public and political debate. They are personally involved in the life of Digital New Deal, especially in the choice of reports and their editors. They are the guarantors of our academic and economic independence.

SÉBASTIEN BAZIN
PDG AccorHotels

NATHALIE COLLIN
General Manager, Consumer and Digital Division La Poste Group

NICOLAS DUFOURCQ
DG of Bpifrance

AXELLE LEMAIRE
Former Secretary of State for Digital Technology and Innovation

ALAIN MINC
President AM Conseil

DENIS OLIVENNES
DG Libération

YVES POILANE
DG Ionis Education Group

ARNO PONS
General Delegate of the Digital New Deal think tank

JUDITH ROCHFELD
Associate Professor of Law, Panthéon Sorbonne

OLIVIER SICHEL
President Digital New Deal
DGA Caisse des Dépôts

BRUNO SPORTISSE
PDG Inria

ROBERT ZARADER
PDG Bona fidé

OUR PUBLICATIONS

Cybersécurité, vigile de notre autonomie stratégique | Arnaud Martin, Didier Gras - *June 2022*

RGPD, acte II : la maîtrise collective de nos données comme impératif | Julia Roussoulières, Jean Rérolle – *May 2022*

Fiscalité numérique, le match retour | Vincent Renoux - *September 2021*

Défendre l'état de droit à l'ère des plateformes | Denis Olivennes and Gilles Le Chatelier - *June 2021*

Cloud de confiance : un enjeu d'autonomie stratégique pour l'Europe | Laurence Houdeville and Arno Pons - *mai 2021*

Livres blancs : Partage des données & tourisme | Fabernovel and Digital New Deal - *April 2021*

Partage de données personnelles : changer la donne par la gouvernance | Matthias de Bièvre and Olivier Dion - *September 2020*

Réflexions dans la perspective du Digital Services Act européen | Liza Bellulo - *March 2020*

Préserver notre souveraineté éducative : soutenir l'EdTech française | Marie-Christine Levet - *November 2019*

Briser le monopole des *Big Tech* : réguler pour libérer la multitude | Sébastien Soriano - *September 2019*

Sortir du syndrome de Stockholm numérique | Jean-Romain Lhomme - *October 2018*

Le Service Public Citoyen | Paul Duan - *June 2018*

L'âge du web décentralisé | Clément Jeanneau - *April 2018*

Fiscalité réelle pour un monde virtuel | Vincent Renoux - *September 2017*

Réguler le « numérique » | Joëlle Toledano - *May 2017*

Appel aux candidats à l'élection présidentielle pour un #PacteNumérique | *January 2017*

La santé face au tsunami des NBIC et aux plateformistes | Laurent Alexandre - *June 2016*

Quelle politique en matière de données personnelles ? | Judith Rochfeld - *September 2015*

Etat des lieux du numérique en Europe | Olivier Sichel - *July 2015*

contact@thedigitalnewdeal.org

September 2022

www.thedigitalnewdeal.org