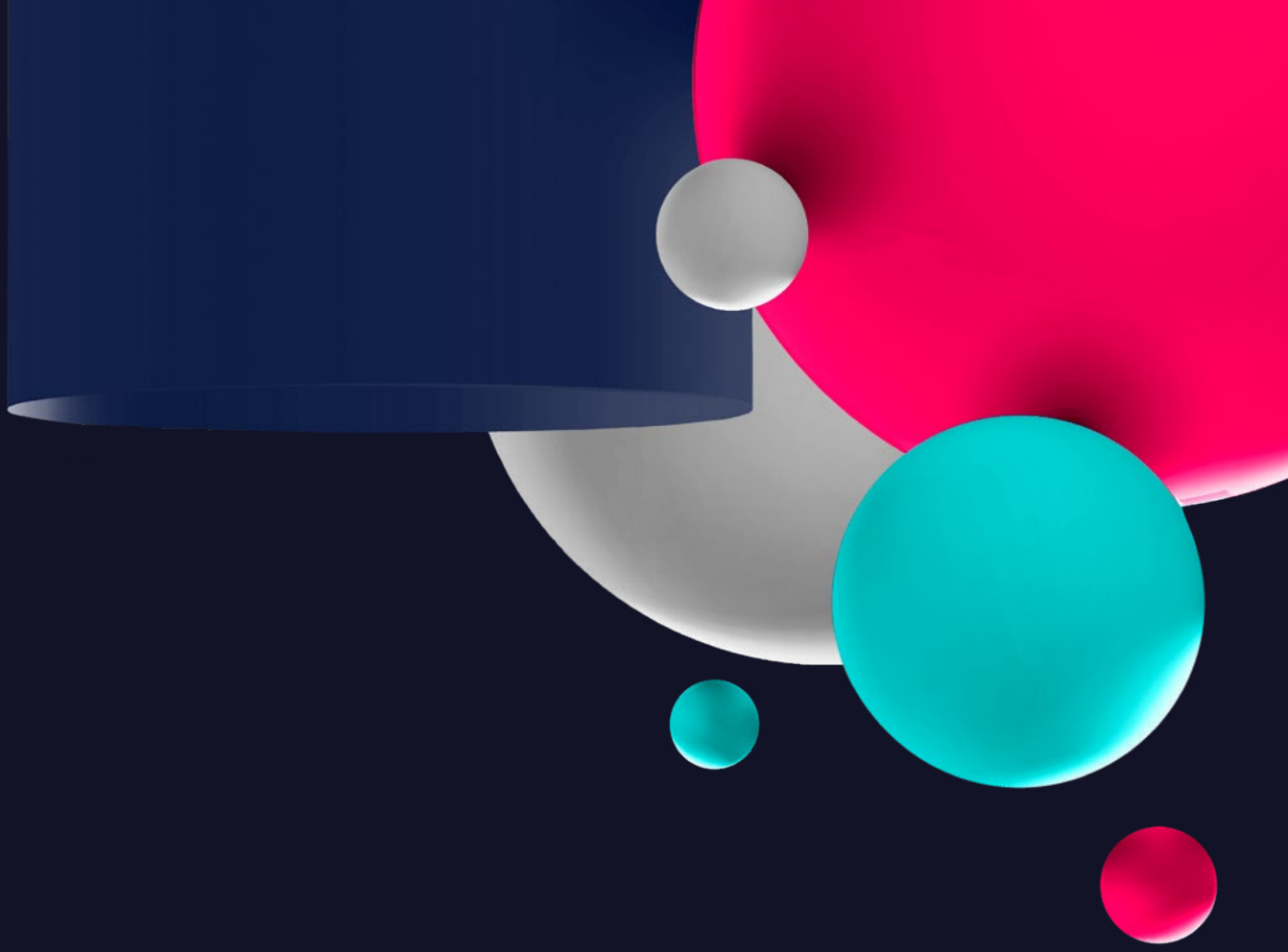


ARTEFACT

ARTIFICIAL INTELLIGENCE & ETHICS

New challenges and data-driven solutions for the deployment of trustworthy AI in organizations.

Silver
Business
Partner

IBM

ARTEFACT

We transform **data** into **value** and business **impact**.

15
COUNTRIES

+1000
EMPLOYEES

+300
MAJOR BRANDS

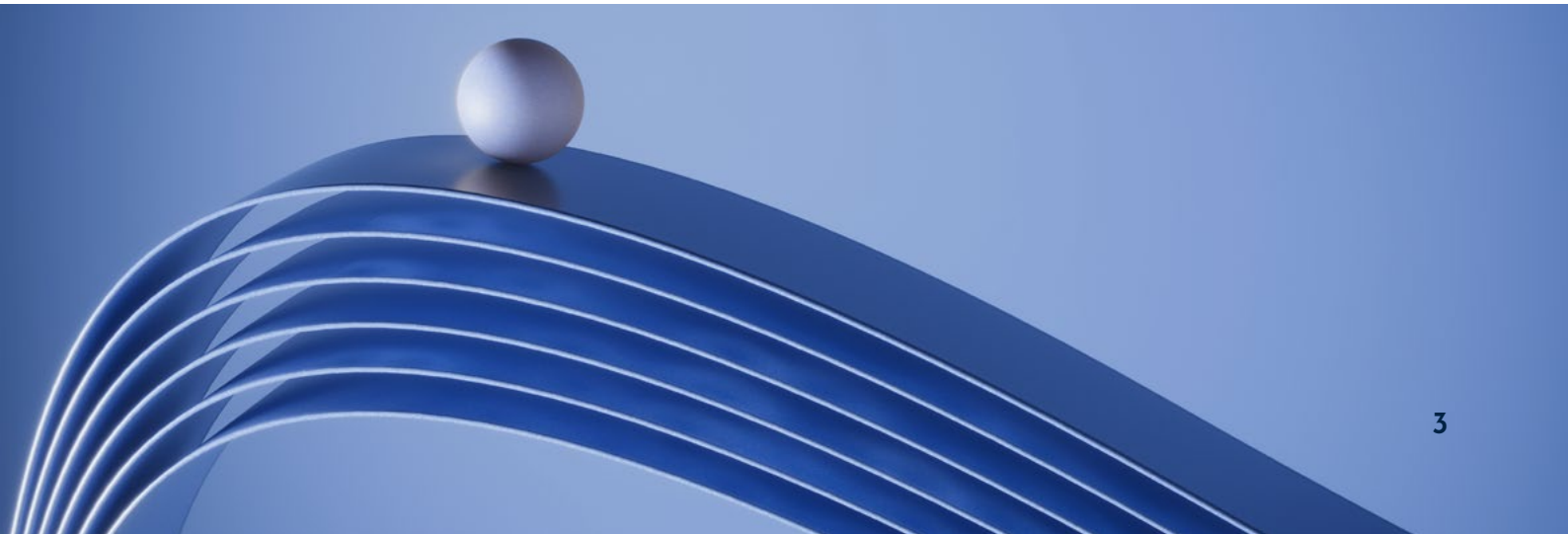
Artefact is a global data-driven services company. Our offers sit at the intersection of consulting, marketing and data science, putting consumers at the heart of enterprises' digital transformation.

DATA CONSULTING | DATA & DIGITAL MARKETING | DIGITAL COMMERCE

Artificial Intelligence & Ethics: new challenges and data-driven solutions for the deployment of trustworthy AI in organizations.

Vincent Perrin
Partner Ecosystem Technical Leader
IBM

Hanan Ouazan
VP Data Science
ARTEFACT



1 – Solutions for trustworthy Artificial Intelligence by design - a dialogue between ethics and technology

In recent years, AI and Ethics have regularly fueled debates as controversies involving Big Tech players have erupted. To cite only some of the most noteworthy cases:

- Microsoft's chatbot, designed to mimic the behaviors of Twitter users, adopted racist attitudes after hundreds of Tweeters fed the AI discriminatory language shortly after it went live on Twitter.
- At an intersection, Tesla's autonomous car failed to stop when it encountered a truck because it had learned to identify a vehicle from the front or back, so a truck in profile was interpreted as a sign.
- Google's AI identified photos featuring black-skinned people as content about gorillas.
- Apple Card gave higher credit lines to men than women.

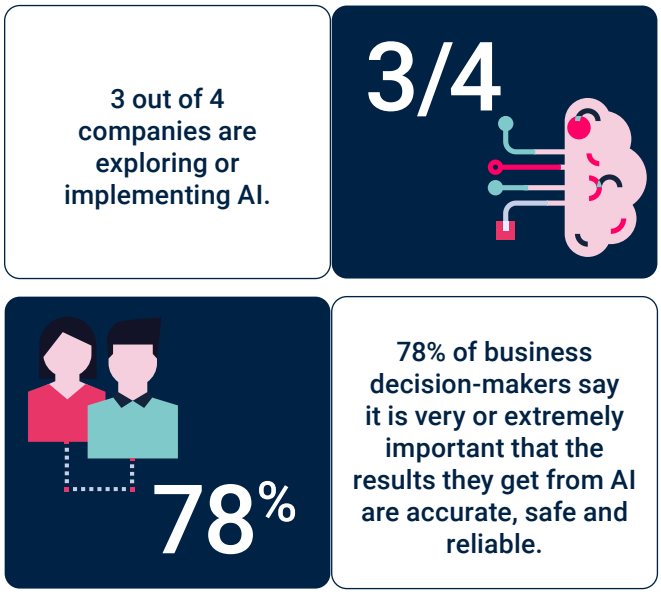
These incidents recall certain 20th century scenarios in which Artificial Intelligence systems revolt and begin behaving according to their own obviously evil value systems. Think of HAL 9000 from Space Odyssey or SKYNET from Terminator. Though rebellious AIs remain a fantasy of science fiction writers, cases of malfunctioning AI have become common, often with very real consequences.

The real risk linked to the mass use of AI is not that algorithms rebel and adopt behaviors counter to those they were programmed for. Rather, problems arise precisely when AI behaves exactly like we asked it to, mimicking our biases, repeating our mistakes, amplifying our uncertainties and inaccuracies.

Given that algorithms are widely used in everyday activities such as shopping online, signing banking contracts or consulting content on social networks, we cannot ignore the ethical implications raised by the use of Artificial Intelligence.

For trustworthy Artificial Intelligence, the real challenge is to ensure that it is designed in a way that does not reproduce our cultural biases.

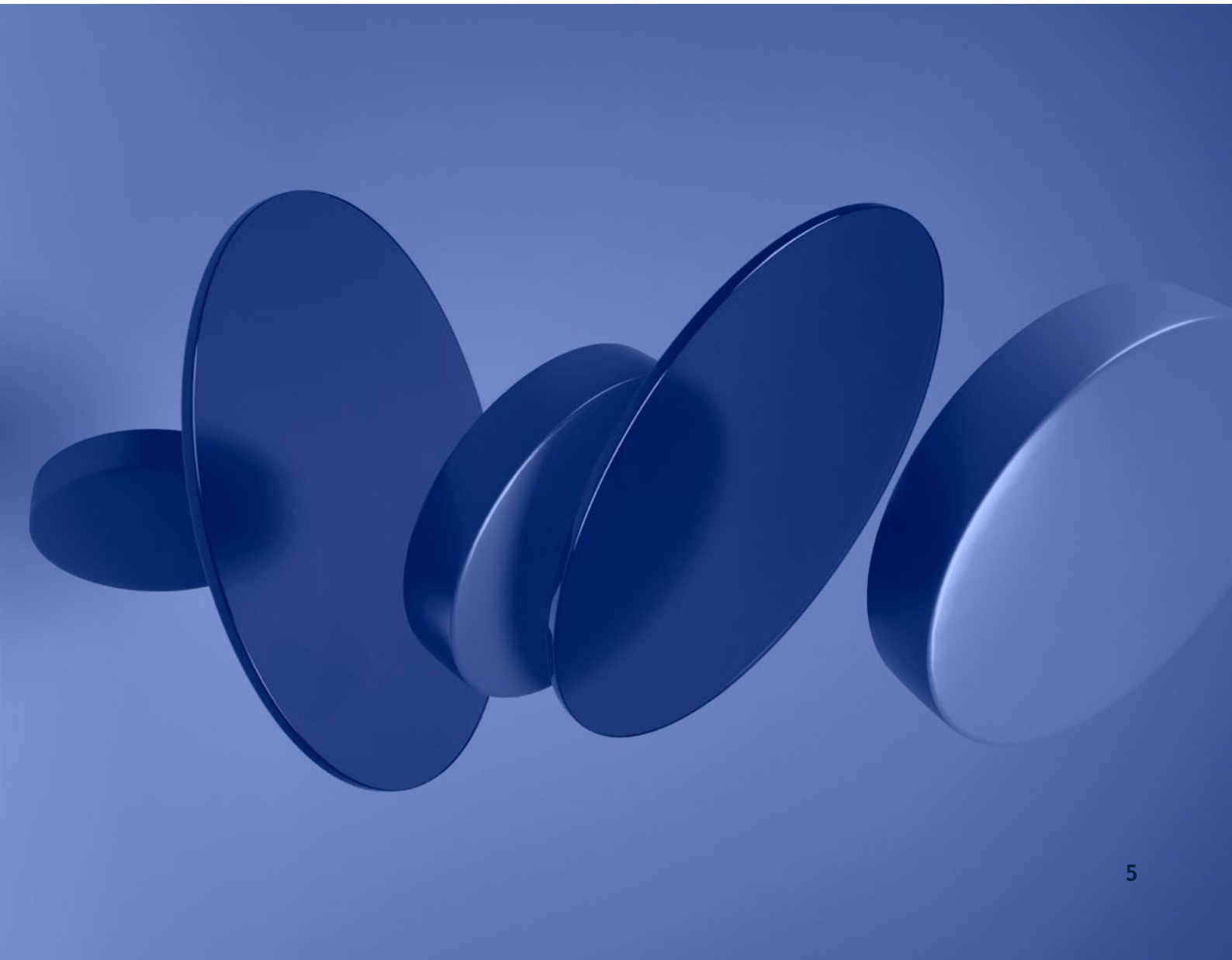
From now on, it is essential to prepare for the design and maintenance of trustworthy AI systems that comply with current and upcoming regulations. But this is a vast undertaking, encompassing the technical, legal, organizational and cultural dimensions of the company. Moreover, ethics is a subject that is built by design throughout the lifecycle of every product.



Source: IBM's From Roadblock to Scale: The Global Sprint Towards AI Study 2020

Many companies are uncertain about what strategy and roadmap to adopt. To help them in their journey, we've prepared this white paper which addresses these issues around three pillars:

- Legal compliance covers the principles of respect for the privacy of individuals and the accountability of companies.
- Technical design of AI and its lifecycle explores both technical challenges and existing solutions.
- Transformation management, or change management, is a transversal project and is particularly focused on «societal and environmental well-being».



II – Between Soft Law and regulatory premises, a legal environment in construction

The European Union has been at the forefront of technological legislation, most prominently and recently with the General Data Protection Regulation. Its consequences and applicability have had implications reaching far outside the EU's borders, making it relevant for global tech actors. Though the EU has not yet adopted any binding regulation regarding AI, draft legislation is being debated. It is thus critical for businesses dealing with AI systems, whether in the EU or not, to understand the EU's position on AI, and its strategy towards achieving trustworthy AI.

The Commission's work lists seven essential requirements for a trustworthy AI:

1 - Human agency and oversight

2 - Technical robustness and safety

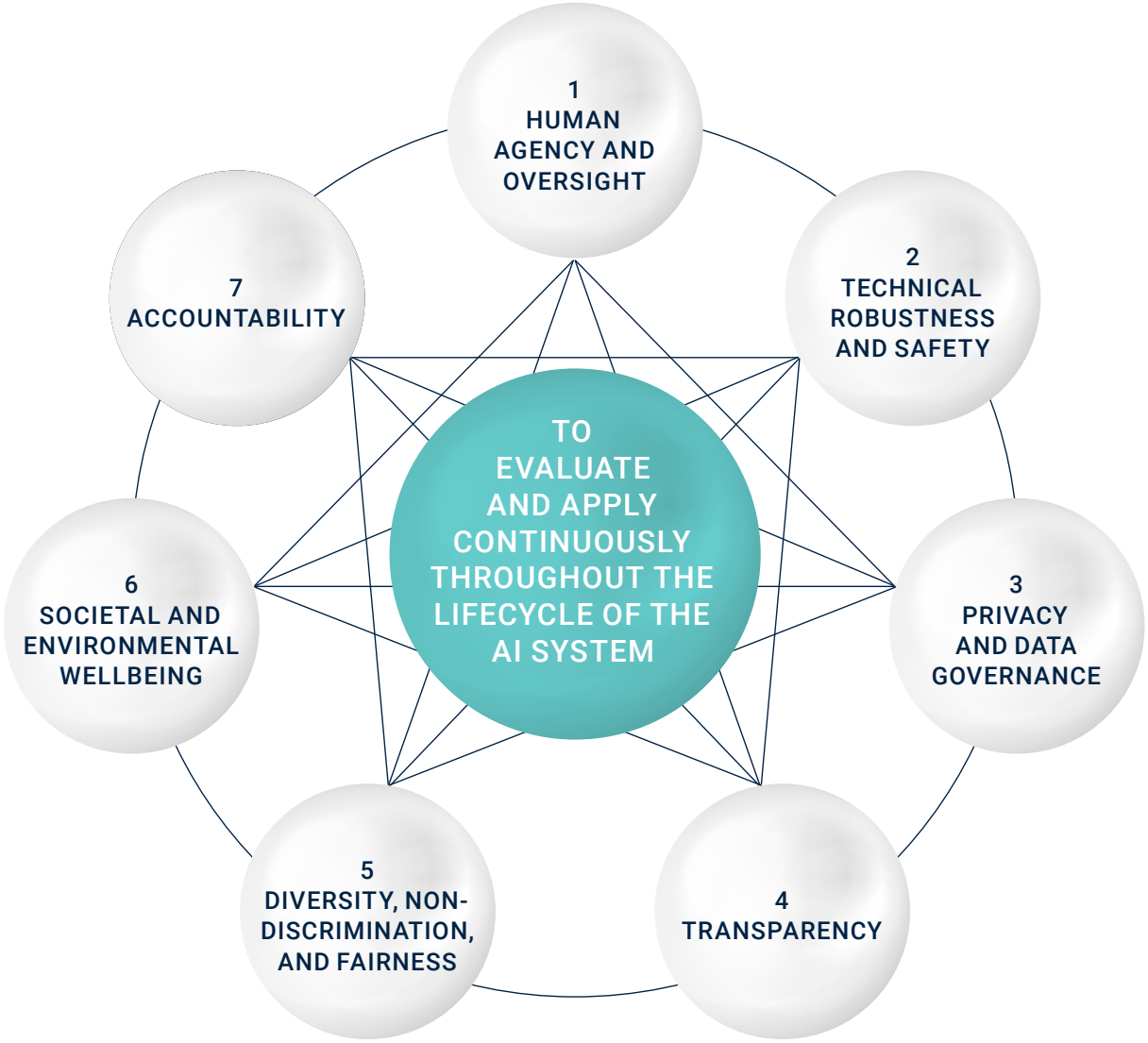
3 - Privacy and data governance

4 - Transparency

5 - Diversity, non-discrimination, and fairness

6 - Societal and environmental wellbeing

7 - Accountability



Source: Ethics guidelines for trustworthy AI | Shaping Europe's digital future

1. The 7 Key Requirements for Trustworthy AI [1]

1 Human agency and oversight

AI systems must support human agency and fundamental rights. They cannot diminish, limit, or misdirect human autonomy. The general well-being of the user is thus at the heart of the system's functionality.

Human surveillance and oversight guarantee that an AI system does not reduce or hinder human autonomy, nor does it cause other undesirable effects. Given the wide range of AI systems and their use cases, each system needs specifically tailored levels of appropriate control measures, including adaptability, accuracy, and explainability of the system. Oversight can best be achieved through adequate governance mechanisms.

Exemple: *After using an algorithm in the context of a bank loan, a bank agent confirms or rejects the AI's recommendations, and clearly states the reasoning behind the final decision given to the customer.*

2 Technical robustness and security

AI systems must be designed with a preventive approach and incorporate a risk assessment throughout their lifecycle. Their decisions must be accurate, and their results must be reproducible. AI systems must integrate safety and security mechanisms by design to reduce the risk of exposing private and sensitive data.

Exemple: *A bank uses an image detection algorithm to automatically sort through and process checks. With enough ink and smudging on the check, a malicious user can transform a 5 into an 8.*

3 Privacy and data governance

Users must be able to fully control their personal data: they cannot be used for purposes other than those they were explicitly collected for. Data integrity must be ensured throughout the system's lifecycle. Finally, data access must be adequately governed and controlled.

Exemple: *Give users the ability to download their data and assist users in exercising their right to be forgotten.*

1 - *Building Trust in Human-Centric Artificial Intelligence*, Communication from the Commission to the European Parliament, the Council, the European Economic and Social Committee and the Committee of the Regions, Brussels, April 2019.

4 Transparency

The traceability of AI systems must be ensured: this can be achieved by recording and documenting both a system's development process and the decisions it takes. Moreover, explainability of the algorithmic decision-making process must be provided to stakeholders as far as possible.

This task can be complex, as some systems, known as black-box models, have reached such levels of abstraction that even their designers cannot fully comprehend their inner workings. Technical explainability should be coupled with explaining how AI influences the final human decision, the design choices made, and the rationale for deployment. All this information needs to be easy to find and understand for all stakeholders.

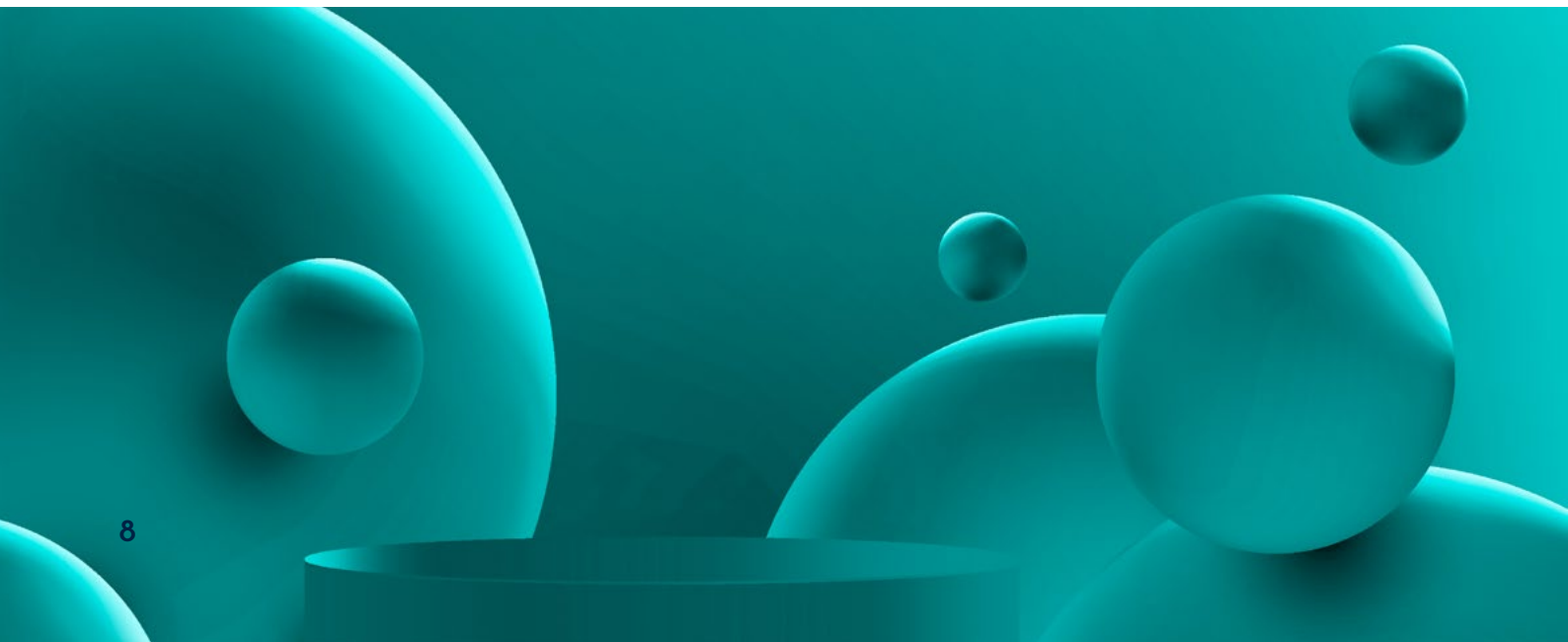
Finally, AI systems must be identifiable as such, so that users know they are interacting with a robot and which people are responsible for it.

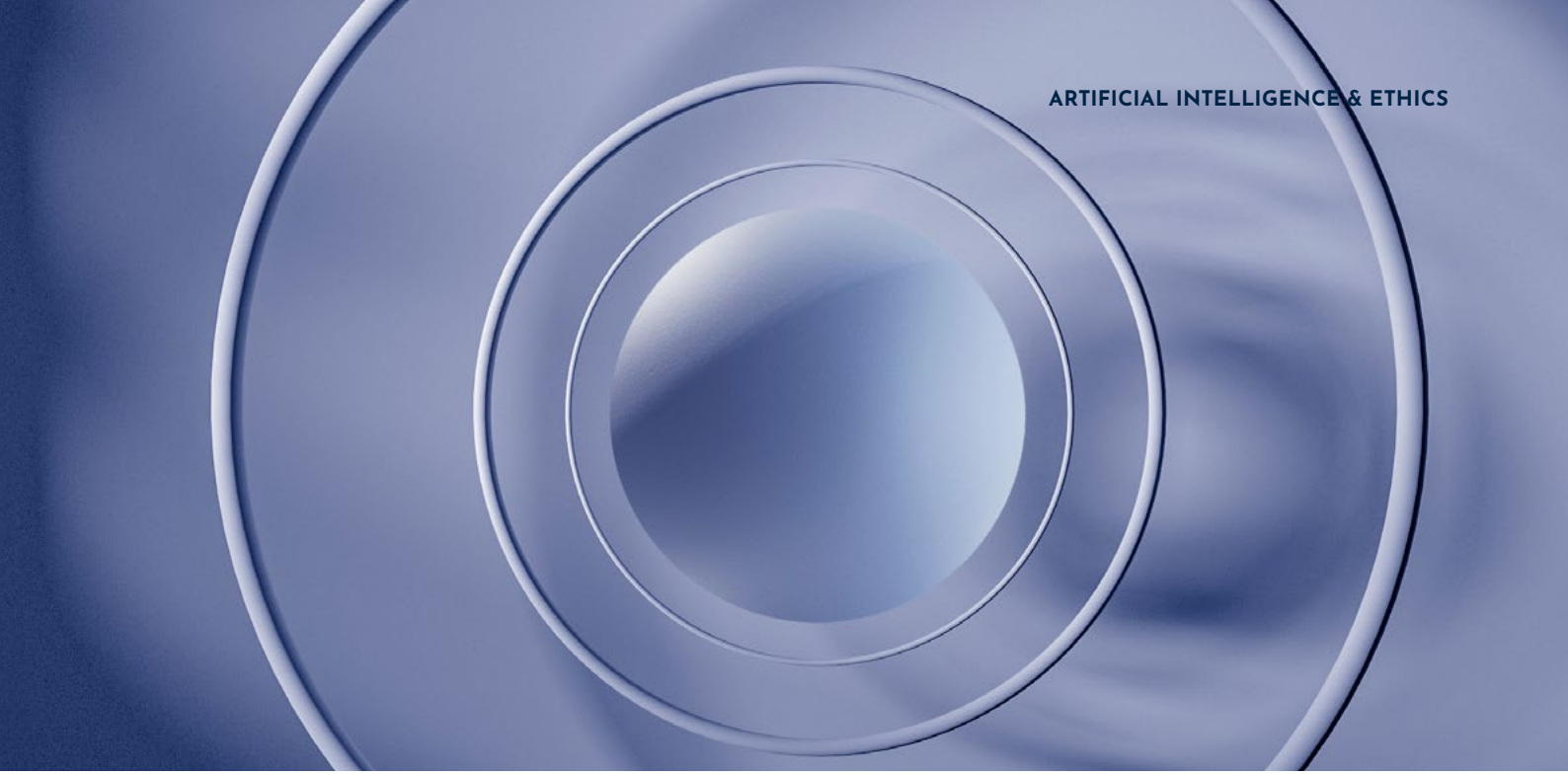
Counterexample: *Users have a fragmented or erroneous view of the performance of an AI system, of the data it uses to operate, and how its recommendations are established.*

5 Diversity, non-discrimination, and fairness

Datasets used by AI systems and design choices for these systems can be influenced by and suffer from involuntary historical biases, incompleteness of data, or poor governance. The persistence of such biases can lead to direct or indirect discrimination. Prejudice can also result from intentionally exploiting consumer bias or engaging in unfair commercial competition. These concerns must be addressed at the system design stage. Note that the complexity of fairness lies in the fact that there may be a correlation between discriminatory characteristics and others that are completely innocuous.

Conterexemple: *Filtering candidate resumes based on nationality or gender.*





6 Societal and environmental well-being

For AI to be trustworthy, its social and environmental impact is paramount. The sustainability and ecological responsibility of AI systems are therefore encouraged. Furthermore, their impact must be considered not only from an individual point of view, but also on a societal scale. Particular attention should be paid to use cases touching upon the democratic process, such as opinion-forming, political decision-making or electoral contexts.

Exemple: *A system that uses image analysis to identify non-recyclable waste in a recycling center.*

7 Accountability

Mechanisms need to be put in place to guarantee a chain of accountability for decisions made when designing or deploying the system, and for the outcomes of the system. Additionally, auditability is a must, whether it be done by employees of the company that designed the AI system or by third parties. Audits must be easily available to contribute to the trustworthiness of AI technology.

Specific documentation must be produced to identify, evaluate and minimize the potential negative impacts of AI systems. Finally, dedicated mechanisms must be put in place to mitigate any known malfunctions.

Counterexample: *Not defining a chain of responsibility to deal with an algorithm dysfunction.*

2. A global legal framework in development

The seven principles published by the European Commission are only non-binding recommendations and as such do not create any new legal obligations. However, some of these principles reflect existing European legal provisions, especially those concerning security, personal data protection, respect of private life, or environmental protection.

Of greater importance is a draft regulation on AI, called the "Artificial Intelligence Act," that was proposed by the European Commission on April 21, 2021. Its goal is to provide a legal framework for the rights and duties of those responsible for AI systems, as well as to define sanctions. It is scheduled to be adopted by EU Member States in 2023 and to come into effect between 2025 and 2026.

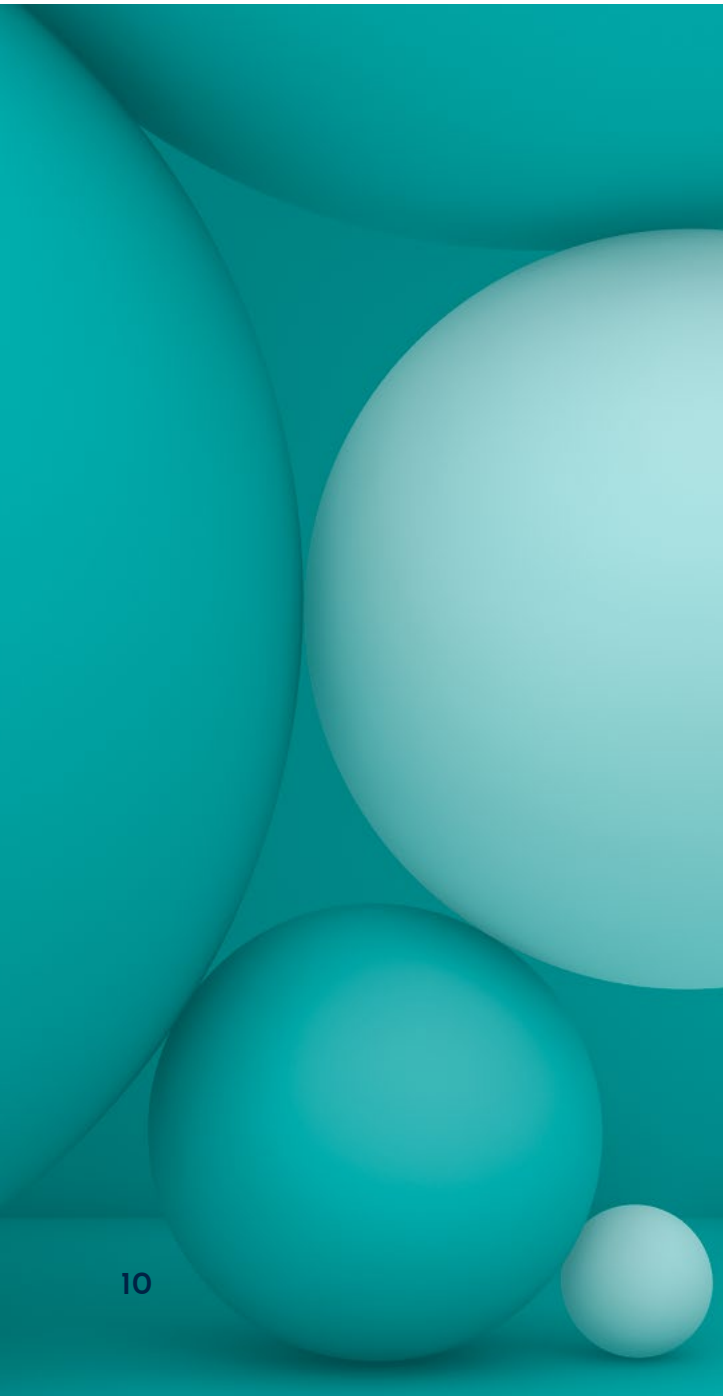
The Act has undergone numerous amendments and modifications, with a second version being currently debated. This new version emphasizes two points:

- **Semantics:** What is Artificial Intelligence? Will regulation cover the data used by AI systems? What types of data are included?

- **Governance:** Will the regulation cover AI systems designed outside the EU? How might this affect or stifle innovation in the EU? How can the development of new technologies be fostered whilst guaranteeing the confidence and trust of affected individuals? What governance? Who audits, oversees, and sanctions?

At the international level, important work is being led by other multilateral organizations, such as the Council of Europe, UNESCO, the OECD, the WTO, or the ITU (International Telecommunications Union).

The European Union has taken a particularly active role in developing the OECD's Ethical Principles for AI, the contents of which were endorsed by the G20 during its June 2019 ministerial meeting on trade and the digital economy. Within the UN, the EU participates in the follow-up to the report of the High-level Panel on Digital Cooperation, and particularly its AI recommendations.



3. Complying with current regulations and anticipating future ones

Because the regulatory framework for AI systems remains to be defined, it is challenging for companies to fully invest in it. Which projects should be prioritized? How can one prepare legal compliance given changing laws? Which geographic scopes are relevant?

The relatively recent case of the General Data Protection Regulation (GDPR) demonstrated that these projects require such large transformational projects from businesses that they need to be well anticipated. This remains true even when considering the grace period that often comes with such new regulations.

Although implementation of the Artificial Intelligence Act is not expected until 2025/26, a number of regulations and standards are already in place to frame and clarify processes and responsibilities, such as:

- Collection and use of personal data (RGPD)
- Contracting with service providers and sharing of responsibilities
- Procedures in case of AI system drifts
- Processes for monitoring legal issues and updating corporate standards

Whether you need an audit, development of a governance policy or deployment of documentation tools, the company that advises you on your AI strategy and implements your technical solutions can assist you in initiating your action plan.

The adoption of ethical behavior should not only be a reaction to potential sanctions, but should above all meet the requirements of your stakeholders and users, who have become increasingly demanding with regard to these issues over the last few years.

III – Technical solutions to correct biases from conception onwards

1. Why Artificial Intelligence systems may be biased

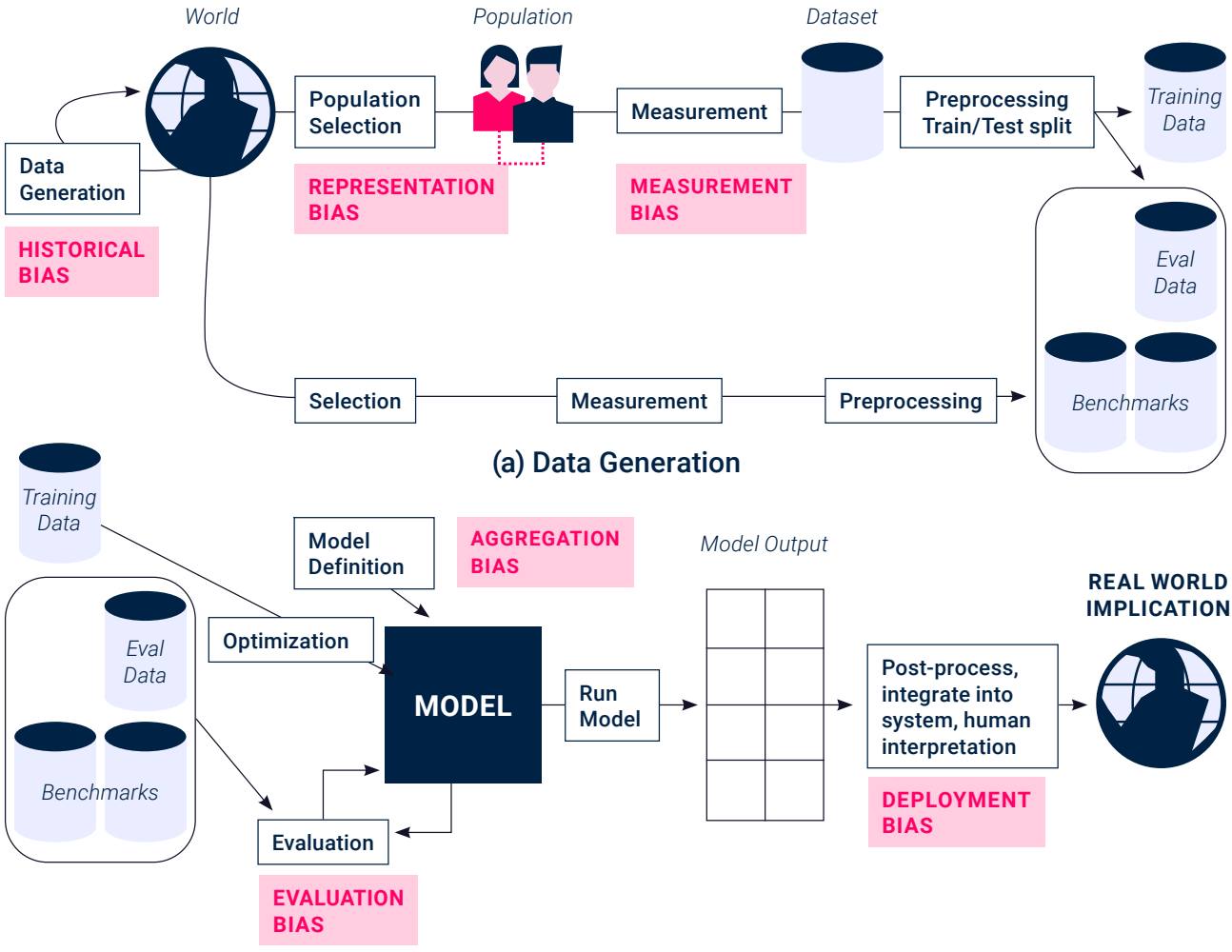
Artificial Intelligence is a set of theories and techniques that allow a machine to make decisions when faced with a predefined situation. In other words, it simulates human intelligence. Such programs can either be:

Deterministic: successive rules are applied according to a decision-making scheme.

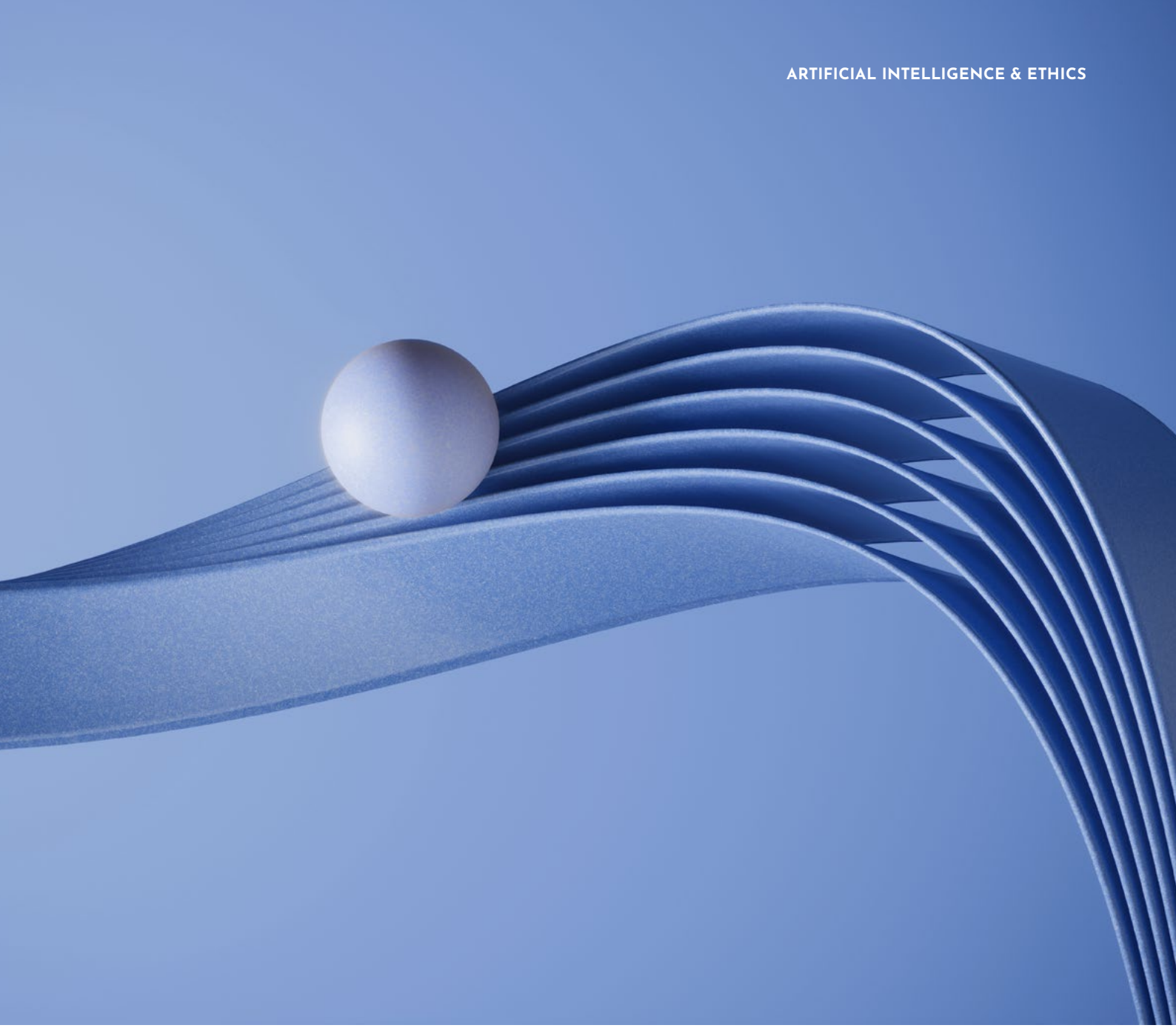
Probabilistic: after learning from similar past situations, the program can infer what the 'right' decision should look like. This is also called machine learning.

Regardless of the program type, AI is the automatization of decision-making. The ethical risks do not reside in the fact that AI may automate badly, but on the contrary, that it perfectly mimics human decisions, with its burden of errors and biases.

These dysfunctions come from the cultural context in which the AI was developed, when it replicates the historical value systems of a society (gender or ethnic inequality, for example) or those of its designers (education, political sensitivity, religious practices, etc.).



Source: [1901.10002] A Framework for Understanding Sources of Harm throughout the Machine Learning Life Cycle



These biases are particularly prevalent in machine learning algorithms since they are trained on legacy databases, inferring the future from the past.

To rectify training sets, it might seem logical to erase all traces of sensitive data. However, this would be as useless as it would be dangerous: on the one hand, because non-sensitive data can be insidiously correlated with sensitive data, and on the other because this type of shortcut encourages limiting regular checks on data and processes.

The great challenge of machine learning is to identify the cause of bias rather than to dismiss its consequences. This requires dedicated technological tools and reliable corrective methods.

Technological solutions can detect and measure biases all along the data processing chain: from collection to processing, including transformation and modeling. The solution must then be implemented throughout the product lifecycle, both at the design and development stage as well as in production, in order to ensure permanent correction.

There are other methods for correcting biases at each stage of data processing: collection, transformation, modeling and exploitation.

2. How to design trustworthy AI throughout the system lifecycle

The essential questions linked to conceiving and developing an AI lifecycle concern both technical and ethical issues:

- How to detect potential biases in an AI system?
- How to interpret a model's results?
- Can we accurately detect an AI model's biases?
- Can we measure the evolving performance of an AI system?
- Can we react to a malfunction? If not, what can we do?
- Can we perfectly track used data throughout the data processing journey?
- Can we organize an archive of results and different versions of an AI system?
- What safety guarantees can we offer?

This non-exhaustive list reflects the diversity of challenges to be faced during the production of an AI. To meet them, many technical solutions have been put on the market.

For several years, IBM Research has been working on a range of technical products to ensure that future AI systems are equitable, robust, explainable, accountable, and conform to society's values throughout their lifecycles.

Among the solutions developed by IBM Research are four toolboxes that bring a technical solution to the four dimensions linked to the design and lifecycle of AI systems:

- **Explainability through AI Explainability 360**
- **Equity with AI Fairness 360**
- **Robustness with the Adversarial Robustness toolbox**
- **Transparency thanks to AI FactSheets 360**

These open-source projects are made freely available for all through the LF AI & Data Foundation, a Linux Foundation entity that supports open-source innovation for AI and Data.

1 - Explainability

High-performing AI systems, such as artificial neural networks, are very efficient and used in multiple use cases. However, they are difficult to interpret, which is why they are often referred to as "black-box models."

To solve this issue and bring explainability functionalities to these models, IBM Research's AI Explainability 360 toolkit brings together ten algorithms and methods for interpreting datasets and machine learning models.

35 % - 50 %	Reduced model monitoring effort
Increased models in production	3 - 8
15 % - 30 %	Increased accuracy of models

Source: New Technology: The Projected Total Economic Impact™ Of Explainable AI And Model Monitoring In IBM Cloud Pak For Data, Forrester, 2020.

"The fairness of an AI depends on the accuracy of the data used to train it."

2 - Equity

To ensure that a machine learning model is not a source of discrimination, good practices must be followed, teams must be diverse, and ad hoc tools must be employed. IBM's **AI Fairness 360** integrates 70 fairness metrics and 10 bias mitigation algorithms developed by the research community.

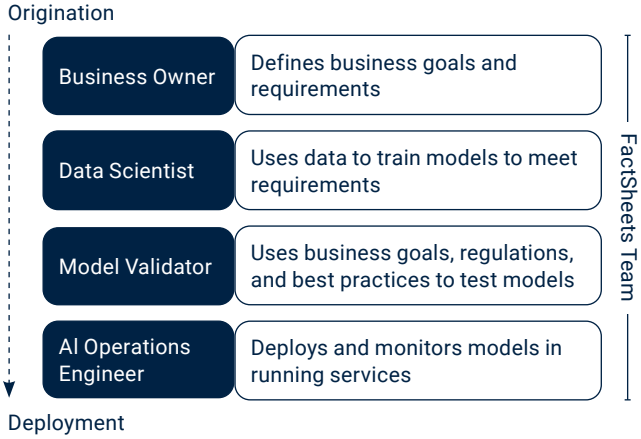
3 - Robustness

Sources of malfunction in Artificial Intelligence models can be accidental, when data is corrupted, or intentional, when caused by hackers for example. Both of these sources of bias lead AI models into error by providing incorrect predictions or results.

The Adversarial Robustness toolbox provides tools that enable developers and researchers to evaluate, defend, and verify machine learning models and applications against conflicting threats that might target the data (data poisoning), the model (model leakage) or the underlying infrastructure, both hardware and software.

A contradictory entry, superimposed on a typical image, can lead a classifier to wrongly categorize a panda as a gibbon.

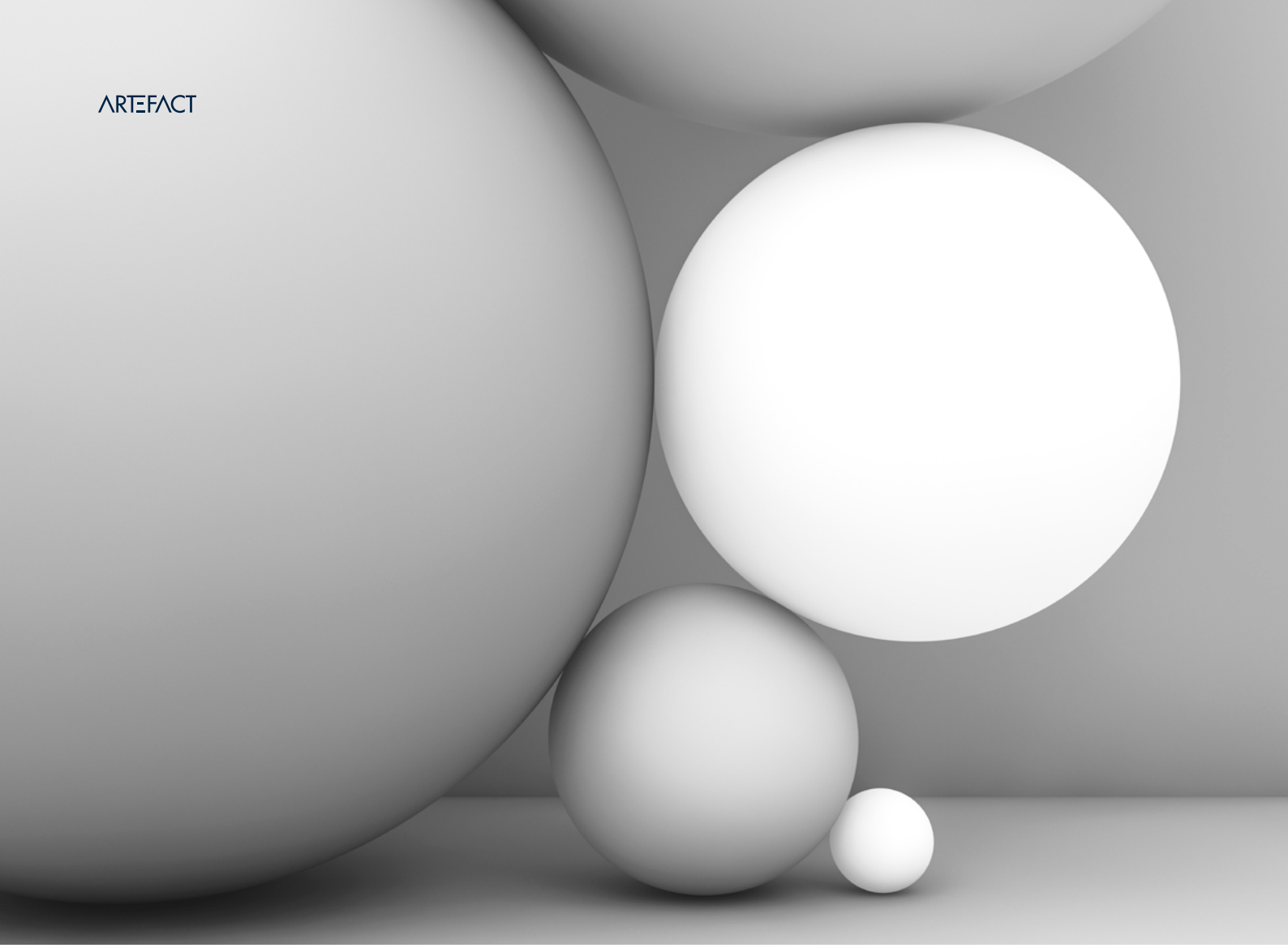
Les rôles clés dans le cycle de vie d'un modèle IA



Source: A Methodology for Creating AI FactSheets
https://arxiv.org/pdf/2006.13796.pdf

4 - Transparency

Governance in the lifecycle of AI models is fundamental. It allows for the specification and application of precise policies in product design and deployment. This helps avoid undesirable situations, such as the use of unapproved data, or models that are biased or have unexpected performance variations. To structure these processes, **AI FactSheets 360** brings greater transparency to all stakeholders involved in the AI lifecycle with information sheets that collect and collate data on the system in one place.



IV – Change Management: Training employees, adopting dedicated governance and systematically documenting processing

The concept of trustworthy AI cannot be reduced to its legal and technical conception. Remediation actions are partly linked to the organization and transformation of the company. The seven principles described above are a good example: they must be understood by all stakeholders, implemented throughout the creation process and followed over time.

Too often ignored or underestimated, the human and organizational aspect is crucial to the success of an ethical approach to Artificial Intelligence. This means transforming the company's culture in order to profoundly integrate ethical issues and approaches. thereby guaranteeing the sustainability of solutions by all stakeholders and contributing to ethical AI by design.

Changing the way AI systems are designed and operated involves all layers of the organization and the governance of a project. For example, it involves assigning new responsibilities, creating new positions, recruiting... It also requires training all employees on both the ethical issues and the processes and methods used to guarantee them.

Change management requires the implementation of many actions and procedures, such as:

- Establishing roles and responsibilities in the organization of AI-based projects
- Implementing a governance structure
- Defining the scope of use of AI and documenting its systems
- Bringing diversity to project teams, from product design through testing
- Drafting an ethical and environmental policy, signed by all employees
- Creating a chain of command in case of AI malfunctioning
- Building and distributing training and awareness campaigns on the importance of trustworthy AI.

Governance processes must include comprehensive documentation of AI processing. This can include:

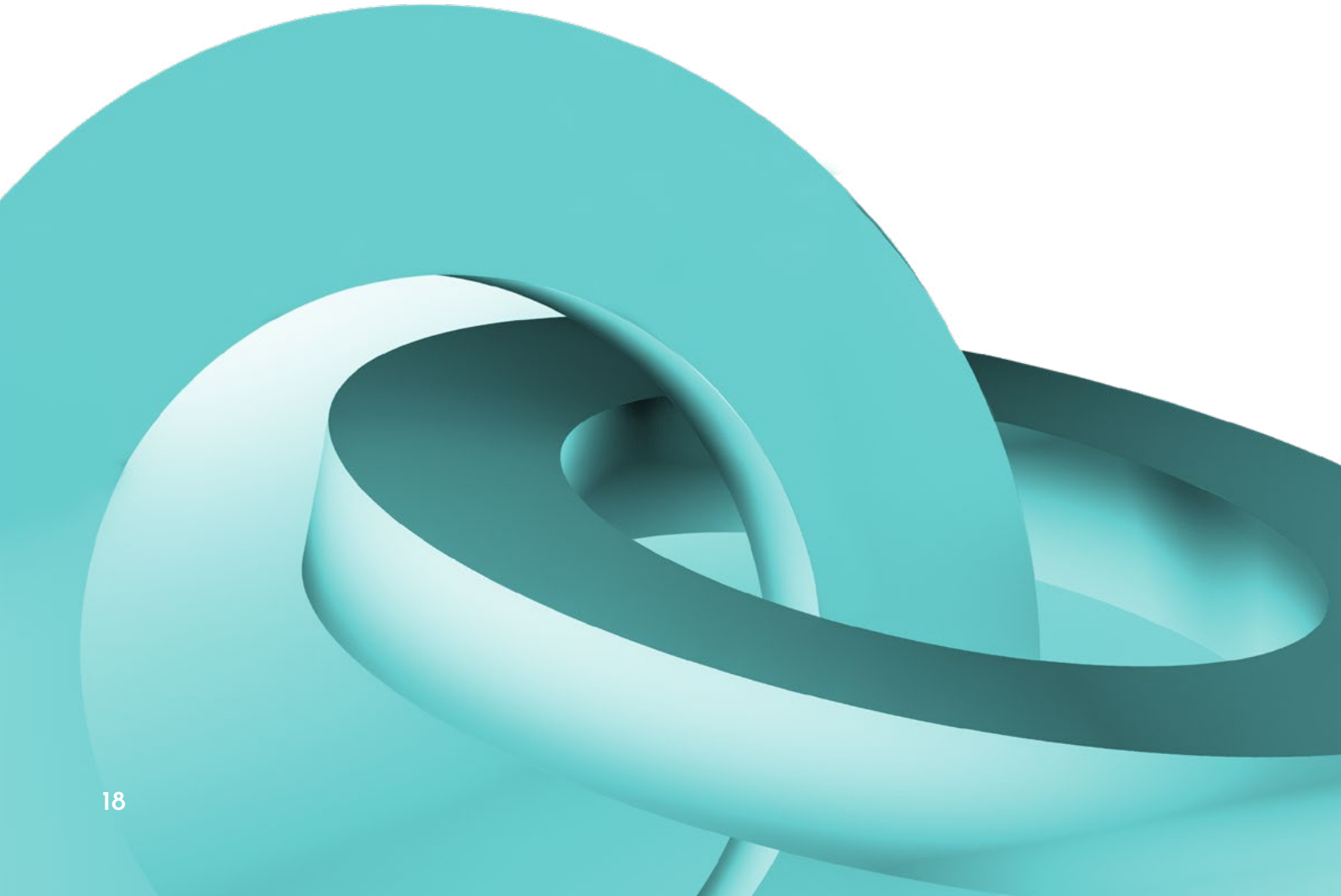
- **The different AI models and systems designed:** What data is used? What are the functional outputs and the practical outcomes of the system?
- **The inherent utilization of the AI systems:** Within which bounds and use-cases can these AI systems be used? What needs do the AI systems fulfill?
- **The protocol to follow in case of a malfunction:** Should the system be stopped? Are there alternative systems? How and in which cases should users be alerted?

Since 2016, IBM has placed an Ethics Committee at the center of its trustworthy AI initiatives. This group forms a central body for all processes relating to governance, review and decision-making for policies, practices, communications, research, products, and services pertaining to ethics at IBM. Each department has "focal points" whose role is to support the work of the committee and engage local teams around compliance with IBM's Principles of Trust and Transparency. Local teams and focal points can rely on the committee's advice and action plans for support. There is also a network of volunteers that helps promote an ethical, responsible and trustworthy culture within the company.

V – Developing less human, but more humanistic systems

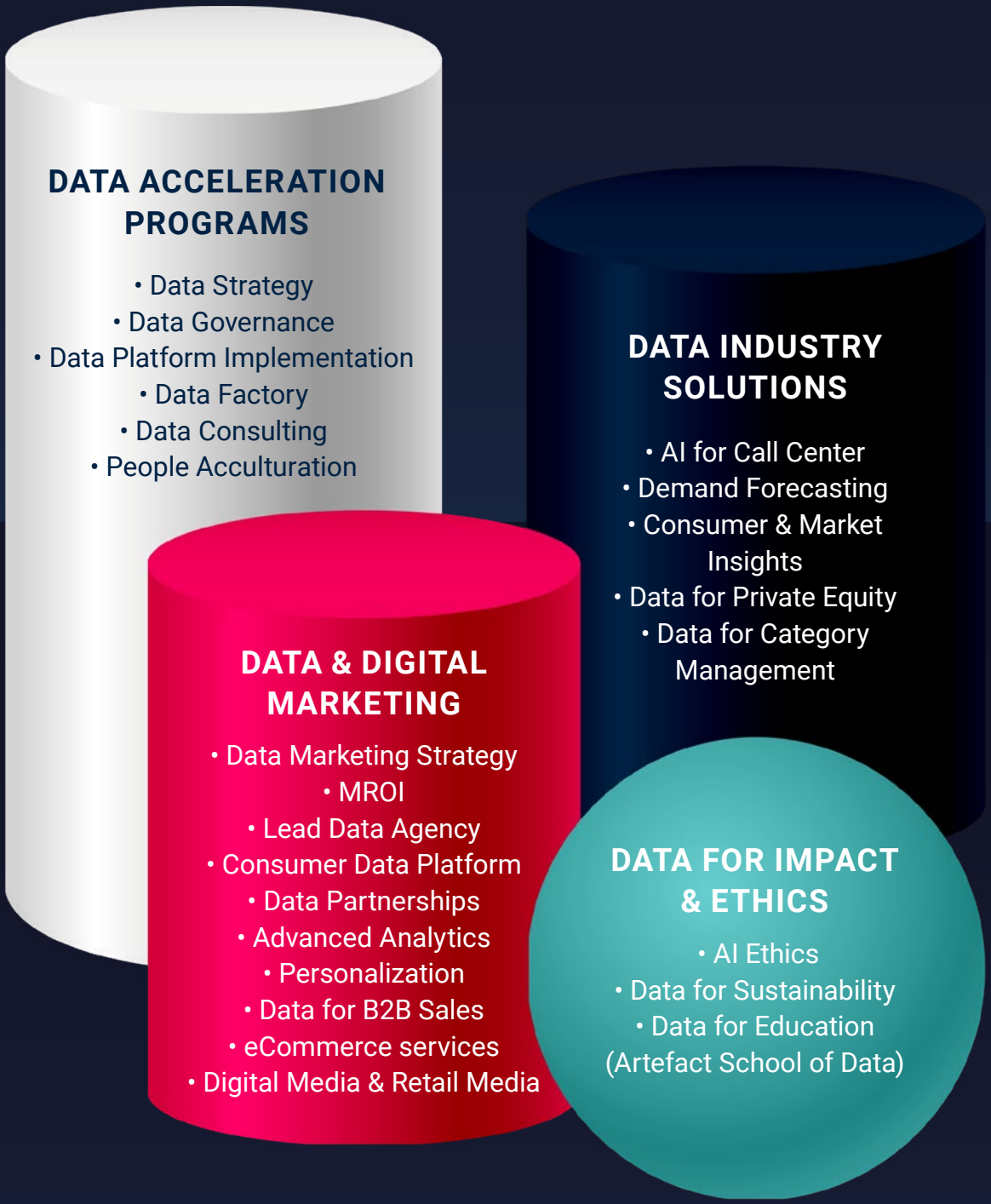
A number of initiatives must be rapidly defined and implemented if we are to respond to the issues raised by trustworthy Artificial Intelligence. This is all the more urgent given the ongoing development of the regulatory framework expected to come into force in 2023.

Because we believe that the design and operation of trustworthy AI depends first and foremost on the empowerment of organizations, IBM and Artefact are joining forces to support companies in their efforts to achieve trustworthy AI. These organizations will benefit from the open-source tools developed by IBM for explainable, fair, robust and transparent AI systems, and from Artefact's expertise in preparing their compliance with regulations and defining new governance.



ARTEFACT

Every company talks about data. At Artefact, we don't talk, we act.



ARTEFACT
VALUE BY DATA

CONTACT

hello@artefact.com
artefact.com/contact-us

ARTEFACT HEADQUARTERS

19, rue Richer
75009 – Paris
France

artefact.com

Artefact and IBM are partnering to make machine learning results more tangible and effective by combining predictive and prescriptive capabilities.

Silver
Business
Partner

IBM