dihnamic

For the deployment of a digitized, resilient, and ethical production industry in Nouvelle-Aquitaine Region in France

Trustworthy AI Guidelines

2025

Cofinancé par l'Union européenne

RÉGION Nouvelle-Aquitaine

Projet cofinancé par la Région Nouvelle-Aquitaine

Introduction

For a Trustworthy Artificial Intelligence innovation

> A dynamic context for Artificial Intelligence and innovation

The adoption of the AI Act in 2024 marked a turning point in the field of Artificial Intelligence (AI)[1] at the global and European level.

The adoption of this historic **legal framework** to frame the development and use of **AI and Trustworthy AI systems** in Europe illustrates a desire to reassure and unite stakeholders around a shared vision in a field where recent exploits were still unimaginable two years ago.

Among these stakeholders are political decision-makers faced with the need to understand AI and Trustworthy AI to legislate and make informed decisions, business leaders whose services and revenues are challenged by the new AI-based tools, employees whose professions are impacted, researchers at the heart of innovations, and the general public, who is the end consumer of products and information generated or not by AI.
This technology, with its high growth potential, also has the potential to disrupt the day-to-day activities of companies and public authorities, from the individual to the structures themselves.

This disruption can arise from the complex ethical challenges that emerge from the earliest phases of ideation to the design and deployment of AI-based systems. These include, but are not limited to, data traceability and governance, cybersecurity, hallucinations and biased behavior of AI-based systems, lack of transparency and explainability, manipulation for malicious purposes, and misinformation.

The AI Act can be considered a structuring legal framework for innovation in AI, as well as for AI itself, to ensure that it is, henceforth, responsible, transparent, and sustainable.

> Dihnamic, a trustworthy innovation player supported by the Nouvelle-Aquitaine region and Europe

In this context, **Dihnamic** (Digital Innovation Hub for Nouvelle-Aquitaine Manufacturing Industry Community) **has been supporting the innovation of public authorities and manufacturing and industry service companies since 2022**. Its expertise is rooted in four technology nodes—AI, Internet of Things (IoT), Robotics, and Digital Twins— and aims to promote **ethical AI** through **the development of Trustworthy AI** systems.

"Trustworthy AI" and "ethical AI" should be understood here as emerging properties of a system in a business and application context linked to a level of risk.
These properties need to be characterized from the earliest stages of innovation projects. As technical approaches alone are insufficient to achieve a coherent and adequate level of trust, it is essential to employ a multidisciplinary innovation strategy that considers the business context and associated risk.

[1] The glossary at the end of this document provides the definitions discussed by the "Artificial Intelligence" and "Responsible Digital" expert hub for this document

> Our guidelines to raise awareness of Trustworthy AI

To accompany this critical stage of AI projects, Dihnamic proposes **guidelines to raise awareness of Trustworthy AI, presenting simplified yet robust recommendations on essential values**, such as human control, security, data governance, transparency, sustainability, and fairness.

These guidelines are the result of :

- the consultation with a hub of experts specialized in AI, ethics, and responsible digital,
- the exchanges with over 200 companies and partners in the Nouvelle-Aquitaine ecosystem,
- the review of numerous open-source methodological and technical works and productions by the technical and scientific communities on Trustworthy AI and related themes,
- the work of regional programs[2], and significant national programs run by French and European public-sector organizations on the Responsible Digital Economy and the AI Act,
- Dihnamic partners' expertise in cutting-edge scientific and technical innovations.

Through **eight recommendations** detailed in this document, the guidelines highlight the essential axes for **a "Trustworthy AI-company" collaboration** that respects the rights and well-being of employees in both the public and private sectors.
For each of these, Dihnamic's partners offer concrete services with complementary expertise, including acculturation, training, consulting, prototype development, and funding research, to meet the tailored needs of each company or public authority.
Through these guidelines, Dihnamic aims to bridge the gap between the new European regulations and the logistical, financial, and industrial realities of manufacturing companies by creating a **shared vision of Trustworthy AI that federates the entire innovation ecosystem**.

The Inria center of the University of Bordeaux
Trustworthy AI manager of the Dihnamic project
for Dihnamic partners
April 2025

[2] We can cite three regional actions: (i) The Chaire IA digne de confiance, an industrial chair on Trustworthy AI founded in 2023 by KEDGE, the University of Bordeaux Foundation, and the Enseirb Matmeca -Bordeaux INP engineering school, (ii) the Numérique Responsable guidelines and the research work of the Institut du numérique responsable in La Rochelle, as well as (iii) the 3IA chairs led by ANITI and the University of Toulouse for "acceptable AI", "certifiable AI" and "collaborative AI". For further information, please refer to the "Additional resources" section

Dihnamic expert Hub

Many thanks to the high-level AI expert group, as presented below, for its contributions to the discussions, analysis, and advice on the guidelines and recommendations.

Catherine Tessier

Research Director, ONERA

Scientific Integrity and Research Ethics Referent

Member of the National Digital Ethics Steering Committee (2019-2024)

Claude Kirchner

Emeritus Research Director, INRIA

Chairman, Comité Consultatif National d'Éthique du Numérique

Frédéric Alexandre

Research Director, INRIA

Head of the Mnemosyne team

Researcher in AI, cognitive modeling, and computational neuroscience

Laurent Simon

University Professor at Bordeaux INP's ENSEIRB-MATMECA engineering school and LaBRI researcher at Bordeaux University

Regional Councilor,
CESER de Nouvelle-Aquitaine

CHAIRE industrielle IA digne de confiance

Michèle Barbier

European project coordinator, INRIA (digital twin)

Ethics expert for the European Commission and EU-funded projects

Member of the External Advisory Board of the Ocean Enterprise Initiative

Nicolas Roussel

Research Director, INRIA

Researcher in Human-Computer Interaction

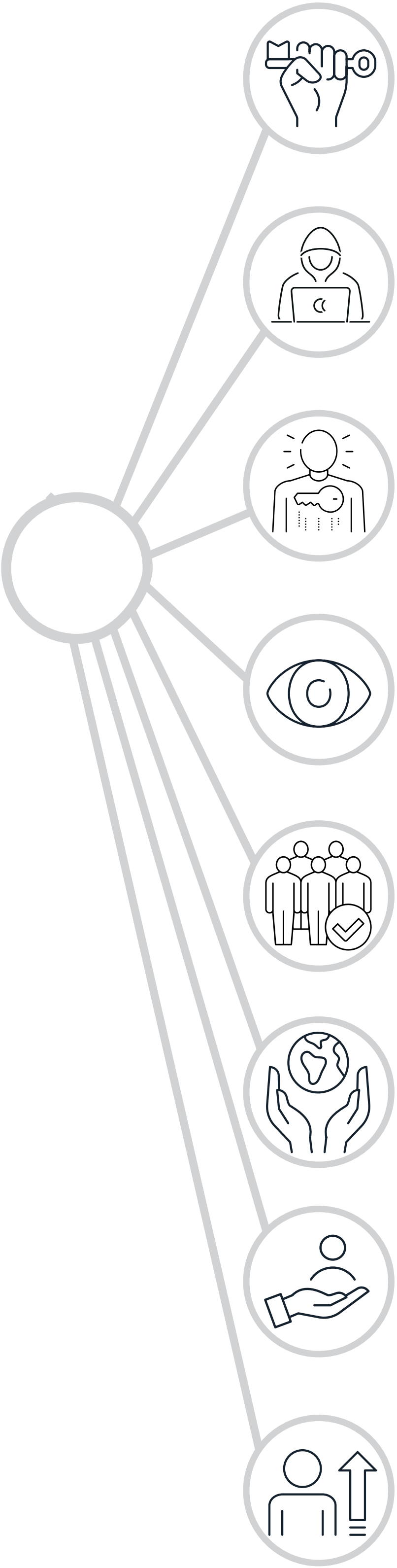
Director of the Inria Center at Bordeaux University

Pierre Etienne Legoux

Head of AI @ Thales Avionics

Dihnamic Guidelines

Trustworthy Artificial Intelligence & Digital Ethics



Human action and control

Includes fundamental rights, human autonomy

Technical robustness and security

Includes resilience to attacks and security, contingency plans and general safety, accuracy, reliability and reproducibility

Privacy and data governance

Includes data quality and integrity, access to data

Transparency

Includes traceability, explainability and communication

Non-discrimination, equity and diversity

Includes absence of unfair bias, accessibility and universal design, and stakeholder participation

Societal and environmental well-being

Includes sustainability and respect for the environment, frugality, social impact, society and democracy

Social responsibility and societal accountability

Includes auditability, minimization and communication of negative impacts, arbitration and appeals

Dignity

Includes absence of any kind of malice or harm towards human beings, protection of their physical and mental integrity, their sense of personal and cultural identity, and satisfaction of their basic needs

Extended edition

Human action and human control: autonomous decision-making at all times

To ensure respect for fundamental rights, humans must have the option to opt out of using an AI system (AIS) if they believe that ethical or safety standards are not being met. Governing AISs, protecting autonomy and decision-making, and developing human-centered designs with AI supervision and control are all essential.

Specifically, a company needs to:

- Establish protocols to define the AI context, levels of delegation, and the role of humans (identify human control points). Ensure ongoing monitoring;
- Create a process for detecting, alerting, and correcting failures (such as violations of fundamental rights or risk situations);
- Develop a process for reclaiming control of the AIS if failures occur;
- Set up emergency access points into the AIS for correction, monitoring, and reporting to users.

Technical robustness and security

The reliability and reproducibility of decisions amid contextual variations and disturbances help build robustness and resilience against attacks (cybersecurity threats) and the unpredictability and volatility of the real-world environments where AISs are integrated.

In practical terms, it means to:

- Secure learning by identifying AI system vulnerabilities (including application cases and beyond) and creating a checklist;
- Develop protocols for testing and monitoring system behavior throughout its life cycle to ensure that the risk of drift is minimized or eliminated;
- Establish robustness evaluation standards or metrics and define an acceptable error margin;
- Set performance thresholds indicating when the AIS has reached its operational limits and implement control measures when data is compromised;
- Maintain system stability by ensuring reliable results and physical protection against unforeseen events;
- Monitor and manage all risks and anticipate environmental changes with appropriate procedures communicated upstream.

Privacy and data governance

The protection, quality, and integrity of private data, as well as data governance in compliance with the European Union's Data Governance Act[3] ensure the principle of preventing breaches of personal privacy.

In concrete terms, this means:

- Ensuring legal and regulatory compliance for personal data processing, in line with the General Data Protection Regulation, RGPD (legally binding controls), and the Data Governance Act / CNIL: anonymize data and perform regular monitoring to maintain the data's anonymous nature over time.
- Making sure that the legal and regulatory conditions for collecting, storing, processing, and sharing data are clear and transparent to the individuals whose data the AIS processes. Obtain their consent.
- Implementing data governance, including traceability methods for the data lifecycle, from collection and creation to use, along with verification methods, procedures, and operations that impact or utilize the data.

Transparency: traceability and explainability are essential for understanding, explaining, and justifying results

Being able to justify, understand, and interpret the models and methods underlying the AIS enables humans to clarify the system's results and behavior, thereby promoting adoption. Tracking the data used also helps clarify AIS behavior and results, making them transparent and easier to comprehend.

Specifically, it involves:

- Setting up and maintaining a repository where all AI-related information is stored, including model design, performance, results, as well as technical limitations and choices;
- Developing related documentation and creating layouts with graphical tools to visualize key criteria;
- Establishing internal standards and rules with regular reviews to anticipate risks;
- Justifying AIS operation by comparing AI functions with human reasoning in the same use case;
- Evaluating and documenting the trade-off between performance and transparency when implementing an AIS;
- Ensuring that suitable AIS explanations are generated and tailored to the needs of the target audience(s), and regularly assessing their acceptability and clarity.

[3] https://www.cnil.fr/en/artificial-intelligence-opinion-cnil-and-its-counterparts-future-european-regulation (last acces : 11/13/2025)

Non-discrimination, equity, and diversity: ensuring an ethical AIS for everyone

Depending on business and technological constraints, as well as the level of risk associated with AIS, it is advisable to limit potential biases that could harm a given population. If this is the case, it must be documented and the affected stakeholders informed. Additionally, to guarantee equitable access and prevent any human harm, it is also essential to ensure AIS accessibility, universal design, and the participation of all stakeholders potentially involved or impacted by AIS throughout the software development cycle, from the initial ideation phases (needs analysis) to deployment.

In concrete terms, this means:

- Implementing methodologies for detecting and limiting statistical biases in data sets through verification and validation processes for bias-related adjustments and decisions;
- Raising awareness among teams about the cognitive biases involved in AIS ideation, design, and testing;
- Selecting and justifying the type of equitable transformation to adopt when AIS influences decision-making for a given population, by collaborating with specialists in the social sciences and/or statisticians to develop an ethical framework;
- Practicing and publishing regular self-assessments to ensure data scientists' work complies with regulations, corporate values, and technical constraints.

Societal and environmental well-being: two core missions of any responsible AIS

Sustainable and environmentally friendly AI must be aware of its potential harm to society and democracy, while incorporating energy efficiency to reduce its environmental impact. AI applications should contribute to building a desirable future for the company, its consumers, and its citizens.

In practical terms, a company needs to:

- Identify AIS risks to users' cognitive functions and mental health and eliminate or minimize them;
- Recognize and define the socio-economic values to be respected in its development and use;
- Determine evaluation metrics and methodologies for assessing impacts, such as choosing suppliers, service providers, and subcontractors involved in the AIS lifecycle, monitoring carbon consumption, and implementing data archiving, expiration, and deletion policies;
- Apply these measures, considering the technical constraints of the domain and the company, both at the algorithm level and across all cases of application;
- Educate teams and users about sustainable impacts on the environment, health, and socio-economics;
- Follow bioethics principles if the AIS has a healthcare mission.

Responsibility: constantly recognizing social and societal obligations

When a player in the production chain takes responsibility for an AIS, they also accept accountability for the outcomes. Transparency, documentation, and communication of negative impacts, along with arbitration and recourse, ensure autonomy and accountability related to AI systems and their results.

In practical terms, this means:

- Establish protocols to assess the context of AIS use, the level of delegation, and the role of the human, and create a process for identifying responsibilities in the AIS design and implementation chain;
- Implement protocols and incorporate them into all AI projects: appoint a "model owner" (or person responsible for the model) who will be accountable for the model or system in production;
- Determine liability criteria for each participant in the chain: designer, developer, supplier, host, other subcontractors, and end-user;
- Offer introductory and refresher training courses on digital technologies aligned with SIAs of cybersecurity, explainability, frugality, and human-machine interaction.

Dignity: respect for human dignity

The absence of malice or harm toward humans, the protection of their physical and mental integrity, personal and cultural identity, and the satisfaction of their essential needs are key to respecting human dignity. Generally, an AI system is considered ethical if it can uphold, throughout its entire life cycle, fundamental human rights such as dignity, mental and physical integrity, freedom, autonomy, fair treatment, intimacy, and privacy.

In practical terms, this means:

- Assessing and qualifying AI systems based on fundamental rights;
- Educating consumers and employees about the capabilities and limitations of the AI systems they use, and obtaining their consent for automation and its purposes;
- Creating an evaluation guide for end-users of AI systems;
- Implementing an accessible, understandable, and customizable information message for end-users.

Glossary

Artificial Intelligence

In the context of these guidelines, Artificial Intelligence is defined as any tool used by a machine capable of "reproducing human-related behaviors, such as reasoning, planning, and creativity". The applications of AI are vast, as they extend to many aspects of daily life, permeate all areas of society, and pose a technological challenge that affects the economy, research, and education. Consequently, this requires a set of measures and approaches to ensure trustworthiness.

Responsible digital

Responsible digital is an ongoing approach that tackles the ecological and social footprint of digital technology. This includes Green IT, which seeks to minimize the environmental impact of IT departments, and IT for green, which uses digital technology to support sustainable development and the responsible design of digital services.

In summary, the goal of implementing a Responsible Digital approach is to:

1. Reduce environmental impact;
2. Enhance social impact;
3. Serve as an economic driver;
4. Foster innovation;
5. Encourage employee and corporate engagement through Corporate Social Responsibility (CSR) initiatives.

How to cite this work?

EDIH Dihnamic (April 2025) “For a digitized, resilient and ethical production industry in Nouvelle-Aquitaine, Trustworthy AI Guidelines” 2025

For further information

Below are some additional, non-exhaustive resources in French that have helped us reflect on and develop these guidelines. (Last access date for all links: 10/24/2025).

At the regional levell

- The industrial chair on trustworthy AI, established in 2023 by KEDGE BUSINESS SCHOOL, the Fondation Université de Bordeaux, and Bordeaux INP's ENSEIRB Matmeca engineering school: https://www.fondation.univ-bordeaux.fr/projet/chaire-ia-digne-de-confiance
- The work and guidelines of the Institut de Recherche en Numérique Responsable in La Rochelle: https://institutnr.org/charte-numerique-responsable
- The 3IA Chairs supported by ANITI and the University of Toulouse, focusing on acceptable AI, certifiable AI, and collaborative AI: https://aniti.univ-toulouse.fr/chaires-3ia/
- IA Cluster ANITI chairs renewed in 2024: https://aniti.univ-toulouse.fr/les-chaires-ia-cluster-aniti/

At the national level

- Circular no. 6425-SG, dated November 21, 2023, outlines the State's dedication to ecological transformation: https://www.legifrance.gouv.fr/download/pdf/circ?id=45511
- The MiNumEco website of the interministerial mission for eco-responsible digital development: https://ecoresponsable.numerique.gouv.fr/a-propos/
- Numeum's practical guide to ethical AI (2021): https://ai-ethical.com/wp-content/uploads/2021/09/2021-SN-Guide-Me%cc%81thodo-IA-Ethiques-version-imprime%cc%81e.pdf
- The Hub France IA white paper (2023): https://www.hub-franceia.fr/wp-content/uploads/2023/05/Livre_Blanc_IA_Ethique.pdf
- The Confiance.ai program white paper: https://www.confiance.ai/wp-content/uploads/2023/09/LivreBlanc-Confiance.ai-Octobre2022-1.pdf

At the European and international level

- Ethical guidelines published by the European Commission's High-Level Expert Group in 2018: https://digital-strategy.ec.europa.eu/en/library/ethics-guidelines-trustworthy-ai
- Recommendations in European legislation on AI and the AI Act: https://digital-strategy.ec.europa.eu/fr/policies/regulatory-framework-ai
- UNESCO recommendations on ethics in AI: https://www.unesco.org/en/artificial-intelligence/recommendation-ethics

www.dihnamic.eu

DIHNAMIC @dihnamic

contact Dihnamic
contact@dihnamic.eu

13 partners at your service

Partenaires associés :

Agri Sud-Ouest Innovation | Cosmetic Valley |
École Nationale Supérieure d'Arts et Métiers (campus Bordeaux-Talence) |
Pôle Européen de la Céramique | Université de Bordeaux

Cofinancé par :

Cofinancé par l'Union européenne

Projet cofinancé par la Région Nouvelle-Aquitaine

Project n° 101083886 Dihnamic

__Disclaimer__ : Dihnamic project is co-funded by the European Union and the Nouvelle-Aquitaine Region. Views and opinions expressed are however those of the author(s) only, and do not necessarily reflect those of the European Union. Neither the European Union nor the granting authority can be held responsible for them.