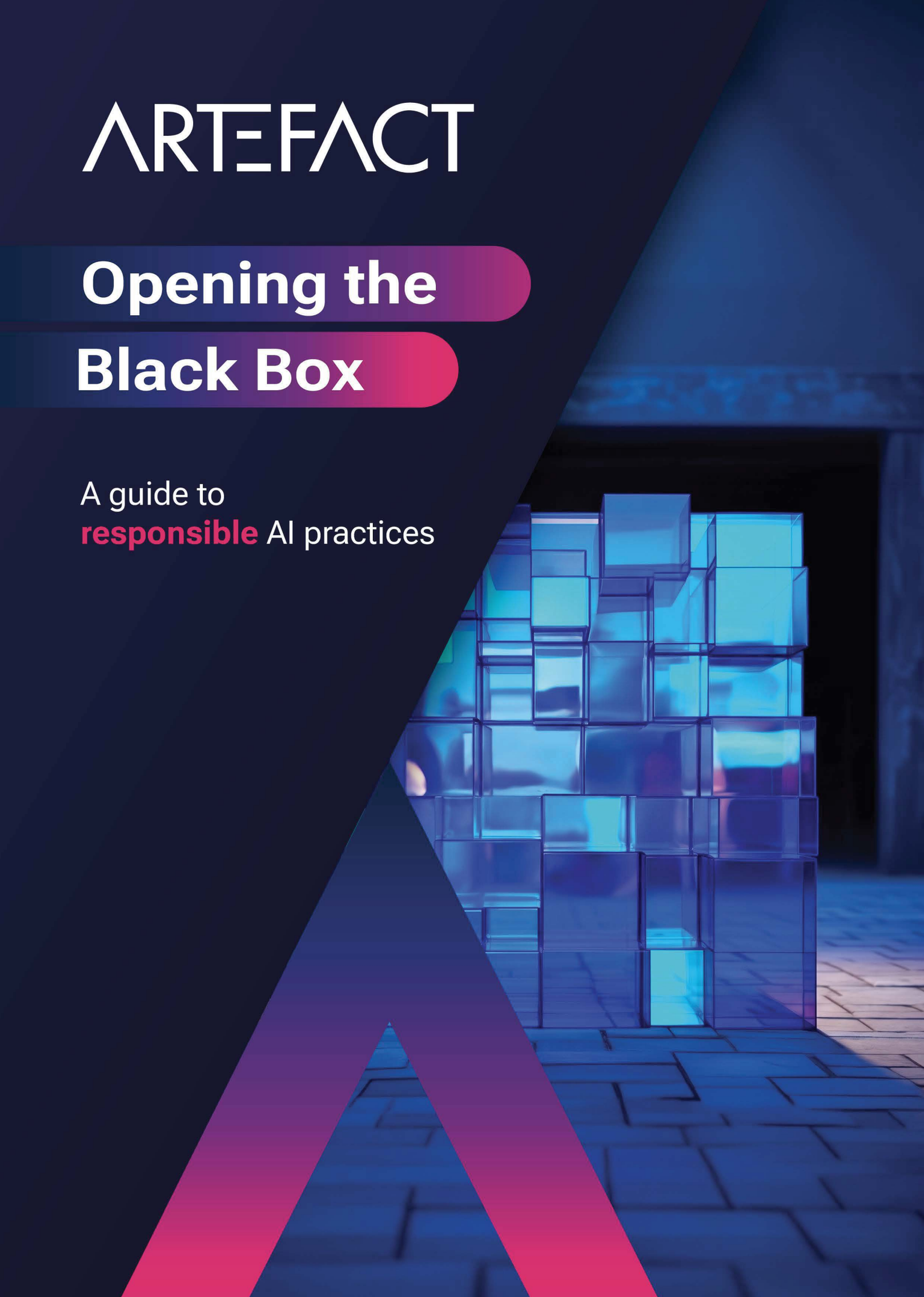
ARTEFACT

Opening the Black Box

A guide to **responsible** AI practices



ARTEFACT

We make AI work. Anyone can buy the model. No one can buy your context. Discover how our 3,000 AI makers Shape, Build, Deploy, and Transform businesses with measurable impact on growth and productivity.



27 COUNTRIES 2,700 EMPLOYEES 1,000+ CLIENTS

About Artefact

Artefact is the AI deployment team to the world's leading companies. Guided by our vision "We make AI work", we build our clients' AI moat from decision to adoption. Operating across more than 30 countries, our 3,000 AI makers (industry strategists, applied AI engineers, AI builders, and adoption specialists) work as one team to shape, build, deploy, and transform businesses with measurable impact.

FMCG • RETAIL & ECOMMERCE
LUXURY & COSMETICS • HEALTHCARE •
BANKING & INSURANCE • TELECOMMUNICATIONS

SPORTS & ENTERTAINMENT
TRAVEL & TOURISM • PUBLIC & GOVERNMENT
REAL ESTATE • MANUFACTURING & UTILITIES

Foreword

The promise of AI is enormous, and on the path to its large-scale impact, we will need to build trust in this technology.

A system that can be explained, governed and stood behind is one people come to rely on, both in the decisions that matter and at the scale where value is created. This reliance transforms a promising technology into a lasting capability.

We also know that AI trustworthiness can be built, because building AI is our work.

We design and deploy these technologies with organizations every day, and our research develops the methods that make them reliable. This work has given us one conviction: trustworthiness is built. It arises from deliberate choices made at every stage, from model design to deployment: what data a model learns from, whether its architecture allows its output to be explained, what it is evaluated against, who is accountable for its decisions, and what infrastructure it runs on.

These choices also change what leaders can expect from AI.

An organization that understands how its systems obtain their results can measure the value they create, identify and correct failures before they spread, and extend those systems to more consequential decisions with justified confidence.

Responsibility and performance are therefore a single program of work: the practices that make a system accountable are the same ones that make it reliable enough to scale.

This paper brings together what we have learned alongside practitioners, researchers and regulators shaping responsible AI in practice.

It is an attempt to make responsible AI concrete. Rather than restate what it should be, it examines what it takes in practice, from the data that feeds a model, through the governance that holds the whole chain together, to the environmental footprint that growth in AI now carries. It draws on the people who build, deploy, regulate, and audit these systems every day, and on research that turns principles such as explainability and fairness into usable methods.

This is a moment to be ambitious in deciding which AI we want for our future society.

The opportunity belongs to those who build trust in AI early. Systems are growing more capable and more autonomous, and the organizations that master transparency, governance, and control now will move faster, earn deeper trust, and scale that advantage over time.

We hope this paper helps you take that step.



Emmanuel Malherbe

— Director, Artefact Research Center

ARTEFACT



Médéric Chomel

— Managing Partner,
Group Responsible AI Lead

ARTEFACT

Who should read this paper?

This white paper is intended for Executive Committee and board members, Chief Data and AI Officers, and leaders across risk and compliance. More broadly, it is designed for readers who need to understand what responsible AI usage entails in practice, across the full AI value chain, from data and models to governance, decision-making, and environmental impact. Ultimately, it serves as a practical guide to translating that understanding into concrete solutions drawn from state-of-the-art research and applied engineering.

Opening the BLACK BOX

A guide to responsible AI practices

CONTENTS

Foreword 3
Executive summary 5
Introduction 6
Methodology 7

1. Ethical AI: embedding transparency and fairness into model design 8
   - 1.1 AI transparency as the foundation of organizational responsibility
   - 1.2 AI fairness is becoming a core requirement for AI alignment

2. AI governance: establishing trust across the model lifecycle 14
   - 2.1 A robust evaluation framework can protect organizations from AI risks at scale
   - 2.2 Maintaining control over AI-empowered processes requires model explainability, data lineage and auditability

3. AI and the Environment 20
   - 3.1 Strategic AI deployment under environmental constraints
   - 3.2 Matching model size to the value of each use case
   - 3.3 Decarbonizing the infrastructure underpinning AI

4. Responsible AI at Artefact 28
   - 4.1 Our commitments
   - 4.2 The Artefact Research Center

Special thanks and acknowledgments 30

AUTHORS



Oualid Mokhantar
Research Scientist
ARTEFACT



Emmanuel Malherbe
Director, Artefact Research Center
ARTEFACT



Lucie Filiol de Raimond
Group CSR Manager
ARTEFACT

EXECUTIVE EDITOR Emmanuel Malherbe, Director, Artefact Research Center

Executive summary

Six findings, from human accountability to the footprint of AI infrastructure.

01

Responsibility cannot be attributed to an AI system: it is inherently human.

Whatever the level of automation, a person or an organization must remain accountable for every decision a model informs or executes, and for the safeguards around it. Human in the loop is therefore not a technical option but a governance principle.

Model accuracy alone does not establish trustworthiness.

02

Technical benchmarks cannot determine whether a system is suitable for a particular deployment. Organizations should evaluate model capabilities, robustness, systemic risks, regulatory compliance, cybersecurity, and suitability for intended use, with appropriate human oversight.

03

Transparency increasingly depends on verification and operational safeguards.

As AI systems become more complex, organizations cannot blindly rely on AI outputs. Source traceability, systematic output verification, and hallucination detection are therefore essential.

OUR RECOMMENDATION

The Artefactual package provides a practical method for identifying potentially unreliable outputs of LLMs.

Discrimination by AI, also called fairness, is a damaging risk that persists even when the AI has no access to sensitive variables.

04

Models can reproduce disparities through proxy variables, thus leading to demographic groups receiving unequal treatment. Organizations should integrate fairness constraints into model training and assess disparities across affected populations.

OUR RECOMMENDATION

OT-FairBoost and FairGroups packages offer complementary methods for mitigating bias and identifying disadvantaged groups.

05

AI explainability enables accountability.

Organizations should establish clear responsibilities for reviewing, challenging and auditing model behavior. A black box AI is impossible to be assessed or judged by an internal or external party, meaning human interpretability is necessary for accountability.

OUR RECOMMENDATION

The WoodTapper package provides an auditable overview of an AI prediction mechanism, which can thus support business users, risk functions, and auditors.

Beyond generalized efforts to reduce AI costs and consumption, the environmental impact is significantly increasing.

06

Rebound and induction effects can offset computational savings, while infrastructure expansion increases demand for energy, materials and water. Organizations should prioritize AI applications according to their value and environmental footprint, monitor absolute resource consumption, and integrate lifecycle assessment, low-carbon construction, electricity sourcing, and efficient cooling into infrastructure decisions.

Introduction

Artificial intelligence is becoming more capable and more deeply integrated in business processes.

Advances in general purpose models have extended AI from single-task prediction to systems that execute multi-step tasks, retrieve and process information from large and complex data sources, interact with external tools and operate with growing autonomy within enterprise environments. Adoption has expanded accordingly. In 2025, 88% of surveyed organizations reported using AI in at least one business function, and the number of incidents recorded in the AI Incident Database rose from 233 in 2024 to 362 in 2025 (Stanford HAI, 2026).

At the same time, regulatory frameworks are moving from principles to requirements.

The EU AI Act, in force since August 2024 with obligations applying in stages, requires providers and deployers of high-risk AI systems to implement risk management, data governance, technical documentation, record keeping, human oversight, and post-market monitoring. It also sets specific obligations for providers of general purpose AI models. Similarly, the OECD Due Diligence Guidance for Responsible AI sets out a risk based process through which enterprises identify, prevent, mitigate, and account for adverse impacts across the AI value chain (OECD, 2026).



These developments increase both the business value of AI adoption and the importance of managing its risks.

AI systems inform or execute decisions in high-stakes domains such as finance, healthcare, employment, consumer services, and public administration.

Their outcomes depend on (1) the data used to train and operate them, (2) the processes in which they are embedded, and (3) the human oversight applied to them. Risks can arise at each stage of the AI system lifecycle, including post-deployment, as data and uses may change over time.

As a result, a growing share of firms are formalizing their commitments:

The share of businesses with no responsible AI policy fell from 24% in 2024 to only 11% in 2025 (Stanford HAI, 2026). However, responsible AI practices continue to lag behind the pace of rapid AI deployment, and evidence on how organizations translate principles into technical tools, operational processes, and governance arrangements remains limited.

This white paper examines how to implement responsible AI in practice, with lessons from organizations addressing it and recommendations of available tools.

It approaches responsible AI as a collective and technical capability: the combination of tools, practices and governance arrangements through which organizations evaluate AI systems, determine appropriate uses, allocate accountability, maintain human oversight, and adapt systems to a changing environment.

Plan for our paper: The analysis covers three areas. The first concerns model-level practices with a focus on transparency and fairness. The second concerns the governance and evaluation of AI systems, including regulatory requirements, auditability, and explainability. The third concerns environmental considerations regarding AI infrastructure and models.

Methodology

Responsible AI generally refers to the set of principles, practices, and governance mechanisms through which organizations seek to ensure that AI systems are trustworthy, explainable, fair, accountable, and environmentally sustainable.

Anchored in this conceptual understanding, for our analysis, we draw on semi-structured interviews with experts from nine organizations operating across finance, technology, infrastructure, AI consulting, academia, climate, sustainability, and public policy. The interviews examine how organizations approach Responsible AI, how they translate principles into operational practices, and how they allocate responsibility across technical, managerial and governance functions.

Participants were selected purposively according to three criteria: their expertise in a specific dimension of Responsible AI; the relevance of their organization to the development, deployment or governance of AI systems; and the diversity of perspectives required to examine Responsible AI across sectors and institutional settings. The selection includes perspectives from AI research, financial services, technology consulting, data center infrastructure, climate strategy, and public policy.

The interviews are analyzed alongside recent peer-reviewed research from the Artefact Research Center, published at the most selective international venues in artificial intelligence and machine learning, including the Association for the Advancement of Artificial Intelligence (AAAI), the European Conference on Artificial Intelligence (ECAI), the Journal of Open Source Software (JOSS), and the European Conference on Information Retrieval (ECIR).

Interview panel: ten experts across industry, finance, research and policy

E Environmental footprint of AI · S Ethical models: transparency and fairness · G Governance of models and workflows

	PROFILE	THEMES DISCUSSED	AXES
E1	Sustainability Director · Data center operator	Energy and water use of AI infrastructure; low-carbon data centers	E S G
E2	Research Director · National digital research institute	Model evaluation, fairness and bias; audit and AI regulation	E S G
E3	Chief Impact Officer · Business network for climate action	Corporate decarbonization; sustainable use of AI	E S G
E4	Manager · Climate strategy consultancy	Carbon footprint of AI; rebound effect; where AI use is justified	E S G
E5	Professor, Director of an AI research center · Business school	Governance of AI-enabled workflows; accountability; adoption	E S G
E6	Senior AI and Innovation Policy expert · International organization	AI policy and standards; adoption by firms; trust	E S G
E7	Managing Director · Global banking group	Model risk management; explainability; audit	E S G
E8	AI Research Scientist · Global consumer goods group	Detection and measurement of bias; fairness in practice	E S G
E9	Data Scientist, Chief of Staff Data Science · Private investment firm	Transparency of LLMs in investment workflows	E S G
E10	AI Research Scientist · AI research center	Interpretable machine learning; rule extraction for audit	E S G

1

Ethical models

How can organizations embed transparency and fairness in their models?

AI now shapes decisions in multiple domains including credit, hiring and pricing. Those decisions must be fair and transparent. This chapter sets out the importance of ensuring transparency and mitigating bias, and provides an easy to use package to help organizations improve the transparency of their AI use cases built on tabular data.

ESG PILLAR

E S G Social





IN THIS CHAPTER

1.1 AI transparency as the foundation of organizational responsibility 10

RECOMMENDATION 01 Visualize hallucinations in large language models with Artefactual 11

1.2 AI fairness is becoming a core requirement for AI alignment 12

RECOMMENDATION 02 Ensure that fairness is embedded directly into model training with OT-FairBoost 13

1.1 — AI transparency as the foundation of organizational responsibility

Providing transparency in AI recently reached a new level of complexity. Historically, statistical and classical machine-learning models could, at least partially, be rendered interpretable through access to their parameters and decision rules. In the era of neural networks and large language models (LLMs), this paradigm can no longer be leveraged, since these models do not admit such inspections. In fact, users of LLMs have no access to the training data employed, the biases that may have been introduced, the decisions taken, or the number of intermediate steps involved in the inference. The system operates as a black box, and users have only a general understanding of its functioning, rather than a deterministic view of why a given output was produced.



"You will never truly explain it. What we can do is put things in place around it; call them guardrails for instance, which allow us to reassure ourselves about how the AI will behave, what it can do, how it does it, and how we can qualify the results it produces."

Arthur Garnier
— IT Chief of Staff & Data Scientist, Ardian

ARDIAN

1.1.1 Transparency as verification and user ownership

In practice, transparency in AI based systems refers less to obtaining insight into the model than to the user's capacity to verify the results it produces. Practitioners readily acknowledge that the reasoning behind an answer cannot be fully inspected. A user can check whether the answer meets the need, whether the figures it contains are accurate, and whether any errors are apparent. Verification thus shifts from the process, which is inaccessible or intractable, to the result, which is not.

This practice rests on a principle of ownership. Work produced with the help of AI remains the user's work. As such, responsibility for what is passed on stays with the user rather than the tool. The principle has a practical purpose as well as an ethical one. AI-generated outputs can appear remarkably polished, and the habit of systematic checking is what prevents that polish from being mistaken for accuracy.

1.1.2 Transparency as operational assurance under model opacity

A second understanding of transparency is centered on the controls placed around the model rather than on the model itself.

Practitioners do not expect current architectures to become fully explainable. Their attention turns instead to safeguards that make outputs assessable even when the model remains opaque.

Source citation is the most frequently cited example. When a model indicates the document, page, or paragraph from which a figure is drawn, the user does not need to repeat the work in order to trust it. A single check against the original source is enough to confirm the figure or to reveal that something has gone wrong. Hallucination detection performs a complementary function. Rather than relying on the user to spot an error, it flags content that the system itself identifies as doubtful, before that content moves to the next stage of a process and becomes harder to trace.

#01 OUR RECOMMENDATION

ECIR 2026

Visualize hallucinations in large language models with Artefactual

The following section is based on an open-source package, "Artefactual: a Python package for detecting hallucinations in LLMs", jointly published with the following affiliations:

IN COLLABORATION WITH

ARTEFACT RESEARCH CENTER · CentraleSupélec · université PARIS-SACLAY · ARDIAN

The challenge

Large language models produce fluent, confident text even when the underlying facts may, in some cases, be wrong. For instance, in a financial workflow, a plausible but incorrect figure is more critical than a visible error, because it will be acted upon. Existing detection methods ask the model the same question several times and compare the answers. This works, but it multiplies computation and flags a full response without identifying which part of it is wrong, forcing users back into verifying everything.

The solution

Artefactual is a Python module that assigns a language model's answer a probability of being a hallucination. It reads the answer that has already been generated, together with the token probabilities returned alongside it, and needs nothing else from the model. At each generation step, the model computes a probability distribution over possible next words; when it hesitates, that distribution spreads out, and this spread correlates with the likelihood that the word is incorrect.

How to use it

01

STARTING POINT

Open weights or API

02

THE TOOL

Artefactual analysis of each generated token

03

THE BENEFITS

User sees when tokens have a high risk of being hallucinated

GITHUB.COM/ARTEFACTORY/ARTEFACTUAL

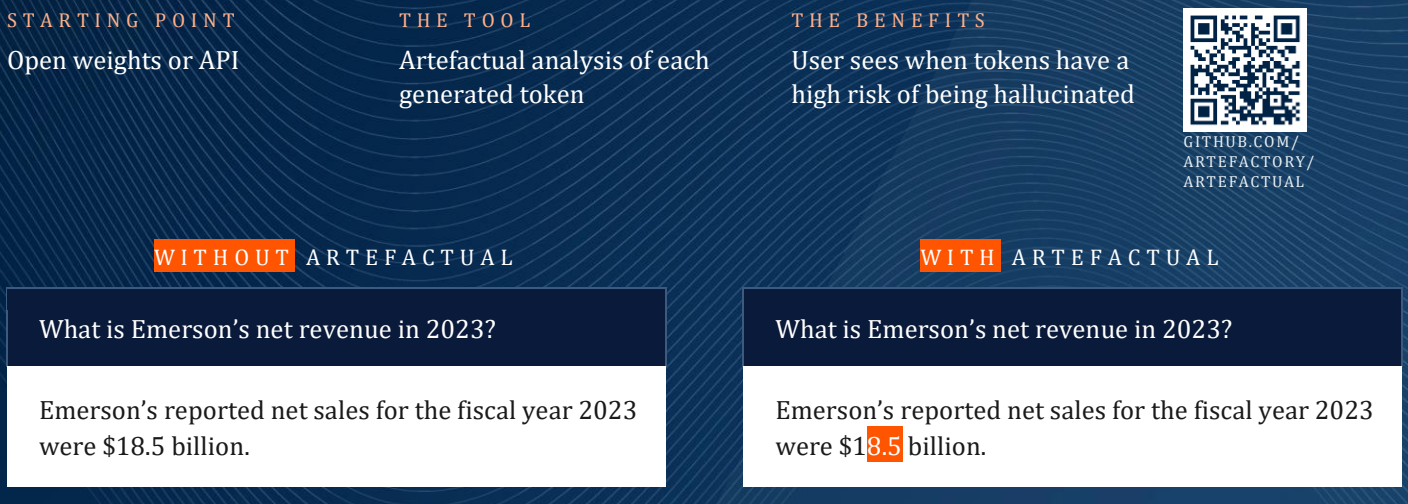


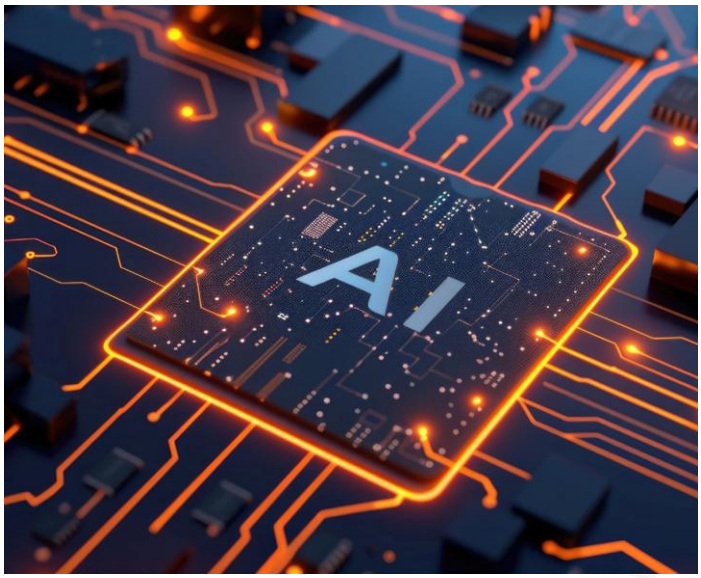
Figure 1: Artefactual flags the uncertain tokens of an answer, in a single pass of about 80 microseconds, so users verify the right part of it

1.2 — AI fairness is becoming a core requirement for AI alignment

Models built on customer, employee, financial, or operational data increasingly support decisions such as credit scoring, recruitment, pricing, and customer segmentation. Because these models are trained on organizational records, they can reproduce past commercial practices, market inequalities or historical exclusions while presenting their outputs as objective and data-driven.

These systematic disparities in approval rates, prices, or false positives can generate legal exposure, customer loss, reputational damage, and weakened employee or stakeholder trust. As such, businesses must improve fairness without undermining predictive performance, operational efficiency, or commercial objectives. Importantly, a model should not only produce fair outcomes, but its performance and errors should be acceptable across the groups affected by the decision. Ideally, organizations should establish explicit decision rules for managing the trade-off between accuracy and disparity, test models across relevant and intersectional groups, document the rationale for model choices, and monitor fairness after deployment. Fairness should consequently be treated as a component of model quality.

Recent regulatory frameworks, most notably the EU AI Act, further accelerate this operational shift by legally mandating that organizations ensure fairness and mitigate bias in their systems. What was previously approached primarily as a voluntary "nice-to-have" initiative, has now transitioned into a matter of AI conformity.





Jean-Michel Loubès
— Research Director, Inria

" AI is a mirror of the flaws in our societies. We had not seen them, or did not want to, and the use of AI forced us to confront them. This brought about a realization, and as a result, we are correcting them."



Veronika Shilova
— Senior Research Scientist, Artefact Research Center

" Fairness has many definitions, and most of them cannot be satisfied at the same time. Before we can measure or mitigate bias, we must decide which notion of fairness we are aiming for. That choice is as important as any correction we apply."

#02 OUR RECOMMENDATION

AAAI CONFERENCE
AAAI Conference on Artificial Intelligence

Ensure that fairness is embedded directly into model training with OT-FairBoost

The following section is based on an open-source package, "OT-FairBoost: a Python package improving fairness in gradient boosting training", jointly published with the following affiliations:

IN COLLABORATION WITH

ARTEFACT RESEARCH CENTER · cnrs · Inria · UNIVERSITÉ TOULOUSE III PAUL SABATIER · SORBONNE UNIVERSITÉ

The challenge

AI models for tabular data are widely used for high-stakes applications such as credit assessment, recruitment, insurance, and risk management. Yet their predictive accuracy does not guarantee that they produce fair outcomes: these models can produce systematically different predictions across demographic groups, thus leading to potentially discriminative treatments. Such disparities may arise from historical data, proxy variables, or the model's focus on accuracy optimization. This presents a growing challenge for organizations facing rising expectations around non-discriminatory AI. Existing fairness strategies typically modify data before training or adjust predictions afterward, which can reduce information, complicate implementation, or weaken the balance between accuracy and fairness. Methods that embed fairness directly into the training of the models are theoretically more reliable but remain limited.

The solution

OT-FairBoost embeds fairness directly into the model training. This reduces disparities between demographic groups while preserving predictive performance, making fairness an explicit modeling goal rather than a post-processing adjustment. Across tested settings, OT-FairBoost achieves stronger accuracy–fairness trade-offs than existing approaches, offering organizations a practical and effective way to integrate responsible AI into established modeling workflows.

How to use it

01

STARTING POINT

An AI use case on tabular data

02

THE TOOL

Train OT-FairBoost model on this data instead of XGBoost, LightGBM or scikit-learn

03

THE BENEFITS

The resulting model predicts without discrimination

GITHUB.COM/ARTEFACTORY

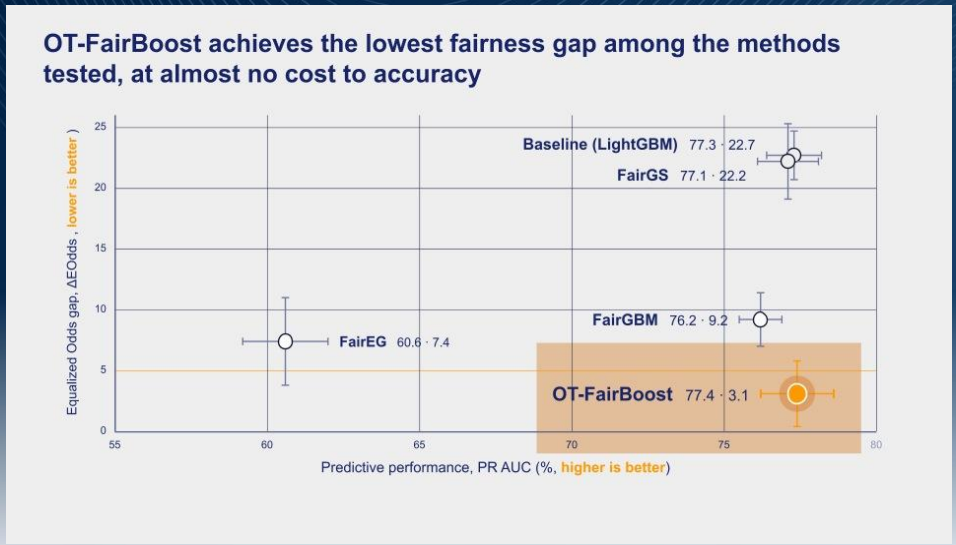


Figure 2: OT-FairBoost reaches the lowest fairness gap among the methods tested, at almost no cost to accuracy

2

AI governance

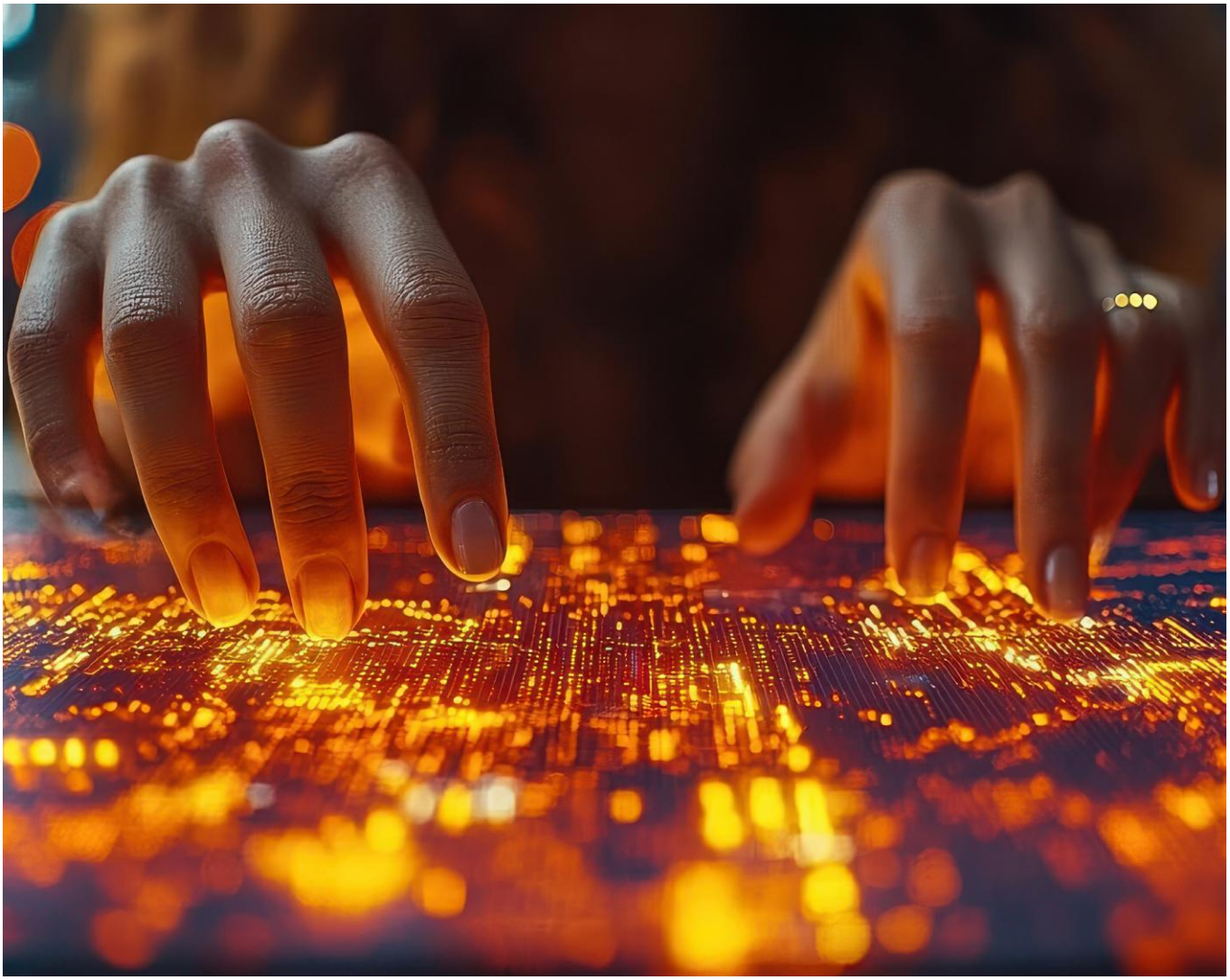
How can organizations demonstrate control over their AI?

Regulators, auditors and boards expect evidence, not assurances. This chapter defines how to evaluate, approve, audit and justify AI systems, a requirement made pressing by the AI Act and by banking supervision.

ESG PILLAR

E S G Governance





IN THIS CHAPTER

2.1 A robust evaluation framework can protect organizations from AI risks at scale 16
17

RECOMMENDATION 03 Discover unknown biases in your AI with FairGroups

2.2 Maintaining control over AI-empowered processes requires model explainability, data lineage and auditability 18
19

RECOMMENDATION 04 Turn model opacity into a short list of auditable decisions with WoodTapper

2.1 — A robust evaluation framework can protect organizations from AI risks at scale

As a first step to decide whether an AI model can be trusted, organizations need to assess both what the model can do and what risks may arise from its deployment. This includes risks that emerge at scale, risks of non-compliance, and cybersecurity risks created by integrating AI into existing information systems.

There are some historical evaluation practices but they were developed for a specific purpose. For instance, benchmark testing measures the accuracy of a model on a defined task against a fixed dataset, and establishes how that model performs under specified conditions. However, it does not establish whether an organization should deploy the model in a given context. Addressing this gap requires rigorous evaluation practices, covering the scope of assessment, its sequence, and the evidence required at each stage. Expert interviews recommend an assessment in three stages:

Assessment of capabilities: The capabilities of a model determine which risks are relevant to a given deployment. An assessment conducted before capabilities have been established is liable to examine risks that do not apply to the system in question and to omit those that do.

Assessment of risks: Three categories are relevant to most deployments:

- Systemic risks: Individual outputs may be defensible while their aggregate effect is not. Such risks are not detectable through case-level testing.
- Compliance risks: Some risks arise when an AI system produces outcomes that conflict with legal or regulatory requirements, even without deliberate misconduct by the organization or its users. Algorithmic collusion is one example: independently operated pricing systems could converge on anti-competitive outcomes without any explicit agreement between firms.
- Cybersecurity risks: The integration of an AI model into an information system creates an additional surface of attack. This category is frequently omitted from model assessment, as responsibility for it generally sits with functions other than those conducting the evaluation.
- Determination of trust: Once risks are identified, two criteria apply:
  - Reliability: Understood as the consistency of a model's behavior, including under conditions not represented in its training data. This is a property of the model.
  - Suitability for use: Understood as the correspondence between the model and the decision, population, and process in which it is deployed. This is a property of the relationship between the model and its context, and is not established by technical performance alone.



Figure 3: A three-stage assessment, from capabilities to a decision to deploy (Source: authors' own elaboration)

Methods for assessing these criteria remain unevenly developed. While the measurement of accuracy is well established, the measurement of suitability for use, of systemic effects and of resistance to attack is not, and no agreed standard specifies the level of evidence sufficient to support a deployment decision. This absence constitutes a principal obstacle to consistent evaluation practice across organizations, and evidence on how organizations conduct such assessments in practice remains limited.

#03 OUR RECOMMENDATION

Discover unknown biases in your AI with FairGroups

The following section is based on an open-source package, "FairGroups: a Python package", jointly published with the following affiliations: Artefact Research Center, Institut de Mathématiques de Toulouse, CNRS, Université de Toulouse, L'Oréal Research and Innovation, Inria, ANITI.

IN COLLABORATION WITH

ARTEFACT RESEARCH CENTER · cnrs · UNIVERSITÉ TOULOUSE III PAUL SABATIER · Université de Toulouse · Inria · L'ORÉAL RESEARCH & INNOVATION

The challenge

For companies using online recommender systems, personalized tools such as virtual try-on and foundation-shade matching create value only when they perform reliably across populations. Otherwise, customers receive unsuitable recommendations, lose trust in the brand, and expose the company to reputational and regulatory risks. The EU AI Act, for example, emphasizes reporting performance across demographic groups. Yet unequal performance, and the biases it may reflect, can be difficult to detect. Fairness assessments typically compare predefined demographic groups, leaving open the possibility that a model performs poorly for an overlooked population.

Detecting such biased recommendations a posteriori, without specifying the affected group in advance, can help companies identify problems earlier and mitigate associated risks.

The solution

FairGroups, developed in a research partnership between the Artefact Research Center and L'Oréal Research & Innovation, reverses the usual order. It takes as an example the grouping of users according to their skin tone to discover the discriminated groups. The most disadvantaged ranges of users can thus receive a special treatment to mitigate this discovered bias.

How to use it

01

STARTING POINT

An AI use case with risk of bias on an unknown group of users

02

THE TOOL

Run FairGroups on the predictions of your AI on your data to discover discriminated groups

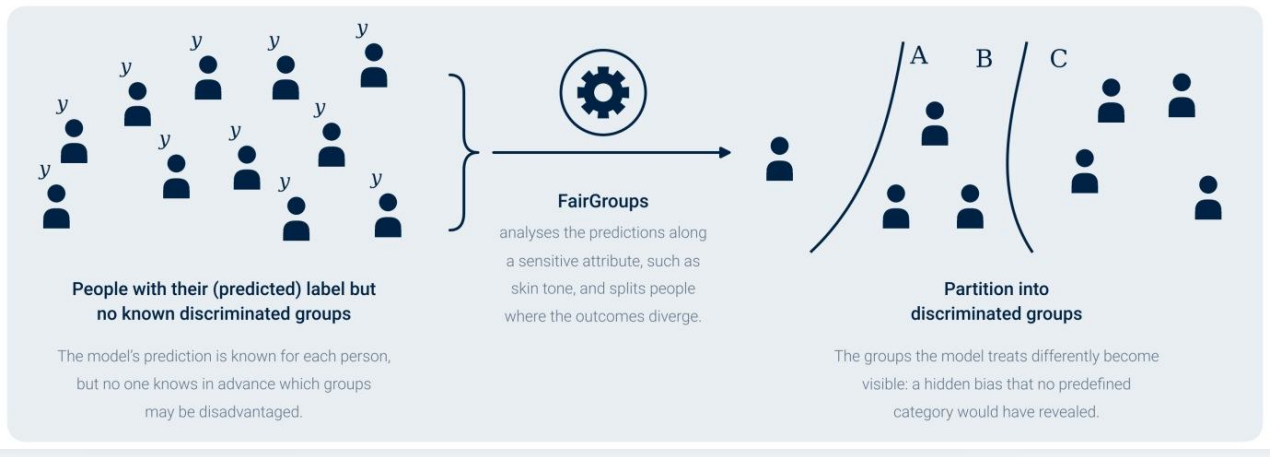
03

THE BENEFITS

Analyze the level of bias, and take actions if needed (on the data, on the model or a posteriori)

GITHUB.COM/ARTEFACTORY

Figure 4: FairGroups partitions users into groups according to how the model performs for them, without knowing the discriminated groups in advance



2.2 — Maintaining control over AI-empowered processes requires model explainability, data lineage and auditability

The principles below apply to AI governance across the banking sector. They do not describe the specific systems, models, or control environment of any individual institution.

An AI model's business value depends on how well its outputs can be monitored and controlled. AI is increasingly used to support decisions across sectors: risk assessment and fraud detection in finance, diagnosis and patient management in healthcare, and forecasting, quality control and operational optimization in manufacturing and logistics. The challenge is to create value while keeping control over critical decisions. Errors can lead to financial loss, inappropriate access to services, regulatory exposure, operational disruption, or reduced customer confidence.

Four areas deserve particular attention:

Data quality.

Organizations must assess whether the data is complete, accurate, up to date, traceable and suitable for the intended use. Even technically precise model outputs may be unreliable if the underlying data are not fit for purpose. The output itself may not reveal the problem. This makes data provenance and lineage important for model oversight.

Information inferred by the model.

Organizations must examine both the variables used by a model and the information it can infer from them. Models may rely on indirect variables, reconstruct information removed from the dataset, or identify correlations that are unsuitable for a particular decision. Removing a sensitive variable does not necessarily remove the information associated with it if other variables act as proxies.

Suitability of outputs.

Technical performance shows how accurately a model identifies a relationship or predicts an outcome. It does not, on its own, establish whether that relationship is appropriate for the intended use. That assessment requires human judgment, regulatory knowledge, and an understanding of the people affected by the decision.

Auditability.

Organizations must be able to trace how data entered the system and how they were transformed. They must also identify which relationships contributed to the output, how the model was validated, and what decisions followed. This record must remain available throughout the model's lifecycle.

A model that is hard to explain can also be harder to adopt, validate, and oversee. Business users may hesitate to rely on it; control functions may need to carry out further reviews, and regulatory expectations may call for further evidence of governance, performance and explainability. If a result is unexpected, tracing its source and deciding how to respond can also be more difficult.

Banking is a useful example because decisions need to be accurate, traceable, and defensible. Potential AI use cases include credit, fraud detection, risk management, and regulatory reporting. Banks also handle large volumes of data.

Using these data for AI requires appropriate governance and controls. In banking, explainability helps maintain operational control. Depending on the use case and applicable requirements, sufficient explainability may be needed to support business understanding, independent review, auditability and regulatory oversight. Without it, validation, monitoring, and governance can become more difficult.



SOCIETE GENERALE

Valérie Vercoutre

— Global Head Risk & Finance Data Management, Societe Generale

"AI can support decisions, but people remain responsible for them. Responsible AI requires human judgment, strong governance and effective controls at every stage of its lifecycle."



ARTEFACT RESEARCH CENTER

Abdoulaye Sakho

— Senior Research Scientist, Artefact Research Center

"AI models no longer have to be a choice between performance and transparency. Most AI models now have dedicated explainability tools that can help establish trust with practitioners and business leaders."

#04 OUR RECOMMENDATION

Turn model opacity into a short list of auditable decisions with WoodTapper

The following section is based on an open-source package and paper named "WoodTapper: A Python package for explaining decision tree ensembles", published in the Journal of Open Source Software, and co-authored with the following affiliations:

IN COLLABORATION WITH

ARTEFACT RESEARCH CENTER

SORBONNE UNIVERSITÉ

SOCIETE GENERALE

The challenge

As organizations deploy increasingly complex AI models in regulated and high stakes settings, they need practical methods to examine how those models generate predictions. This requirement applies to different types of AI systems, and methods required for evaluation depend on the type of model being assessed. This challenge is especially pronounced with tabular data, when information is structured and arranged in standardized rows. For example, industrial scenarios such as credit risk scoring, insurance underwriting, and clinical outcome forecasting are increasingly solved by AI on tabular data, for which practitioners and stakeholders require transparent, mathematically sound methods to inspect model behavior.

The solution

WoodTapper, an open-source Python package developed by Artefact Research Center in collaboration with Societe Generale addresses the explainability of these AI models. It is designed to make them easy to be analyzed and reviewed by a human. The tool translates the AI into simple rules in the form of "if … then …", based on variables and thresholds that contribute to the model's predictions. The easy-to-read list of rules supports the examination of the model's behavior, with no hidden patterns. First, data scientists can use it to examine model behavior in the model design phase. Second, business users can use the information to interpret individual outputs. Model-risk and control functions can use it to assess assumptions and investigate potential risks. Auditors and external regulators can use the resulting analysis to examine in detail how the model operates.

How to use it

01

STARTING POINT

An AI use case on tabular data

02

THE TOOL

Train a WoodTapper model on your data instead of a scikit-learn black box model such as random forest

03

THE BENEFITS

Read and understand the complete WoodTapper model's decisions mechanism as a list of simple if-else rules

GITHUB.COM/ARTEFACTORY/WOODTAPPER

3

AI and the Environment

How to deploy AI under environmental constraints?

AI consumes compute and energy. This chapter provides the criteria to prioritize use cases by the value they create and to reduce the footprint of the underlying infrastructure, keeping AI aligned with climate commitments.

ESG PILLAR

E S G Environmental





IN THIS CHAPTER

3.1 Strategic AI deployment under environmental constraints 22

3.2 Matching model size to the value of each use case 24

3.3 Decarbonizing the infrastructure underpinning AI 26

3.1 — Strategic AI deployment under environmental constraints

Companies increasingly face decisions about whether to automate activities, expand existing applications, adopt AI functions embedded in enterprise software, or invest in more computationally intensive systems. These decisions are traditionally evaluated through productivity gains, cost savings or technical performance. Environmental constraints introduce another consideration: the marginal value created by additional computation relative to the resources it requires.

A strategic deployment strategy would therefore distinguish between applications whose value justifies their environmental demands and those for which additional computational capacity produces limited benefits. This may lead companies to replace general-purpose models with specialized systems, evaluate vendor offerings, or question deploying agentic AI, large language models, or machine learning, where simpler technologies achieve equivalent results.

However, the environmental implications of AI cannot be addressed through model efficiency and computational cost optimization alone, including approaches such as FinOps. While these practices can reduce the resources required for individual workloads, they do not determine whether an AI application is necessary, whether its economic and societal value justifies its environmental footprint, or whether efficiency gains translate into lower absolute resource consumption. For companies, the more consequential challenge is therefore one of strategic prioritization of use cases: deciding which ones justify computational resources, at what scale, and under what environmental constraints.



carbone4

“For companies adopting AI, the priority is to question and prioritize its uses: assess the relevance of each AI integration, push back against the default inclusion of AI features in software, concentrate efforts on a limited number of high-value applications, and favor smaller models whenever the task allows.”

Zénon Vasselin

— Head of IT Practice, Carbone4



leaders for climate action

"

The relevant question is which model is good enough for the use case, considering its capability, cost and resource requirements."

Piotr Drozd

— Chief Impact Officer, Leaders for Climate Action

Rebound and induction effect

As the cost of computation declines, companies can expand AI applications, increase their frequency of use, and automate activities that previously required greater justification. Lower resource consumption per task may therefore coexist with rising aggregate energy demand and emissions. This rebound effect is particularly relevant to corporate AI strategies built around rapid deployment. A company may reduce the computational footprint of individual applications while increasing the number, complexity and frequency of AI-enabled tasks.

Evaluating efficiency at the application level would then provide an incomplete account of the environmental consequences of the overall portfolio. Companies therefore need to distinguish efficiency gains from absolute environmental effects.

The more difficult question concerns the economic activities AI makes possible or more profitable. AI may improve the efficiency of resource-intensive production, accelerate content generation or fossil fuel exploration, or strengthen the commercial viability of carbon-intensive activities. These effects cannot be assessed solely through the electricity consumed by the AI system. They require consideration of how AI changes production volumes, demand and investment incentives, and how AI catalyzes other sectors.

This implies monitoring total computational demand, establishing deployment thresholds, and evaluating AI investments against corporate decarbonization objectives. It also places responsibility beyond IT departments: business units, procurement teams, and executive committees must be able to assess whether expanding an AI application remains justified as its scale and effects evolve. Finally, it requires high-order effect analysis and regulation to ensure AI is used to accelerate the low-carbon transition rather than undermining even faster climate security.

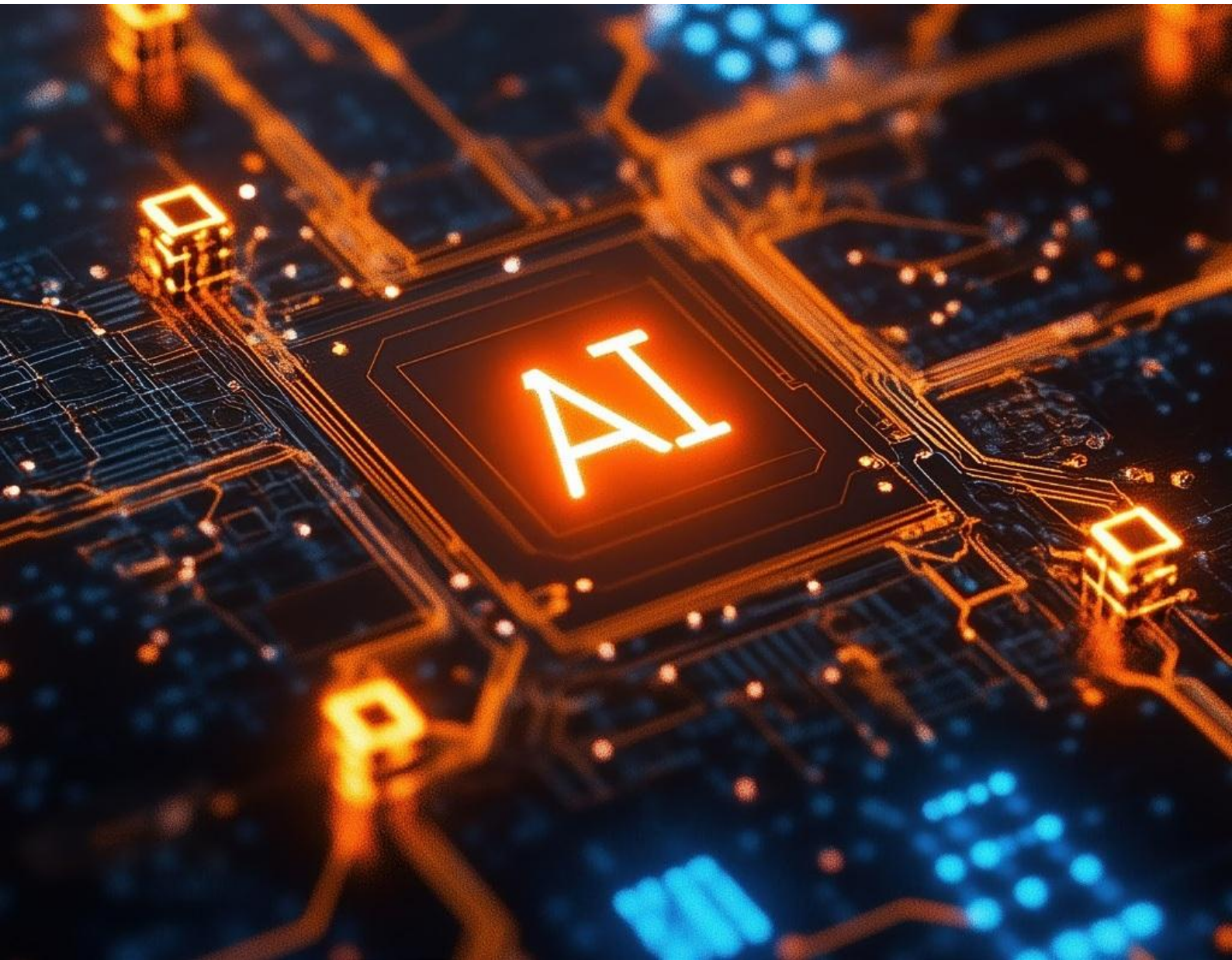


3.2 — Matching model size to the value of each use case

Prioritizing use cases determines where AI should be deployed. The choice of model then determines how much each deployment consumes. For a given task, organizations can select among models of very different sizes, and the resources required grow sharply with the number of parameters.

Model selection therefore carries a direct environmental consequence. This impact can be quantified over the full life cycle of a request, including the manufacturing of the hardware that runs it. As such, the footprint of a model becomes comparable with its cost and performance. Three factors drive it: the energy per generated token, the share of hardware manufacturing attributed to each request, and the electricity mix of the data center.

Between a large model and a small one serving the same task, the resulting footprint differs by several orders of magnitude. For a single application, these volumes remain limited. Across a portfolio, they compound: each new use case adds requests, and AI agents multiply the requests behind each task. Right-sizing (selecting the smallest model that meets the requirements of each use case) is therefore a structural and measurable lever of sobriety.



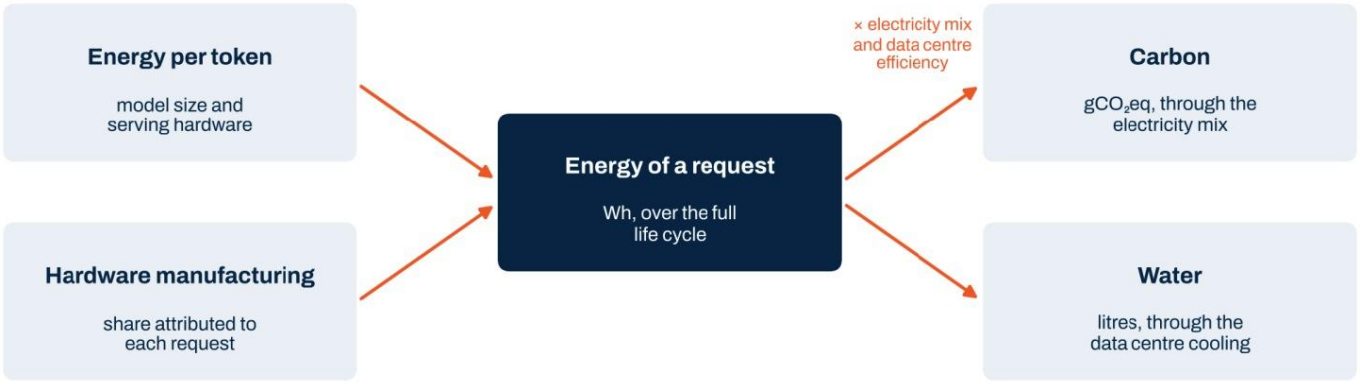
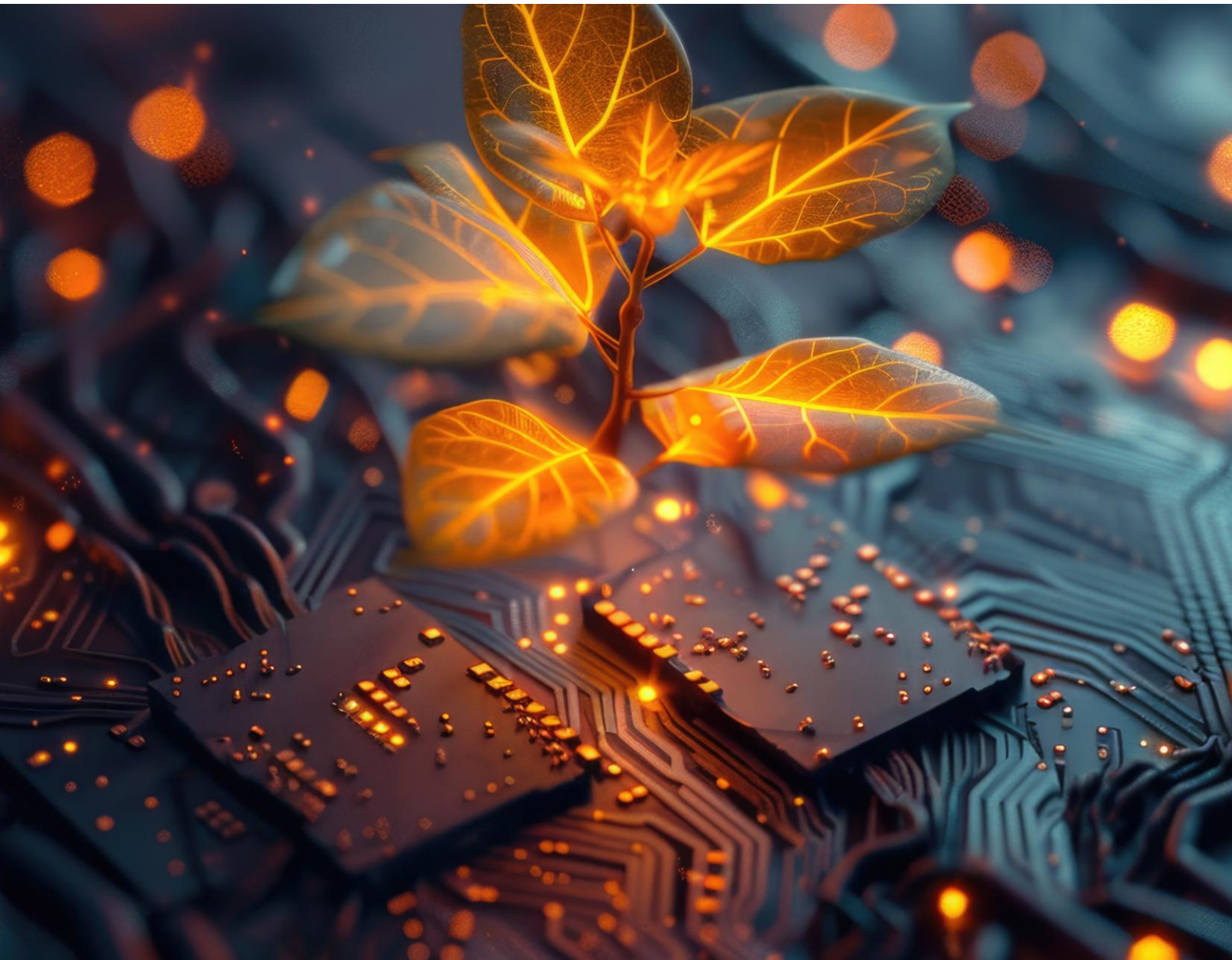


Figure 5: How the footprint of a request is built, from the energy per token to carbon and water



3.3 — Decarbonizing the infrastructure underpinning AI

The environmental footprint of AI extends beyond the electricity consumed during computation. As demand for computing capacity grows, the construction of data centers creates additional pressures through material production, land use, cooling requirements, and infrastructure development. These impacts raise strategic questions for operators, whose environmental performance increasingly depends on decisions made before facilities become operational.

The experience of Data4 illustrates how life cycle assessment (LCA) can inform these decisions, from construction materials and site design to energy sourcing and cooling technologies.



Insights from Data4 Group: building the next frontier of low-carbon infrastructure for AI

Linda Lescuyer

— Head of ESG & Sustainable Innovation, Data4 Group; Board member, Net Zero Innovation Hub for Data Centres

The growth of AI translates directly into demand for new computing capacity, and that capacity has a physical footprint before it processes a single request. For a data center operator, the environmental question therefore extends beyond energy use to the way infrastructure is designed and built. Data4 addresses it across the full life cycle of a site, from construction materials to operation and the reuse of residual heat.

Measurement as the starting point. Data4's approach rests on life cycle assessment, which allows impacts to be located before they are addressed. Its assessment of a representative site shows that operations are the largest source of carbon impact, while producing construction materials, mainly concrete and steel, comes close behind. Building and running are thus two distinct problems, each requiring its own set of solutions.



Low-carbon construction. Construction is the lever most specific to the expansion of capacity. Data4 has made eco-design a standard for new sites, relying on low-carbon concrete and prefabricated slabs, which reduce concrete volume and transport. Where construction emissions cannot yet be avoided, the company supports carbon sinks, including a local manufacturer of bio-based insulation made from recycled paper.

Efficient and low-carbon operation. On the operational side, the company sources low-carbon electricity through long-term power purchase agreements, mostly renewable, and relies on free cooling and dry cooling to limit both energy and water use, a growing concern as computing density increases.



4 — Responsible AI at Artefact
4.1 — Our commitments



ARTEFACT

> *"Responsible AI is no longer solely a CSR matter; it is a company-wide strategic responsibility. It must be embedded across all business functions, from technology and data development to corporate governance and decision-making. Only by making responsibility everyone's business can we ensure that AI is developed and deployed in a way that is both sustainable and beneficial to society."*
>
> Lucie Filiol de Raimond
> — Group CSR Manager

01 Trusted and ethical models

— An ethics charter of 10 principles, including transparency, explainability, bias avoidance and accountability. Everyone at Artefact is invited to read and sign it.

— Checklists and control points for technical teams, covering ethical data collection, data handling and model building.

— Labelia certification (intermediate level, 2022 and 2024 at the level of HQ), an independent label for responsible and trustworthy AI.

— Positive AI. Exchange with peers on building ethical AI

02 Governance and accountability

— ISO 27001 certification for Artefact France and Artefact UK (2025), which embeds security across business processes.

— Mandatory training on data protection (GDPR) and anti-corruption, supported by a whistleblowing framework.

— External assessment: EcoVadis Bronze medal in 2025, in the top 35% of companies rated, and membership of the UN Global Compact since 2025.

03 Environmental footprint

— A target of −50% greenhouse gas emissions by 2030, with a five-year roadmap and a commitment to the Science Based Targets initiative (2025). Scope 3 represents 98% of the footprint.

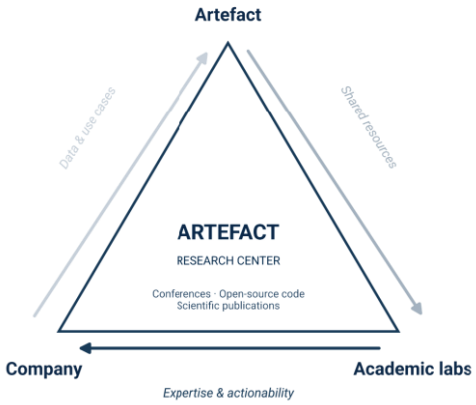
— Eco-design guidelines for AI solutions.

— A carbon dashboard for projects, in development, to track emissions down to technical components, with an internal handbook for sustainable data.

— Coalition for Sustainable AI: Artefact contributes to standards for AI aligned with environmental goals.

4.2 — The Artefact Research Center: trustworthy AI, built from the field

The Artefact Research Center bridges academic research and real business use cases. Everything it produces is open source: conferences, scientific publications and code that companies can use.



01 Research that starts in the field

Research questions come from the real challenges companies face when they deploy AI.

02 Data that comes from the field

Real data and use cases are shared with researchers, to test methods at scale.

03 An ecosystem rooted in academia

Professors and PhD candidates from leading universities and schools work with our scientists.

OUR FOCUS

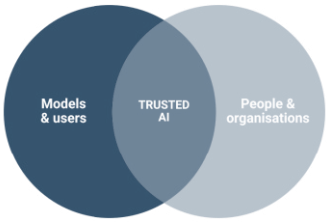
Trustworthy AI, seen from both sides

The Research Center works on the two sides of trust: the models and the people who use them, and the organizations that deploy them and answer for their decisions.

Control & fairness

Explainability & transparency

Uncertainty & bias



AI adoption in business

Auditability

Ethics & accountability

Open-source tools

Artefactual

Hallucination detection for large language models.

SEE RECOMMENDATION 01

OT-FairBoost

Fairness built into gradient-boosting training.

SEE RECOMMENDATION 02

FairGroups

Bias detection for continuous sensitive variables.

SEE RECOMMENDATION 03

WoodTapper

Rules extracted from tree ensembles, for audit.

SEE RECOMMENDATION 04

ChoiceLearn

Easy-to-use and scalable discrete choice modeling.

AVAILABLE ON GITHUB

MGS-GRF

Better AI scoring on imbalanced data.

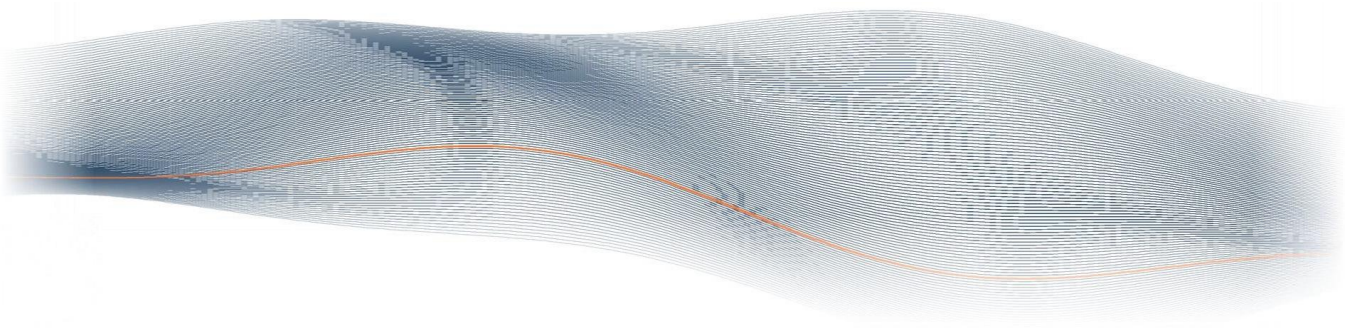
AVAILABLE ON GITHUB

Special thanks and acknowledgments

We would like to thank the following companies and organizations for their direct and indirect contributions throughout the project:

CONTRIBUTORS

- Abdoulaye Sakho, Senior Research Scientist, Artefact Research Center
- Antoine Gallier, Senior Data Scientist, Artefact
- Arthur Garnier, IT Chief of Staff & Data Scientist, Ardian
- Jean-Michel Loubès, Research Director, Inria
- Linda Lescuyer, Head of ESG & Sustainable Innovation, Data4 Group; Board member, Net Zero Innovation Hub for Data Centres
- Lorena Blasco Arcas, Professor and Head of TRACIS, ESCP Business School
- Médéric Chomel, Managing Partner, Group Responsible AI Lead, Artefact
- Piotr Drozd, Chief Impact Officer, Leaders for Climate Action
- Tom Giraudet, Director, Artefact
- Valérie Vercoutre, Global Head Risk & Finance Data Management, Societe Generale
- Veronika Shilova, Senior Research Scientist, Artefact Research Center
- Zénon Vasselin, Head of IT Practice, Carbone4



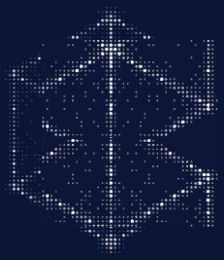
ARTEFACT

We make AI Work.



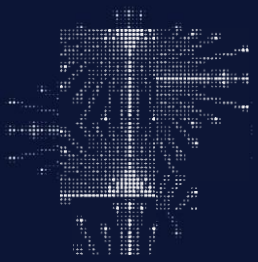
Shape

Define where AI creates value and what to do first (AI strategy, maturity assessment, operating model).



Build

Architect the secure Data & Context platform every team builds on, accelerated by SKAFF, our open-source library of Data & AI accelerators.



Deploy

Industrialize autonomous agents at scale across functions: from marketing, sales, and customer service to supply chain, manufacturing, procurement, R&D, and finance.



Transform

Reshape the organization for AI, turning your teams into AI-native builders through the Artefact AI Academy.

FMCG • RETAIL & ECOMMERCE
LUXURY & COSMETICS • HEALTHCARE
BANKING & INSURANCE • TELECOMMUNICATIONS

SPORTS & ENTERTAINMENT
TRAVEL & TOURISM • PUBLIC & GOVERNMENT
REAL ESTATE • MANUFACTURING & UTILITIES

ARTEFACT

CONTACT

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

ARTEFACT HEADQUARTERS

19, rue Richer 75009 — Paris France

artefact.com

