Trustworthy & Responsible AI

Turning Principles into Practice across Security, Explainability, and Fairness

bitkom

Table of Contents

1 Foreword & Executive Summary

Foreword

Artificial Intelligence has moved rapidly from experimentation into operational deployment across industries and critical domains. With this shift, the focus is no longer only on technological capability, but on **trust, responsibility, and governance**.

The workshop *»Industry meets Research: Trustworthy & Responsible AI«* on February 26, 2026, brought together experts from industry, academia, and public institutions to address a key question: *How can abstract principles such as trustworthiness, in light of security, explainability, and fairness, be translated into practical implementation?*

This publication reflects a shared conclusion: AI systems, especially generative and agentic systems, are dynamic, adaptive, and embedded in complex environments. Ensuring trustworthiness is therefore not a one-time exercise, but a continuous lifecycle responsibility.

This paper is the result of a joint effort by experts from industry and academia, as well as current students and PhD candidates, bringing together a diverse range of perspectives on the state of the art in trustworthy AI. It aims to reflect cutting-edge developments, current challenges, and the latest thinking in this rapidly evolving field. The publication has been developed as a collaborative initiative between Bitkom and AI Grid, with the goal of providing timely and relevant insights for practitioners and the research community alike. A full list of contributing authors can be found on the final page.

AI Grid is Germany's largest network for young AI researchers, funded by the Federal Ministry of Research, Technology and Space (BMFTR). It brings together AI Master's students and PhD candidates within an interdisciplinary community and supports them in their research and career development through scientific exchange in micro focus groups, mentoring with AI experts from academia and industry, and exclusive events in Germany and across Europe.

Executive Summary

Artificial Intelligence has transitioned from experimental use cases to a core component of operational processes, decision-making, and customer-facing services. This shift fundamentally changes how organizations must approach trust, risk, and accountability.

This paper demonstrates that Trustworthy and Responsible AI cannot be achieved through isolated technical controls or one-time validation efforts. Instead, AI systems,

especially generative and agentic systems, introduce dynamic, evolving risk profiles across their entire lifecycle, driven by interactions between data, models, infrastructure, and human behaviour.

For businesses, this has three key implications:

1. **Trust is no longer a property of a system; it is a managed outcome**
   Trust emerges from continuous oversight, governance structures, and organisational capabilities rather than from the intrinsic qualities of a model or vendor.

2. **Risk is systemic and lifecycle-driven**
   Security, safety, fairness, and explainability risks are interconnected and evolve during data collection, model training, deployment, and operation. Addressing them requires coordinated, cross-functional control mechanisms rather than isolated safeguards.

3. **Governance becomes a strategic capability**
   Organisations must move from principle-based discussions towards **operational AI governance**, embedding policies, roles, monitoring, and escalation mechanisms into day-to-day processes.

This paper provides practical guidance across four key dimensions, security, safety, fairness, and explainability, framed within an overarching AI governance system. It translates regulatory requirements (e. g., EU AI Act) into actionable implementation approaches, enabling organisations to:

- Establish lifecycle-based accountability
- Integrate technical and organisational controls
- Enable continuous monitoring and adaptation
- Align AI deployment with regulatory and societal expectations

The objective is not only compliance, but sustainable and scalable value creation through AI systems that remain trustworthy under real-world conditions.

2 Introduction

Purpose and Scope

This whitepaper provides practical guidance for the implementation of Trustworthy and Responsible Artificial Intelligence in business environments. It is directed at decision-makers, developers, operators of AI systems, including agentic AI, and professionals in governance, risk, and compliance functions who are responsible for ensuring that AI systems are deployed in a secure, safe, fair, and explainable manner.

Artificial Intelligence is increasingly embedded in operational processes, decision-making, and customer-facing services. At the same time, AI systems, particularly generative and agentic systems, introduce new and evolving risk profiles. These risks arise not only from technical limitations, but from the interaction of data, models, infrastructure, and human behaviour within complex organisational environments. As a result, ensuring Trustworthy AI is no longer a one-time validation task, but a continuous lifecycle responsibility.

This publication addresses this challenge by translating high-level principles and regulatory requirements, most notably the EU AI Act, as well as the General Data Protection Regulation and the Data Act, into actionable implementation approaches. It adopts a risk-based and lifecycle-oriented perspective, covering the phases of data, model development, deployment, and operation.

To provide a structured and operational entry point, the paper focuses on three key dimensions of Trustworthy and Responsible AI:

- **Security:** protection against malicious interference and unauthorised access
    - Safety: prevention and control of unintended harm during system operation
- **Fairness:** avoidance of unjustified bias and discriminatory outcomes
- **Explainability:** enabling understanding, oversight, and accountability

These dimensions are treated as interdependent control domains that must be addressed jointly within an overarching AI governance framework.

The scope of this publication is intentionally focused: it does not aim to provide a comprehensive treatment of all aspects of Trustworthy AI. Topics such as data privacy, robustness, or transparency are addressed where relevant but not exhaustively. Instead, the objective is to offer hands-on technical and organisational guidance that supports organisations in operationalising Trustworthy and Responsible AI in real-world settings.

Ultimately, this paper aims to support organisations in moving from abstract principles and regulatory requirements toward concrete, scalable implementation practices, enabling both compliance and sustainable value creation through AI.

Defining Trustworthy & Responsible AI

To ensure legal clarity and aligned interpretation, this publication adopts concise definitions of essential AI terms.

An extensive glossary can be found on page 47.

Artificial Intelligence (AI)

EU AI Act (Article 3(1))
Definition: »AI system means a machine-based system (...) designed to operate with varying levels of autonomy (...) inferring from input how to generate outputs that influence physical or virtual environments.«[1]

German Implementation
Germany adopts the EU definition through its »Gesetz zur Durchführung der KI-Verordnung«, which references the EU definition without alteration for national market surveillance and governance.[2]

Trustworthy AI

EU AI Act (Regulatory Objective)
Defined through purpose statements emphasizing human-centric, safe, transparent, robust, and fundamental-rights-respecting AI.[3]

German Implementation
Germany reinforces the EU's trustworthiness criteria by mandating transparency, rights protection, non-manipulation, and strict controls for high-risk AI.[4]

[1] Artificial Intelligence Act (n.d.) Article 3. Available at: https://artificialintelligenceact.eu/article/3/
[2] Bundesministerium für Digitales und Staatsmodernisierung (n.d.) Gesetz zur Durchführung der KI-Verordnung. Available at: https://bmds.bund.de/service/gesetzgebungsverfahren/gesetz-zur-durchfuehrung-der-ki-verordnung
[3] European Union (2024) Regulation (EU) 2024/1689 (Artificial Intelligence Act). Available at: https://eur-lex.europa.eu/eli/reg/2024/1689/oj/eng
[4] Bundesregierung (n.d.) AI Act: EU einigt sich auf Regeln für Künstliche Intelligenz. Available at: https://www.bundesregierung.de/breg-de/aktuelles/ai-act-2285944

Responsible AI

(Not a legal term in the Act; derived from binding obligations.)

EU AI Act (Operational Requirements)
Responsibility is reflected in mandatory controls: risk management, data governance, documentation, transparency, human oversight, accuracy, robustness, and cybersecurity (Articles 8–15).[5]

German Implementation
National governance structures operationalise responsibility through oversight, accountability, documentation duties, and coordinated enforcement via bodies such as KoKIVO.[6]

[5] AI Act Info (n.d.) AI Act Overview. Available at: https://aiactinfo.eu/
[6] AI Bundesministerium für Digitales und Staatsmodernisierung (n.d.) Gesetz zur Durchführung der KI-Verordnung. Available at: https://bmds.bund.de/service/gesetzgebungsverfahren/gesetz-zur-durchfuehrung-der-ki-verordnung

AI Governance

More recently, the discussion has shifted from »trustworthy« or »responsible« AI towards AI Governance. As AI applications are embedded in day-to-day processes, the focus expands from a tool being »trustworthy« once and for all to a broader context with implications across the full lifecycle for society, business, employees, and customers.

The goal of AI Governance is to manage, avoid and mitigate risks associated with AI applications, including the definition and application of principles, policies, processes, and practices to ensure responsible development, implementation, and operation. It becomes key mechanism for developing and maintaining trustworthy AI across the lifecycle.

For AI product owners and risk professionals, this shift from a static product to ongoing governance requires major changes: product owners need a deeper understanding of law, ethics, and risk management, while risk professionals need a stronger understanding of AI, technical infrastructure, and data. Effective governance requires cross-functional collaboration across the lifecycle. This is a growing challenge, especially for smaller companies and with the increasing use of third-party and shadow AI.

Additionally, the AI risk taxonomy is dynamic: with new models and agentic technologies, risks and attack vectors are constantly evolving, making AI governance a moving target.

3 Security

Safety and Security of GPTs, LLMs, and Agentic AI Systems

Large Language Models (LLMs), including GPT-based systems, and increasingly agentic AI architectures fundamentally change how AI systems are designed, deployed, and operated. Unlike traditional software, these systems are probabilistic, adaptive, and deeply embedded in complex environments. Their behaviour is shaped not only by their internal model structure, but by interactions with data, users, tools, and other systems. As a result, they introduce a dynamic and evolving risk landscape.

A key distinction must be made between **security** and **safety**. Security focuses on protecting AI systems against intentional manipulation, such as prompt injection, model extraction, or unauthorised access. Safety, by contrast, addresses the prevention of unintended harm caused by system behaviour, including hallucinations, incorrect recommendations, or flawed autonomous actions. In practice, both dimensions are closely interrelated and must be addressed jointly.

Modern AI systems should not be understood as isolated components, but as networked systems composed of models, agents, data sources, interfaces, and external tools. In such systems, risks do not remain local. Instead, they can propagate through interactions. A vulnerability in one component, whether technical, data-related, or behavioural, can influence other components and, under certain conditions, escalate into system-wide failures.

From this perspective, risk is not only determined by the presence of individual weaknesses, but by the structure and connectivity of the system. Highly connected systems with strong interaction dependencies increase the likelihood that disturbances, whether caused by attacks or unintended behaviour, can spread and amplify. This is particularly relevant for agentic AI systems, where models are not only generating outputs but are also able to take actions, invoke tools, and interact with other systems. In such settings, a single faulty or manipulated step can trigger cascading effects across multiple processes.

Security risks in LLM-based systems include manipulation through crafted inputs, leakage of sensitive information, or attacks on the model supply chain. These risks are amplified when models are integrated into broader systems via APIs, retrieval mechanisms, or external tools. At the same time, safety risks arise from the inherent uncertainty of model behaviour. Even without malicious intent, models may produce incorrect, biased, or harmful outputs, which can become critical when directly linked to operational decisions or automated actions.

Agentic AI systems further increase both safety and security risks by introducing autonomy and execution capability. Decisions are no longer merely advisory but can directly trigger actions. This reduces the time available for human intervention and

increases the potential impact of errors or manipulations. In interconnected environments, such effects may propagate rapidly across system boundaries.

For these reasons, safety and security cannot be treated as static properties verified at deployment. They must be managed as continuous control problems. Effective mitigation requires a combination of structural, technical, and organisational measures. These include limiting system connectivity where possible, enforcing least privilege access to tools and data, continuously monitoring system behaviour, and maintaining meaningful human oversight for high-impact decisions.[7]

Finally, the safety and security of GPTs, LLMs, and agentic AI systems emerge from the interaction of their components rather than from any single element. Ensuring trustworthy operation therefore requires a system-level perspective, where risks are understood, monitored, and controlled across the entire lifecycle of the AI system.

Relevance and Risks

Building on the system-level perspective outlined above, security in AI systems focuses on protecting models and their surrounding ecosystems against intentional manipulation and misuse. Unlike traditional IT security, this requires addressing risks that emerge from the interaction between data, models, and human-AI interfaces.

LLM-based systems introduce new attack vectors, including prompt injections, model extraction, and adversarial inputs, which call for adapted and context-aware safeguards. At the same time, relying on isolated technical measures (e. g., prompt scanners) can create a false sense of security if they are not embedded within a broader, system-oriented security and governance framework.

AI systems introduce new and evolving attack surfaces across the lifecycle

The rapid development of AI systems and their deep integration into everyday life demonstrate that they have become indispensable. In addition to the many opportunities these systems offer across a wide range of applications, they also present vulnerabilities throughout the entire lifecycle of an AI model, from the data foundation through model training and operation to model adaptation[8]. During the data phase alone, manipulated training data can deliberately distort a model's behaviour or create hidden backdoors that can harm the model's performance later. Additional risks also arise during training and the integration of pre-trained models due to insecure supply chains. Manipulated models or components often go undetected and only reveal harmful effects once the system is in operation. During the operational phase, adversarial attacks occur, in which minimally altered inputs lead to incorrect decisions. At the same time, attacks such as model extraction or membership inference allow

[7] Bitkom e. V. (2025) Security of AI Agents. Available at: https://www.bitkom.org/Bitkom/Publikationen/Security-of-AI-Agents
[8] National Institute of Standards and Technology (NIST) (2025) Adversarial Machine Learning: A Taxonomy and Terminology of Attacks and Mitigations. Available at: https://csrc.nist.gov

sensitive training data to be inferred[9]. With the proliferation of generative AI, the attack surface is expanding further: techniques such as prompt injection or API abuse demonstrate that not only the models themselves, but entire AI ecosystems are affected.

Especially with the emergence of agentic systems, the attack surface increases at an even faster pace. Agentic AI systems often connect LLMs with user interfaces that were originally designed for human interaction. Although these systems are typically equipped with basic security measures such as input sanitisation and access management, their security is not designed for agentic use by LLMs. Many of their features assume that a reasonable human is interacting with them. This assumption cannot be made when securing agentic systems and is the main reason why prompt injections are so widespread in such systems.

Overall, AI systems create dynamic, constantly evolving attack surfaces across all phases. Accordingly, security strategies must address the entire lifecycle and be continuously adapted.

Prompt injection and jailbreaks can manipulate system behaviour and bypass safeguards

The enormous success of LLMs and their widespread adoption in everyday applications are not only driving productivity gains in many areas (e. g., customer service, software development, etc.) but also creating new entry points for external attacks. A growing threat stems from so-called evasion attacks, in which attackers attempt to manipulate the LLM during operation and induce undesirable model behaviour[10]. An example of such an attack is (indirect) prompt injections, in which the model's behaviour can be significantly influenced by unnoticed external inputs. This attack affects LLMs that access data from external sources or use unverified third-party documents as input. The attacker hides instructions for the LLM on websites or within documents to bypass security and protection mechanisms. This can, for example, result in an autonomous agent equipped with an LLM for instruction processing granting an attacker access to data or systems. Similarly, jailbreak attacks on LLMs can also pose a threat. These attacks attempt to circumvent predefined security barriers and model restrictions by using specific phrasing within a prompt to cause information that is normally protected to leak from the LLM (e. g., malicious instructions, prohibited content). These and other adversarial attacks must be considered during the training, operation, and use of LLMs, and their effects must be mitigated through practical countermeasures (e. g., robust guardrails) as well as comprehensive governance, security monitoring, and proper management of non-human identities (NHI).

Many security features like guardrails are advertised as plug-and-play security solutions. However, in most cases AI, and especially agentic systems, require a comprehensive security structure surrounding them. Ideally most security features related to prompt security and governance are included in the central AI platform of an

[9] European Union Agency for Cybersecurity (ENISA) (2023) Artificial Intelligence and Cybersecurity Research. Available at: https://www.enisa.europa.eu
[10] Bundesamt für Sicherheit in der Informationstechnik (BSI) (2025) Evasion-Attacks on LLMs – Countermeasures in Practice.

organisation. This allows teams across the organization to innovate at scale while maintaining centralized security oversight and response capabilities. Core systems across organizations and central identity providers (IdPs) must be equipped to handle NHIs at scale, as they require a very short lifecycle, a narrow access scope and automatic provisioning and decommissioning. These measures are required as, to date, prompt injections are not a solved problem and will likely remain a risk in LLM-based systems. Therefore, it is important to minimise their impact and blast radius through other security measures.

Data leakage and data poisoning risks increase with external data sources and APIs

Data poisoning and data leakage are among the key security risks in machine learning and pose a threat to both the integrity and confidentiality of data and models.

Data poisoning refers to attacks that cause machine learning models to malfunction or perform poorly by deliberately manipulating training data. For example, inputs are labelled incorrectly and introduced into the training dataset. Even a small amount of such manipulated data can affect model performance.[11]

An important variant is the backdoor attack, in which a model is trained to respond to a specific, usually inconspicuous pattern in the input data with a false output determined by the attacker, with the manipulation often remaining undetected.

The risks of data poisoning and data leakage increase, particularly when using external data sources and APIs. If training data is automatically sourced from third-party sources, attackers can inject manipulated data without having direct access to the system. At the same time, data leakage can occur if sensitive information is unintentionally included in training data or disclosed via interfaces. External APIs can also serve as an indirect attack vector, for example, when they provide data or labels that could harm the model.

Model-related risks

Model extraction / reverse engineering
Model extraction attacks refer to all attacks aimed at reconstructing the model or information from the training data. In such attacks, prior knowledge of the model's training dataset or access to publicly available parts of it is advantageous to the attacker. For companies that have invested significant resources in developing an AI model, the theft of their model poses a threat to their business foundation. Attackers can copy the model by making numerous queries and storing the respective responses in their own shadow model. Depending on the extent of knowledge regarding a model's internal workings (black-box vs. white-box), various methods can be employed in the context of model extraction to extract specific information about individual data points, class probabilities, or labels, and to enrich custom shadow models. In addition

[11] Souly, A. et al. (2025) Poisoning attacks on LLMs require a near-constant number of poison samples. Available at: https://arxiv.org/abs/2510.07192

to model extraction attacks, which aim to answer the question of exactly how a specific model works, there are similar types of attacks designed to acquire knowledge about a model's internal workings. These include, above all, the membership inference attack, which examines whether a specific data point was part of the training dataset, and the model inversion attack, which attempts to reconstruct the training dataset accordingly. All attack vectors allow relevant and, in some cases, sensitive information to leak from the models and can cause significant downstream damage.

Vulnerabilities in open-weight models

The growing adoption of open-weight models introduces supply chain risks that go beyond the data poisoning threats described above. While data poisoning targets the training process, supply chain attacks can compromise the model itself before it ever reaches an organisation: malicious actors may upload manipulated weights to public repositories, embed backdoors that activate only under specific input conditions, or tamper with the widely used inference frameworks and orchestration libraries that are required to deploy and run these models. Crucially, organisations typically have limited insight into the provenance and integrity of the components they are adopting, making such manipulations difficult to detect. Organisations should therefore treat AI model adoption with the same supply chain diligence applied to other software dependencies, including cryptographic verification of model artefacts and thorough security assessment of all third-party frameworks involved in model serving and integration. Using models exclusively via external APIs avoids some of these risks but introduces its own challenges, as organisations then have no visibility into the underlying model and must place full trust in the security practices of the model provider.[12]

From a consumer perspective, it is equally important to consider how input data is handled after submission. Systems should not only ensure secure processing but also implement clear policies and technical mechanisms for data minimisation, retention, and, where appropriate, the ability to delete or »forget« sensitive information provided by users.

Integration risks

AI systems embedded into existing IT landscapes (e. g., APIs, copilots)

Integrating AI systems into established IT landscapes, through APIs, copilots, or AI-assisted workflows, substantially amplifies the security challenges outlined in the preceding sections. Every new integration point represents an additional attack surface: APIs can serve as vectors for data exfiltration or manipulation, while copilots embedded in productivity environments may inadvertently expose sensitive internal data to external models or third-party services. The complexity of these integrations makes it significantly harder to maintain a comprehensive overview of data flows and access permissions across the entire system, meaning security risks do not simply add up but multiply. Traditional IT security frameworks were not designed with AI-specific threat vectors in mind, making it essential for organisations to revisit their existing

[12] OWASP (2025) OWASP Top 10 for Large Language Model Applications. Available at: https://owasp.org/www-project-top-10-for-large-language-model-applications/

controls and extend them explicitly to cover AI components and their interaction points within the broader IT ecosystem. Retrieval-augmented generation (RAG) architectures, in particular, where AI agents access internal knowledge bases at runtime, pose elevated risks of sensitive information being disclosed without adequate access controls in place.[13]

Increased exposure through automation and agent-based systems

The security stakes escalate even further when AI is deployed in the form of autonomous agents capable of perceiving their environment, planning multi-step actions, and executing decisions with minimal human oversight. As Bitkom's whitepaper »Security of AI Agents« (2025)[14] comprehensively details, agentic systems introduce a qualitatively different risk profile compared to passive AI tools: agents do not merely advise, they act. They can send emails, trigger workflows, access external systems, and interact with other agents, meaning that a single successful attack or misconfiguration can have immediate, real-world consequences that propagate through automated processes before any human intervention is possible. The whitepaper identifies 16 distinct attack vectors specific to agent-based systems, ranging from prompt injection and tool misuse to privilege escalation and identity spoofing. A large-scale red-teaming study cited therein found that 86 percent of tested AI agents executed critical or harmful actions under at least one attack scenario, and over 80 percent of successful attacks relied solely on text manipulation, requiring no technical access or code changes whatsoever.[15] In multi-agent systems, these risks compound further: each agent-to-agent interface is a potential attack vector, cascading failures can propagate across the entire processing pipeline, and accountability becomes considerably harder to establish. Organisations deploying agent-based AI systems must therefore treat them as security-critical infrastructure from the outset, applying zero-trust principles, least-privilege access controls, continuous monitoring, and mandatory human-in-the-loop mechanisms for high-stakes decisions, as recommended both by the Bitkom whitepaper and recent AI governance research.[16]

Operational risks

Overreliance on AI outputs

Even when an AI system is technically secure and correctly integrated, the way in which humans interact with it can introduce significant security risks in its own right. Overreliance on AI outputs, meaning the tendency to accept model-generated results without sufficient critical scrutiny, can lead to consequential decisions being made based on incorrect, manipulated, or out-of-context information. This risk is particularly acute in high-stakes domains such as compliance, financial decision-making, or security monitoring, where a single unchallenged AI output can have far-reaching consequences. Humans therefore remain a decisive component in the security chain: technical safeguards can reduce the likelihood of harmful outputs, but they cannot

[13] Raza, S. et al. (2025) TRiSM for agentic AI: A review of trust, risk, and security management in LLM-based agentic multi-agent systems. Available at: https://arxiv.org/abs/2506.04133
[14] Bitkom e. V. (2025) Security of AI Agents. Available at: https://www.bitkom.org/Bitkom/Publikationen/Security-of-AI-Agents
[15] Zou, A. et al. (2025) Security challenges in AI agent deployment: Insights from a large scale public competition. Available at: https://arxiv.org/abs/2507.20526
[16] Kraprayoon, J. et al. (2025) AI agent governance: A field guide. Available at: https://arxiv.org/abs/2505.21808

substitute for the judgement of informed and critically engaged users. Organisations must actively counteract overreliance through clear usage guidelines, role-specific training, and process designs that keep humans meaningfully in the loop rather than reducing them to passive recipients of AI-generated recommendations.

Lack of transparency and monitoring
Overreliance is further compounded when organisations lack the visibility needed to understand how their AI systems behave in production. Without adequate monitoring, anomalous outputs, unexpected behavioural shifts, or early indicators of an ongoing attack can go undetected for extended periods, significantly increasing the potential damage. This points to a broader challenge: AI systems must be observable throughout their entire operational lifecycle, not only at the point of initial deployment. End-to-end testing plays a central role here, as it allows organizations to validate not just individual components in isolation but the behaviour of the complete system under realistic conditions, including the interfaces between models, tools, data sources, and human users.

Organisational risks
Organisational shortcomings can significantly weaken AI security, even when technical safeguards are in place. In many cases, the issue is not a complete lack of awareness, but rather that knowledge about AI risks is unevenly distributed or not effectively translated into day-to-day practices. Employees and developers may be familiar with general security principles but lack a concrete understanding of how these apply to AI-specific contexts, leading to gaps in implementation.

At the same time, unclear responsibilities for AI security create additional risks. When it is not well defined who is accountable for tasks such as monitoring, risk assessment, or incident response, security-relevant issues can fall between teams or be addressed inconsistently. This is particularly challenging in cross-functional environments where AI systems sit at the intersection of IT, data, legal, and business units.

»Security theatre« Implementing point solutions without systemic protection
Security theatre arises when organisations respond to AI-specific threats with visible but narrow countermeasures that are not embedded in a broader security and governance framework. In such cases, individual controls may appear reassuring while underlying and correlated risks related to data access, model behaviour, third-party dependencies, escalation, and monitoring remain insufficiently addressed. This can create misplaced confidence, weaken risk awareness, and delay the implementation of more resilient lifecycle-wide protection.

Recommendations for the implementation across the AI lifecycle

AI security risks often fall between functions because no single team owns the full lifecycle. Data teams may own inputs, ML teams own training, product teams own deployment, and security teams review only selected controls. Effective implementation thus requires explicit lifecycle-based ownership so that accountability for data integrity, model behaviour, deployment architecture, incident response, and post-deployment monitoring is clearly assigned rather than dispersed across interfaces

and cross-functional teams speak the same language. This section therefore explores recommendations for implementation across both technical and organisational dimensions.

Technical

Methods/Tools

Core technical safeguards should focus on protecting data, controlling access, and ensuring input integrity across the AI lifecycle.

- data encryption (at rest and in transit)
- secure data pipelines and access control
- anonymisation / pseudonymisation where possible
- data validation and filtering mechanisms

Deployment

Control over data, models, and deployment
The degree of control an organisation retains over its AI components has direct security implications. Deploying models within one's own infrastructure, rather than relying exclusively on external APIs, provides significantly greater visibility into model behaviour, data flows, and access patterns, and reduces dependency on the security practices of third-party providers. This control comes with operational overhead that must be weighed carefully against the associated benefits. Complete retraining of a model from scratch is rarely a proportionate response to emerging security or quality concerns and is resource-intensive in most organisational contexts. Targeted fine-tuning of an existing base model, by contrast, offers a more practical path to adapting model behaviour while retaining control over the training data and process. Organisations should evaluate their deployment architecture with this trade-off explicitly in mind, recognizing that greater ownership of the model lifecycle, while demanding, substantially strengthens the overall security posture.[17]

Incident response and model versioning
Even well-tested systems can exhibit unexpected or harmful behaviour in production, whether due to adversarial attacks, distributional shifts, or unforeseen interactions between components. Organisations must therefore establish a clear concept for what happens when a model or agent misbehaves, before such an incident occurs. This includes maintaining versioned snapshots of models and agent configurations so that a controlled rollback to a previously validated state is possible without lengthy redeployment cycles. Incident response plans for AI systems should define escalation paths, specify who is authorized to intervene and in what form, and ensure that affected components can be isolated or shut down rapidly. Treating model and agent

[17] Bommasani, R., Hudson, D. A., Aditi, E., Altman, R., Arora, S., Bernstein, S., et al. (2022). On the opportunities and risks of foundation models. Stanford Center for Research on Foundation Models. Available at: https://arxiv.org/abs/2108.07258

versioning as a standard operational practice rather than an afterthought significantly reduces both the impact and the recovery time of security incidents.[18]

Monitoring & Testing

Continuous monitoring and testing are essential to ensure the security and reliability of AI systems across their lifecycle. Given the dynamic nature of data inputs, model behaviour, and system integrations, organisations must be able to detect anomalies, validate system integrity, and identify potential attacks or failures early. This requires combining robust pre-deployment testing with ongoing monitoring in production environments.

- detection of anomalous data patterns (poisoning attempts)
- integrity checks for incoming data streams
- end-to-end testing and continuous monitoring

Testing AI systems only at the point of initial deployment is insufficient. Particularly for integrated AI systems and agentic architectures, where multiple components interact across complex pipelines, end-to-end testing is essential to validate the behaviour of the complete system under realistic, use-case-specific conditions. Standard benchmarks alone are not adequate for this purpose: they rarely reflect the edge cases, adversarial inputs, or domain-specific failure modes that matter most in practice. Organisations should therefore complement benchmark evaluations with scenario-based tests that closely mirror actual usage contexts, including the interfaces between models, tools, external data sources, and human users. Testing must also be understood as a continuous practice rather than a one-time gate, since model behaviour can shift over time as inputs, integrations, or the broader environment evolve. Continuous monitoring in production is the necessary complement to pre-deployment testing, enabling organisations to detect anomalous outputs, unexpected behavioural changes, or early indicators of an attack before significant damage occurs.

Documentation

Comprehensive documentation is a key enabler for transparency, traceability, and effective risk management in AI systems. It allows organisations to understand how data is sourced, processed, and used, and supports auditing, accountability, and incident response. Proper documentation also facilitates coordination across teams and ensures that assumptions and dependencies remain visible over time.

- data sheets (origin, quality, preprocessing steps)
- data lineage and provenance tracking

[18] Kraprayoon, J., et al. (2025). AI agent governance: A field guide. Available at: https://arxiv.org/abs/2505.21808

Organisational

The organisational embedding of AI security constitutes a key prerequisite for the secure and sustainable deployment of AI systems; however, in practice it is frequently not addressed with a sufficient degree of methodological rigor. Technical safeguards, taken in isolation, provide limited assurance if they are not supported by clearly defined accountability structures, robust governance arrangements, and established organisational processes. Only the deliberate and coordinated interaction of these elements enables a coherent and resilient management of security-relevant risks across the entire AI lifecycle.[19]

Rigor in organisational embedding of AI security is especially important as a central obstacle to implementation is not a lack of awareness of individual risks, but a lack of organisational translation. Technical teams, legal experts, governance functions, risk managers, procurement, and business owners often work with different understandings, concepts and assessment criteria, requiring regular cross-functional reviews across the lifecycle and adaptive governance mechanisms. As AI systems evolve rapidly through new models, integrations, agentic capabilities, and external dependencies, trustworthy and responsible AI cannot be implemented through isolated controls alone. Rather, it requires continuous alignment across functions to build a shared risk understanding, and governance structures that can adapt as the technology and its threat landscape change. Therefore, well-designed governance remains ineffective if organisations do not allocate sufficient expertise, testing capacity, and operational resources to maintain it over time. Governance thus becomes ineffective when control reviews, documentation, and decision processes operate on a slower cycle than the models, tools, and attack patterns they are meant to govern.

Furthermore, it is evident that established IT and risk management frameworks cannot be applied to AI systems without substantive adaptation. The data-centric nature of AI systems, combined with the continuous evolution of models over time, gives rise to specific organisational and control requirements. These challenges are further amplified by the tight integration of AI systems into existing IT environments, necessitating targeted organisational adjustments to adequately address the distinct security requirements associated with AI deployments.[20] Additionally, implementation becomes more difficult when organisations rely on third-party models and tools without sufficient visibility into their security controls, update cycles, failure modes or contractual responsibilities.

[19] Federal Financial Supervisory Authority (BaFin) (2026) Guidance on ICT Risks in the Use of Artificial Intelligence in Financial Institutions. Available at: https://www.bafin.de/SharedDocs/Veroeffentlichungen/EN/Meldung/2025/meldung_2025_12_18_orientierungshilfe_ikt_risiken_en.html

[20] KPMG Germany (2025) Cyber Security im Wandel: Wie Unternehmen die notwendige Transformation gestalten. Available at: https://kpmg.com/de/de/themen/daten-und-technologie/cybersecurity-im-wandel.html#contact-carousel-2fe546137e-item-94bc40f2c5-tabpanel

Roles & Responsibilities

A clear and unambiguous definition of roles and responsibilities across the entire AI lifecycle forms a foundational element of effective AI security governance. In practice, security-relevant weaknesses often arise not primarily from the absence of individual controls, but from unclear ownership and insufficiently defined interfaces between key organisational functions, including IT, data science, business units, and governance-related roles.

The explicit allocation of responsibility for data, models, and operational activities is therefore essential to ensuring that security requirements are addressed in a consistent and lifecycle-spanning manner. Early organisational alignment of these responsibilities supports the systematic consideration of security aspects during the design and ongoing evolution of AI systems, rather than deferring them to reactive measures during live operations.

Governance Structures

A resilient governance framework establishes the structural conditions necessary for a controlled and transparent use of AI systems. This includes, in particular, binding policies governing data usage and further processing, the integration and modification of models, as well as the selection, engagement, and oversight of external service providers.

At the core of such a framework is a risk-oriented governance design, ensuring that security-relevant considerations are embedded in a structured manner within decision-making and approval processes, rather than being addressed in isolation or on an ad hoc basis. The early establishment of clear governance structures provides a stable foundation for the controlled development and scalable deployment of AI applications under traceable and well-defined conditions.

Escalation Mechanisms

Even where governance structures are well established and security requirements are clearly defined, the occurrence of security-related incidents cannot be entirely ruled out. Against this backdrop, formally defined escalation mechanisms are required, specifying responsibilities, decision-making authorities, and communication pathways in a clear and binding manner.

This is of particular relevance for AI systems, as security-related malfunctions or targeted attacks may take effect with immediate impact and, due to systemic interdependencies, propagate rapidly across interconnected processes and systems. A clearly articulated escalation framework supports coordinated response capabilities and enhances transparency in the handling of security-relevant incidents.

Processes

Given the dynamic nature of AI systems, AI security should not be regarded as a static state, but rather as a continuous organisational and control function. Accordingly, organisations should establish processes that ensure the regular review of data sources, models, and system dependencies throughout the lifecycle of AI applications.

Close integration of these processes with existing risk management and control mechanisms is essential to assess AI-related risks within the broader context of enterprise-wide risk governance, rather than treating them in isolation. This approach strengthens the consistency of risk management practices and enhances the traceability of security-relevant decisions.

Awareness & Culture

In addition to structural and procedural measures, the human factor plays a distinct and material role in the security posture of AI systems. The use, interpretation, and further development of AI applications are largely shaped by human judgement, meaning that insufficient awareness of security risks can undermine otherwise well-designed governance and control frameworks.

Training and awareness-raising initiatives should therefore be understood as an integral component of the organisational security architecture; they contribute to establishing a shared understanding of security-relevant considerations and support the long-term embedding of responsible AI usage practices across the organisation.

Excursus: From Security to Safety – Continuous Assurance Across the Lifecycle
The lifecycle-oriented security measures outlined above focus on safeguarding AI assets (data, models, infrastructure, and operations) against intentional misuse and compromise. However, even a well-protected system can still cause unintended harm if its behaviour under real-world conditions is not systematically understood, constrained, and monitored. To complement security and support trustworthy AI overall, continuous safety assurance is needed: it spans the same lifecycle phases – data, models, deployment, and monitoring, but with a different primary question: how to keep the residual risk of harm acceptably low over time.

While security focuses on protecting AI systems against intentional attacks, safety is concerned with preventing and limiting unintentional harm to people, the environment, and critical assets. For high-impact AI systems, inherent uncertainty in models, data, and operating conditions means that safety cannot be ensured exclusively at design time. New usage patterns, data distributions, and updates can expose failure modes that were not fully observable during design. Continuous safety assurance therefore starts with explicit safety goals, an acceptable level of residual risk, and a clearly defined operational context. From these, safety-relevant properties and metrics are derived, fault-tolerant and monitored system architectures are designed, and targeted verification and validation activities collect focused evidence. This evidence is integrated into structured, dynamic safety cases that make assumptions, limitations, and residual risks transparent and connect them to concrete artefacts (requirements, models, tests, and operational data), in line with emerging AI regulation and sector-specific safety standards. Safety thus becomes a runtime concern: monitoring of safety-relevant operational indicators tracks whether key assumptions still hold and whether safety-relevant metrics remain within acceptable bounds. Deviations trigger controlled interventions such as safe modes or updates, and feed back into the dynamic safety case to re-evaluate residual risk.

4 Fairness

Fairness and bias are long-standing risks in the use of AI. While there are intersubjective agreements on fairness, it is important to understand that there can be significant differences in how fairness is perceived by groups or individuals. Group fairness requires an AI system to produce comparable outcomes across all demographic groups, whereas individual fairness demands that similar individuals are treated similarly. Reaching an intersubjective consensus on fairness is particularly challenging in the latter case. Our cultural and socioeconomic background plays a key role in shaping our understanding of fairness, as well as our shared values and regulatory frameworks. This applies to societies as well as to organisations, which can be considered as »social systems«.

Thus, discussing fairness is not solely a technical question, but it also requires societal and normative debates and decisions about its definition, target distributions, and accepted trade-offs. This is particularly important and complex with the advent of GenAI. There are various reasons for this, one of which is that AI can act as an epistemic agent (i. e. an active knowledge worker).

Measuring bias is less a purely technical challenge than a question of selecting appropriate metrics. As a prerequisite, organisations must first define the sensitive attributes relevant to their use case – that is, the dimensions along which discrimination may occur. Some of these are legally defined, for example under European anti-discrimination frameworks (e. g. age, gender, ethnicity, or sexual orientation).

In practice, fairness can often be operationalized through two core approaches: **demographic parity** (equal selection rates across groups) and **equal opportunity** (equal success probabilities for equally qualified individuals, regardless of group membership). However, these concepts can be in tension, particularly when pools of people are imbalanced across groups.

There is no universally correct solution. The choice of metrics should therefore be deliberate, context-specific, and transparently documented.

Relevance and Risks

Bias in AI systems has long been studied. Bias can be introduced into AI systems at various stages of the AI lifecycle and in various forms. According to ISO/IEC TR 24047:2021, the sources of bias in AI systems can be categorized into human cognitive bias, data bias, and bias introduced through the engineering and modelling process, with the various subtypes influencing each other throughout the AI lifecycle. The rise of generative AI systems often reinforces existing biases and introduces new challenges.

Some of them are[21]:

Data bias during training, fine-tuning and usage

Data bias can be introduced during training and fine-tuning of AI systems: Data has to be recorded before it can be processed, and therefore it is inherently historical. As a result, historical and/or societal biases are included in this data, as well as stereotypes. Targeted groups (depending on the purpose of the AI system) may be underrepresented in the data. If AI systems are trained and fine-tuned on this data, they will also include the »inherited« bias or underrepresentation from the data. In addition to bias introduced during training or fine-tuning, bias can be amplified, or introduced, during usage of the AI system, e. g. if existing or new biases are introduced into the model through user interactions, and the model learns from these – intentional or unintentional biases.

Human cognitive bias

Human cognitive bias refers to the subjective construction of reality by individuals, which may influence the data. This means that the data used by AI systems may be »pre-processed« by human interpreters (e. g. data curators).

Model and training bias

Model and training bias pose a key risk, as AI systems can inherit and reinforce patterns present in training data. This may lead to the amplification of existing distortions in decision-making. ISO/IEC TR 24027:2021 provides a structured definition and overview of such bias types in AI systems.

Risks and harmful consequences of bias in AI systems

As a result of bias, there are many different risks when evaluating the outcomes and recommendations of an AI system. These risks encompass sub-optimality, performance, discrimination, but also regulatory non-compliance, to name just a few.

Underrepresentation as a risk

Underrepresenting certain groups in the results and decision-making suggestions, based on data not equally representing all groups, may lead to incompleteness in the results and therefore distort the overall results and quality of an AI system.

Disparate system performance as a risk

Due to biased or underrepresented data, AI systems can perform better or worse for specific groups, i. e., lead to better results for one group than another. This does not

[21] For a comprehensive overview see for example: IBM (n.d.) Document download. Available at: https://www.ibm.com/downloads/documents/us-en/10a99803d8afd656 (Accessed: 27 April 2026)

necessarily have to be discriminatory but can still lead to discrimination if no countermeasures are taken.

Discrimination as a risk

As a consequence of bias, no matter where it originates, AI systems can generate content or make decisions that may discriminate or unfairly represent groups or individuals. This is not only unlawful in many countries (e. g. related to anti-discrimination requirements) but also considered unfair in most cases. AI systems making suggestions for decisions, may unfairly create an advantage for one group or individual over another based on bias (from data or humans) – or even worse: create an explicit disadvantage for one group or individual compared with another.

One should, however, be aware that bias is not necessarily bad in all cases, depending on the purpose of the AI system. For example, AI recommendation engines can increase the visibility of underrepresented groups, but this may result in other groups being overlooked, which might not be considered unfair. Thus, context awareness is crucial.

Non-compliance with human rights as a risk

When AI systems discriminate, e. g., related to skin colour, sex, sexual orientation, religion, political or other opinion, and more, they would be non-compliant with the Universal Declaration of Human Rights and harm fundamental human rights.

Regulatory risk

Bias, especially when the consequence is discrimination, can lead to non-compliance with existing regulatory frameworks, e. g., the EU AI Act. This can result in significant financial sanctions for providers of such AI systems.

Dark Patterns

Another overarching risk is the intentional or unintentional existence of dark patterns in AI systems. These are deceptive techniques used in digital tools to manipulate users' behaviour without their knowledge or consent. This is becoming increasingly important with general-purpose AI, as it incorporates more deceptive techniques than just traditional UX dark patterns. Thus, the option to opt out can become much more difficult, even considering the disadvantages for individuals. Therefore, detecting dark patterns in AI systems is also an organisational problem that needs to be addressed.[22]

If any of these risks occur, they can lead to significant reputational damage for the provider of AI systems, financial damage, and ultimately even criminal prosecution. Therefore, and as AI systems are based on data, which in most cases contains some bias in one way or another – the goal for anyone implementing and operating AI systems must be to detect bias as much and as early as possible, introduce mitigation

[22] Gray et al. (2024) provide an Ontology for Dark Pattern detection: https://ontology.darkpatternsresearchandimpact.com/

actions and make bias transparent through adequate AI governance functions. There are metrics and algorithms available that help (e. g. open source tools such as ↗https://ai-fairness-360.org/) identify, document, and mitigate bias in generative AI models throughout their lifecycle.

Recommendations for the Implementation Across the AI Lifecycle

The fundamental recommendation to avoid and reduce bias in AI systems is to consider fairness »by design«. It will not be sufficient to »remedy« bias once it has occurred, but it needs to be taken into account from the very beginning and then along the full lifecycle.

Technical

Multiple areas are affected and are relevant for detecting, mitigating and monitoring bias:

Data

Related to data, many actions can be taken to avoid or mitigate biases[23]:

Fairness parameters / dimensions in alignment with the deployment context and relevant stakeholders should be defined in advance (e. g. »what is fair«, sensitive attributes or group splits) and monitored throughout the lifecycle of an AI system.

Data provenance, i. e., metadata for data, should be introduced as well as data sources should be selected carefully so that it is possible to trace and track data origins at a later stage. Fairness should be considered from the start of data collection, as it is significantly harder to correct bias later. This means that data needs to be evaluated, filtered, and curated. For example, practitioners should be aware of Eurocentric or otherwise culturally skewed data sources (e. g., web, Reddit, news); open-web data reflects existing societal imbalances.

Appropriate metrics should be applied to the data to detect bias and required countermeasures, again throughout the lifecycle.

Ensuring diversity across multiple dimensions (e. g. dimensions of gender, age, religion, socio-economic status, language/culture) in training data is a good starting point to avoid bias.

If data needs to be removed to create a balanced dataset, this should be done in a deliberate and documented way as removing biased data carries the risk of degrading model performance.

[23] Görge, Rebekka, et al. »Textual Data Bias Detection and Mitigation – An Extensible Pipeline with Experimental Evaluation.« *arXiv preprint arXiv:2512.10734* (2025).

Datasets that are optimised for performance should be carefully evaluated, as they can lead to overly homogenous outputs (diversity collapse) and cause problems in areas that require creativity or broader representation.

Augmenting data to help balance representation is another option, but one needs to be aware that this may itself introduce or spread bias (grammar- and context-aware augmentation approaches might be able to mitigate that risk).

Models und model training

To mitigate bias, careful consideration must be given to the selection of models used within an AI system. In particular, it is important to assess whether models are trained on broad, unverified data sources (e. g. large-scale web scraping) or whether they provide sufficient transparency regarding their training data, thereby enabling a systematic evaluation of potential biases.

Another aspect in this context is to check the origin of the models: Are they potentially influenced by political censorship of training data (e. g. Chinese models)?[24]

Using end-to-end open source models (open data, open models) to mitigate bias over the full AI-lifecycle, e. g. Apertus[25], may be a viable option. Nevertheless, Open Source models also can be biased and should not be used unchecked.

Moreover, using multilingual training to reduce model bias (training on a mix of languages), has shown measurable improvements compared with monolingual training.[26]

Fine-tuning the model can also help reduce bias – based on the detection of bias in the data and model.[27]

A new countermeasure could be using mechanistic interpretability research, which is an emerging tool for bias analysis. It is currently used more widely in academia than in the business sector, but it is worth keeping an eye on developments in this area.

Hence, regarding data, a particular attention should be paid to the following two aspects:

- Be aware of benchmarks, as they themselves can carry biases and yield conflicting results depending on their design; use multiple benchmarks, incorporate human validation, and interpret results critically.
- Be aware of »one-size-fits-all« fairness solutions; model adaptation to specific domains and use cases is often necessary.

[24] Pan, J., & Xu, X. (2026). Political censorship in large language models originating from China. *PNAS nexus*, *5*(2), pgag013., https://doi.org/10.1093/pnasnexus/pgag013.
[25] Apertus, P., Hernández-Cano, A., Hägele, A., Huang, A. H., Romanou, A., Solergibert, A. J., ... & Schlag, I. (2025). Apertus: Democratizing Open and Compliant LLMs for Global Language Environments. *arXiv preprint arXiv:2509.14233*.
[26] Nie, S., Fromm, M., Welch, C., Görge, R., Karimi, A., Plepi, J., ... & Flek, L. (2024, August). Do Multilingual Large Language Models Mitigate Stereotype Bias?. In *Proceedings of the 2nd Workshop on Cross-Cultural Considerations in NLP* (pp. 65-83).
[27] Görge, Rebekka, et al. »Textual Data Bias Detection and Mitigation--An Extensible Pipeline with Experimental Evaluation.« *arXiv preprint arXiv:2512.10734* (2025).

Development & Deployment

During development and deployment of AI systems, additional countermeasures to avoid or mitigate bias are required.

Following through from training and fine-tuning, it is important to continue to define, review, and update the fairness criteria and target metrics for your solution during development, testing, and before deployment, tied to the specific use case and domain, as they may differ from one area to another. For example, a general-purpose model requires different fairness standards than a domain-specific system (e. g., HR, healthcare, customer service), which means use case context is essential.

Using methods from human-centred-design (e. g. personae, scenarios) during the development process, may be a helpful approach to incorporate as many relevant stakeholder perspectives as possible.

This should lead to a shared and precise taxonomy of relevant bias types and fairness criteria for the AI system and the use case within the organisation. In this way, issues resulting from inconsistent terminology across teams can be avoided from the outset. In addition, the ethical standards for an AI system should clearly be defined, ideally together with its purpose and expected value.

An additional measure to mitigate bias, where it cannot be fully addressed at the data, training, or development stage, is the use of sensitive prompting in generative AI systems. By explicitly instructing the model to avoid stereotyping or biased assumptions, the likelihood of biased outputs can be reduced. Such prompting strategies provide a practical and accessible mechanism for targeted bias mitigation at the point of use.

Monitoring

When an AI system is introduced, the work on removing, avoiding, or mitigating bias is not finished. During operation, the system must be continuously monitored to determine whether new biases emerge, existing (and not completely removed) bias increases, or whether it persists.

Therefore, introducing proper AI governance, not only for fairness but more broadly, is an absolute necessity when organisations want to use and provide AI in a responsible way. In some cases, »testing in the wild«, i. e. in production, may be necessary to detect and understand risks that cannot be fully assessed in a lab setting, for example when the system carries public risk at scale.

Overall, planning for AI governance and monitoring is essential, and careful evaluation is necessary, as both monitoring for fairness and mitigation may be quite costly. Mitigation itself as well as measuring mitigation effectiveness can require additional resources for retraining and benchmarking.

A further, meta-level consideration is the potential need to assess evaluation models themselves for bias, given that systems designed to audit other systems may inherently exhibit biases of their own.

Organisational

This paper recommends that standardised, auditable methods for designing, developing, and deploying AI systems and well-defined processes for operating AI systems are implemented to minimise discriminatory features in data, the model and during usage.[28] This should also include feedback loops in all lifecycle stages with stakeholders and users.

This includes the integration of fairness aspects across the entire MLOps lifecycle, enabling the proactive operationalisation of fairness.

Also, the underlying data, information about the model, the applied metrics, the selected sensitive attributes, and mitigation and countermeasures to ensure fairness etc. should be documented and made transparent, for example in model cards / fact sheets about the AI system.

For deployment and operations, an escalation process for upcoming issues, may it be bias or something else, needs to be provided. In addition, an option for employees to opt out of using the AI system, with documented consequences, should be integrated.

For operations, i. e., during usage of the AI system, an AI governance process with appropriate roles & responsibilities must be introduced and managed.

In addition to the technical infrastructure and a sound process foundation, the effective implementation of AI fairness requires robust AI governance (which encompasses the technical and process elements). Alongside management commitment and an overarching AI strategy, organisations must clearly define the necessary resources, including expertise, time and budget allocations, as well as clearly defined responsibilities and roles across organisational levels. Without such coordination, decision-making risks become fragmented and can result in ineffective implementation of AI principles. Only a holistic approach ensures that the strategic goal of »Fair AI« is effectively integrated into day-to-day business operations.[29] Similarly critical is the integration of all functions within an enterprise, including those who manage AI, those who apply AI, those who know the regulatory frameworks, and those who are accountable for security and privacy etc. Only a cross-disciplinary approach to AI governance will lead to success.

Clearly defined roles & responsibilities are one of the cornerstones of AI governance within an organisation. Techniques such as RACI matrices (Responsible, Accountable, Consulted, Informed) can help shape and efficiently set up the different required roles. These roles include stakeholders (i. e., business owners, HR, legal, employees, union councils, work councils, customers, etc.), escalation focal points, as well as those who have operational roles in the AI governance structure. In addition to clearly defined responsibilities, it’s also important to document the intended goals of AI governance and the involved roles, e. g., which specific bias should be monitored and mitigated, as well as practicable measurement approaches.

[28] Hühn, Julia, et al. »Standardizing fairness-evaluation procedures: interdisciplinary insights on machine learning algorithms in creditworthiness assessments for small personal loans.« AI and Ethics (2023)
[29] Horneber, David. »Understanding the Implementation of Responsible Artificial Intelligence in Organizations: A Neo-Institutional Theory Perspective.« Communications of the Association for Information Systems 57 (2025): 185–218.

Table: Types of Bias and Mitigation (Examples)[30,31,32,33]

Pipeline stage	Bias type	Econometric interpretation (Wooldridge lens)	AI manifestation	Example	Main mitigation
Data	Omitted variable bias	Regressors correlated with unobservables in the error term	Sensitive traits or structural disadvantage are indirectly captured through proxies	Credit model uses ZIP code → indirectly captures ethnicity → biased loan approvals	Better covariates, causal reasoning, proxy review, robustness checks, domain knowledge
Data	Measurement bias	Measurement error in (X) or (Y)	Biased labels, weak proxies, noisy annotations, biased human feedback	Healthcare model uses past spending as proxy for health need → underestimates needs of underserved groups	Label audits, improved target design, multiple annotators, construct validation
Data	Sample-selection / representation bias	Non-random sampling or selection into observed data	Underrepresented groups in training, validation, or test data	Facial recognition trained mostly on light-skinned faces → poor accuracy for darker skin tones	Reweighting, stratified sampling, targeted data collection, selection correction
Data	Historical bias	Learned latent structure absorbs socially patterned correlations not explicitly modeled	Historical bias in data is carried through the AI model and manifests itself in AI results	Hiring model trained on historical company data learns to prefer male candidates because past hiring decisions favored men → model replicates and scales this pattern	Careful selection of data sets, filtering and curating of data, replacement with synthetic data
Data	Support / coverage bias	Lack of overlap or weak support in parts of the	Sparse data for rare groups, edge cases,	LLM performs poorly on low-resource	Better coverage, targeted sampling, uncertainty flags,

[30] Barocas, S., Hardt, M., & Narayanan, A. (2023). Fairness and machine learning: Limitations and opportunities. MIT Press. https://fairmlbook.org

[31] Schwartz, R., Vassilev, A., Greene, K., Perine, L., Burt, A., & Hall, P. (2022). Towards a standard for identifying and managing bias in artificial intelligence (NIST Special Publication 1270). National Institute of Standards and Technology. https://doi.org/10.6028/NIST.SP.1270

[32] Suresh, H., & Guttag, J. V. (2021). A framework for understanding sources of harm throughout the machine learning life cycle. In Proceedings of the 2021 ACM Conference on Equity and Access in Algorithms, Mechanisms, and Optimization (EAAMO '21). Association for Computing Machinery. https://doi.org/10.1145/3465416.3483305

[33] Wooldridge, J. M. (2020). Introductory econometrics: A modern approach (7th ed.). Cengage Learning.

Pipeline stage	Bias type	Econometric interpretation (Wooldridge lens)	AI manifestation	Example	Main mitigation
		covariate space; external-validity problem	rare languages, or unusual contexts	languages or niche technical domains	abstention rules, out-of-distribution detection
Model	Aggregation bias	One parameter structure imposed on heterogeneous groups	One model is fit to groups with different conditional relationships or language patterns	Same risk model applied across countries with different economic structures → misestimation	Interaction terms, subgroup models, hierarchical models, context-specific models
Model	Objective / loss bias	Estimated criterion does not match the real policy or welfare objective	Optimising clicks, engagement, or average accuracy instead of fairness, safety, or welfare	Recommender system maximizes clicks → promotes sensational or polarizing content	Better target design, fairness-constrained optimisation, multi-objective training, human oversight
Model	Inductive / regularisation bias	Bias deliberately introduced through structural assumptions to reduce variance	Shrinkage, smoothing, compression, or architectural priors affect groups unequally	Strong regularisation smooths minority patterns → minority signals get ignored	Fairness-aware model selection, subgroup diagnostics, tuning by subgroup
Model	Representation / embedding bias	Learned latent structure absorbs socially patterned correlations not explicitly modeled	Embeddings encode stereotypes, demographic associations, or proxy structure	Word embeddings associate »doctor« with male and »nurse« with female	Embedding audits, debiasing methods, counterfactual testing, representation constraints
Evaluation	Evaluation bias	Estimator evaluated on a non-comparable sample or metric	Benchmarks do not reflect deployment population or mask subgroup disparities	Model shows high average accuracy but performs poorly on elderly users	Group-wise evaluation, external validation, stress testing, intersectional metrics
Deployment	Deployment bias	Model is used in a decision process different from the modeled task	Predictions are treated as decisions; poor human-AI workflow or misuse of outputs	Hiring model used as automatic filter instead of decision support tool	Human factors design, governance, usage constraints, model cards

Pipeline stage	Bias type	Econometric interpretation (Wooldridge lens)	AI manifestation	Example	Main mitigation
Deployment	Feedback / simultaneity bias	Prediction affects the outcome process itself	Predictive policing, recommender loops, dynamic pricing loops, self-reinforcing LLM outputs	Predictive policing sends more patrols → more recorded crime → reinforces model bias	Continuous monitoring, causal evaluation, policy constraints, periodic retraining

5 Explainability

Since the increasing development and deployment of machine learning and AI-powered solutions, explainability has become a prominent issue for researchers and businesses.

Due to the black-box nature of machine learning models, humans are not able to directly inspect or understand the decision-making processes of modern AI systems. To support humans in understanding the decisions and predictions made by AI, it is therefore crucial to deploy explainable AI (XAI) techniques to gain more insights into the computations of modern artificial intelligence.[34]

Instead of interpretability, which aids developers in analysing an AI system's underlying functioning, explainability serves as a pillar of trustworthy AI by enabling stakeholders to understand and calibrate their trust in the system's decisions.[35]

By introducing XAI in companies, AI decisions are made understandable for businesses and stakeholders to guarantee responsible, ethical development and use of AI, while also enhancing trust, reliability, and acceptance of these technologies.[36]
With that, explainability methods can support organisations in meeting various regulations required by the EU AI Act by enhancing system-level transparency, analysing matters of accountability and responsibility of both system and users, and finally enabling effective human oversight.[37]

Relevance and Risks

Explainability is no longer solely a matter of technical best practice. It has become a legal obligation for a growing share of AI deployments in the EU. Under the AI Act (Regulation (EU) 2024/1689), **transparency** is defined as developing and using AI systems in a way that allows appropriate traceability and explainability, while making humans aware that they communicate or interact with an AI system and informing deployers about its capabilities and limitations.[38] Organisations deploying AI systems that fall within the high-risk categories set out in Annex III of the Act, including systems used in employment and HR management, access to education, creditworthiness assessment, and critical infrastructure, are subject to a cluster of directly relevant requirements. Article 13 requires that high-risk AI systems be designed so that their operation is sufficiently transparent to enable deployers to

[34] Arrieta, A. B., Díaz-Rodríguez, N., Del Ser, J., Bennetot, A., Tabik, S., Barbado, A., ... & Herrera, F. (2020). Explainable Artificial Intelligence (XAI): Concepts, taxonomies, opportunities and challenges toward responsible AI. Information fusion, 58, 82-115.
[35] Ali, S., Abuhmed, T., El-Sappagh, S., Muhammad, K., Alonso-Moral, J. M., Confalonieri, R., ... & Herrera, F. (2023). Explainable Artificial Intelligence (XAI): What we know and what is left to attain Trustworthy Artificial Intelligence. Information fusion, 99, 101805.
[36] Sewada, R., Jangid, A., Kumar, P., & Mishra, N. (2023). Explainable artificial intelligence (xai). Journal of Nonlinear Analysis and Optimization, 13(1), 41-47.
[37] Panigutti, C., Hamon, R., Hupont, I., Fernandez Llorca, D., Fano Yela, D., Junklewitz, H., ... & Gomez, E. (2023, June). The role of explainable AI in the context of the AI Act. In Proceedings of the 2023 ACM conference on fairness, accountability, and transparency (pp. 1139-1150).
[38] Regulation (EU) 2024/1689 of the European Parliament and of the Council of 13 June 2024 laying down harmonised rules on artificial intelligence (Artificial Intelligence Act), OJ L, 12.7.2024, Recital 27.

interpret their outputs and use them appropriately, including technical measures to support that interpretation.[39] Article 14 builds on this by requiring that high-risk systems be designed with appropriate human-machine interface tools to enable human oversight, including the ability to monitor operation, detect anomalies, avoid overreliance, and correctly interpret outputs.[40] Article 12 further mandates that high-risk AI systems technically allow for the automatic recording of events over their lifetime, enabling the traceability of system functioning that is a prerequisite for meaningful post-hoc explanation.[41] For organisations deploying general-purpose AI (GPAI) models, Article 53 requires providers to supply downstream integrators with documentation sufficient to enable compliance, which in practice means providing information that makes the model's behaviour interpretable to the next tier of the value chain.[42] Non-compliance carries substantial financial consequences, with penalties reaching up to 35 million euros or 7 percent of global annual turnover for serious violations, and full compliance is required by August 2026.[43] Alongside the AI Act, GDPR Article 22 remains relevant wherever AI systems support automated or semi-automated decisions about individuals, establishing a right to a meaningful explanation that predates and complements the new regulation. Taken together, these obligations mean that XAI is no longer a differentiator: for high-risk applications, it is a baseline compliance requirement.

Beyond legal ramifications, not integrating explainability into an AI-powered product can cause a series of risks that affect not only developers and practitioners, but also stakeholders and end users of AI-based applications, concerning the whole lifecycle of AI-based solutions.[44] Reliance and trust deficits are one of various potential risks to consider, as calibrated trust and reliance are major indicators of successful user experience concerning AI-powered solutions.[45] Safety-related risks also occur when XAI is not implemented, as it can be crucial to make computations transparent to identify bugs, errors and mistakes in the underlying AI model, while also determining whether a decision was made in accordance with procedural and substantive standards to hold responsibility.[46] [47]

XAI is especially important for decision-making in high-stakes situations affecting human lives across diverse sectors, such as healthcare, finance, or the automotive industry.

In high-risk situations, the explainability of AI models is crucial to evaluate and validate recommendations and decisions made by AI, while also ensuring safety, trust and acceptance to facilitate human-machine cooperation.[48]

[39] AI Act, Art. 13(1), Transparency and provision of information to deployers.
[40] AI Act, Art. 14(1), (4) Human oversight.
[41] AI Act, Art. 12(1), (2) Record-keeping.
[42] AI Act, Art. 53(1), Obligations for providers of general-purpose AI models
[43] AI Act, Art. 99 Penalties; see also Regulation (EU) 2024/1689, Art. 113 on the timeline for applicability.
[44] Arrieta, A. B., Díaz-Rodríguez, N., Del Ser, J., Bennetot, A., Tabik, S., Barbado, A., ... & Herrera, F. (2020). Explainable Artificial Intelligence (XAI): Concepts, taxonomies, opportunities and challenges toward responsible AI. Information fusion, 58, 82-115.
[45] Zhang, Y., Liao, Q. V., & Bellamy, R. K. (2020, January). Effect of confidence and explanation on accuracy and trust calibration in AI-assisted decision making. In Proceedings of the 2020 conference on fairness, accountability, and transparency (pp. 295-305).
[46] Sokol, K., & Flach, P. A. (2019). Counterfactual Explanations of Machine Learning Predictions: Opportunities and Challenges for AI Safety. SafeAI@ AAAI, 2301, 1-4.
[47] Doshi-Velez, F., Kortz, M., Budish, R., Bavitz, C., Gershman, S., O'Brien, D., ... & Wood, A. (2017). Accountability of AI under the law: The role of explanation. arXiv preprint arXiv:1711.01134.
[48] Sewada, R., Jangid, A., Kumar, P., & Mishra, N. (2023). Explainable artificial intelligence (xai). Journal of Nonlinear Analysis and Optimization, 13(1), 41-47.

Implementing explainability methods is therefore a valuable practice and a key differentiator for businesses, enabling trust, better decisions, compliance, and effective governance.[49]

Why Explainability Matters

Explainability is a core enabler of trustworthy and responsible AI because it makes AI-supported outputs more understandable, reviewable, and governable for different stakeholder groups.[50] In practice, this is relevant not only for technical validation, but also for calibrated trust, meaningful human oversight, fairness diagnostics, continuous improvement, and compliance readiness.

Explainability of AI supports **trust and acceptance**. Public trust in AI-enabled outputs remains limited. According to the latest Eurobarometer, only 38 percent of Europeans say they trust scientific research and discoveries created with the help of AI, while 25 percent say they distrust them and 35 percent say they neither trust nor distrust them.[51] For organisations, this means that AI-supported decisions cannot rely on technical performance alone; they also need to be understandable enough to be accepted as legitimate by users, customers, employees, and other affected stakeholders.

Explainability strengthens **accountability and human oversight**. AI-related risks do not arise only at the moment of deployment, but across the full lifecycle, including design, development, deployment, operation, and decommissioning.[52] In that setting, explainability helps organisations trace outputs back to relevant assumptions, system behaviour, and documented controls. This is essential if human reviewers are expected to identify limitations, intervene when necessary, and exercise real oversight rather than merely confirm system outputs after the fact.

Explainability improves the detection of **bias and problematic model behaviour**. The landmark *Gender Shades* study demonstrated substantial performance disparities in commercial gender-classification systems: error rates reached 34.7 percent for darker-skinned females, while the maximum error rate for lighter-skinned males was 0.8 percent.[16] The same study found that benchmark datasets were heavily skewed toward lighter-skinned individuals, e. g. at 79.6 percent in the IJB-A dataset and 86.2 percent in the Adience dataset.[53] These findings illustrate why explainability and related diagnostic tools matter: without them, serious disparities can remain hidden behind acceptable aggregate performance metrics.

[49] I. Ahmed, G. Jeon and F. Piccialli, »From Artificial Intelligence to Explainable Artificial Intelligence in Industry 4.0: A Survey on What, How, and Where,« in IEEE Transactions on Industrial Informatics, vol. 18, no. 8, pp. 5031-5042, Aug. 2022, doi: 10.1109/TII.2022.3146552.
[50] National Institute of Standards and Technology (NIST), Artificial Intelligence Risk Management Framework: Generative Artificial Intelligence Profile (NIST AI 600-1), July 2024, available at: https://nvlpubs.nist.gov/nistpubs/ai/NIST.AI.600-1.pdf
[51] European Commission / Eurobarometer (2025) European citizens' knowledge and attitudes towards science and technology, survey detail page. Available at: https://europa.eu/eurobarometer/surveys/detail/3227.
[52] NIST, Artificial Intelligence Risk Management Framework: Generative Artificial Intelligence Profile (NIST AI 600-1), pp. 1–3. (NIST Veröffentliche Serien)
[53] Joy Buolamwini and Timnit Gebru, »Gender Shades: Intersectional Accuracy Disparities in Commercial Gender Classification,« Proceedings of Machine Learning Research 81 (2018), available at: https://proceedings.mlr.press/v81/buolamwini18a.html.

Also, explainability improves **correction capacity and human-AI performance**. In a 2024 *Scientific Reports* study based on preregistered experiments with domain experts, explainable AI improved task performance by 7.7 percentage points in a manufacturing task and 4.7 percentage points in a medical task compared with black-box AI.[54] In the manufacturing experiment, 73.1 percent of domain experts supported by explainable AI even outperformed the standalone AI system.[55] These findings suggest that well-designed explanations can improve how humans detect errors, validate outputs, and collaborate with AI systems in real decision environments.

Finally, explainability improves **compliance readiness and defensibility**. In IBM's 2023 Global AI Adoption Index, 83 percent of IT professionals at enterprises currently exploring or deploying AI said it is important for their business to be able to explain how their AI arrived at a decision. In the same survey population, 83 percent said lifecycle monitoring is important, 82 percent highlighted brand integrity and customer trust, and 81 percent emphasized external regulatory and compliance obligations.[56] These figures indicate that explainability is already treated in practice as part of governance, documentation, and control rather than as a purely abstract ethical aspiration.

From a policy and implementation perspective, explainability should therefore be understood as a **risk-based control**. Organisations should define the required degree of explainability based on use case, impact, and stakeholder group; embed explainability into lifecycle governance and documentation; and test explanation methods for their actual usefulness in human oversight and decision review.[57]

Risks: Proven Implications of Non-Explainability

Where AI systems are insufficiently explainable, organisations face predictable and increasingly well-documented risks. These include weaker trust calibration, undetected bias, reduced ability to identify and correct failures, weaker governance, greater legal and reputational exposure, and lower implementation performance.[58] Non-explainability should therefore be understood as a material organisational risk, especially where AI systems influence decisions with operational, economic, or societal consequences.

An important implication of non-explainability is **reduced trust and weaker adoption**. Opaque systems make it harder for stakeholders to understand whether outputs should be accepted, questioned, or escalated. This can result either in underuse of AI systems or in poorly calibrated overreliance. Enterprise evidence underlines the relevance of this issue: 82 percent of organisations exploring or deploying AI consider

[54] Senoner et al., »Explainable AI improves task performance in human–AI collaboration,« Scientific Reports (2024), available at: https://www.nature.com/articles/s41598-024-82501-9.
[55] Burton, Simon and Herd, Benjamin (2023) Addressing uncertainty in the safety assurance of machine-learning. Frontiers in Computer Science 5 DOI: 10.3389/fcomp.2023.1132580
[56] IBM (2023) Global AI Adoption Index. Available at: https://newsroom.ibm.com/2024-01-10-Data-Suggests-Growth-in-Enterprise-Adoption-of-AI-is-Due-to-Widespread-Deployment-by-Early-Adopters
[57] National Institute of Standards and Technology (NIST), Artificial Intelligence Risk Management Framework: Generative Artificial Intelligence Profile (NIST AI 600-1), July 2024; NIST, Artificial Intelligence Risk Management Framework: Generative Artificial Intelligence Profile (NIST AI 600-1), pp. 1–3
[58] National Institute of Standards and Technology (NIST), *Artificial Intelligence Risk Management Framework: Generative Artificial Intelligence Profile (NIST AI 600-1)*, July 2024,

brand integrity and customer trust important in relation to AI trust and explainability.[59]

Also, the persistence of **hidden bias and unfair outcomes** is noteworthy. Where the drivers of model outputs remain opaque, organisations are less able to identify skewed patterns, proxy effects, or harmful data dependencies. The quoted *Gender Shades* evidence remains highly instructive here.[60] Without sufficient explainability, these disparities are harder to identify, challenge, and remediate because influential features and decision pathways remain obscured.

Another key risk is the reduced **ability to detect, diagnose, and correct failures** via prevailing non-explainability. The 2024 *Scientific Reports* study demonstrates the inverse effect directly: when domain experts were supported by black-box AI rather than explainable AI, performance was measurably worse.[61] This suggests that the absence of usable explanations can reduce organisations' ability to find errors, validate outputs, and improve systems over time.

Another implication is **weaker accountability and governance**. AI risks emerge across the full lifecycle, not only at the point of deployment.[62] If systems are opaque, organisations struggle to connect outputs to training conditions, design choices, model updates, and deployment contexts. That weakens internal review, incident analysis, and the substantive exercise of human oversight. This is reinforced by enterprise survey findings showing that 83 percent of organisations exploring or deploying AI view lifecycle monitoring as important.[63]

What business greatly fear is a **considerable legal, institutional, and reputational exposure**. The Dutch childcare benefits scandal is a prominent public example of what can happen when public-sector decision processes become opaque, discriminatory, and weakly contestable. According to an official European Commission briefing, around 26,000 parents were wrongly affected, and the Dutch government announced compensation of 30,000 euros for affected parents, alongside broader debt relief measures.[64] The parliamentary inquiry *Unprecedented Injustice* documented profound failures of proportionality, legal protection, and administrative fairness.[65] The broader lesson is that insufficient explainability and contestability can transform technical opacity into large-scale harm, financial remediation costs, and loss of institutional legitimacy.

Lastly, an implication considered by practitioners is the **strategic underperformance in AI transformation**. IBM's 2023 survey found that 42 percent of enterprise-scale companies reported active AI deployment, while another 40 percent were actively

[59] IBM, *Global AI Adoption Index 2023*
[60] *Joy Buolamwini and Timnit Gebru, »Gender Shades: Intersectional Accuracy Disparities in Commercial Gender Classification,« Proceedings of Machine Learning Research 81 (2018)*
[61] Senoner et al., »Explainable AI improves task performance in human – AI collaboration,« *Scientific Reports* (2024)
[62] NIST, *Artificial Intelligence Risk Management Framework: Generative Artificial Intelligence Profile (NIST AI 600-1)*, pp. 1–3
[63] Joy Buolamwini and Timnit Gebru, »Gender Shades: Intersectional Accuracy Disparities in Commercial Gender Classification,« *Proceedings of Machine Learning Research* 81
[64] European Commission, ESPN Flash Report, *The Dutch childcare benefit scandal: a cautionary tale for the use of algorithms in social policy*, 2021, available at: https://ec.europa.eu/social/BlobServlet?docId=24723&langId=en.
[65] Dutch House of Representatives, *Unprecedented Injustice* (*Ongekend onrecht*), Parliamentary Inquiry Report on the Childcare Benefits Affair, available at: https://www.houseofrepresentatives.nl/sites/default/files/atoms/files/verslag_pok_definitief-en-gb.docx.pdf.

exploring AI.[66] At the same time, explainability, monitoring, trust, bias minimisation, and compliance all ranked as highly important concerns.[28] This points to a structural tension: organisations are deploying AI at scale while still requiring stronger mechanisms for oversight and control. Where explainability is weak, that tension can slow scaling, weaken internal adoption, and reduce the business value realized from AI investments.

From a policy and implementation perspective, organisations should treat non-explainability as a **risk factor in its own right**, especially in higher-impact systems. They should require explanation methods that are usable for the relevant stakeholder, document limitations and uncertainty, and establish escalation paths when outputs cannot be meaningfully interpreted or contested. In higher-risk contexts, explainability should be linked explicitly to oversight, auditability, and incident response.[67]

Limitations of Explainable AI

Besides all advantages that XAI methods can bring to organisations, there are certain limitations to XAI practices that need to be considered.

While post-hoc methods like LIME or SHAP are prominent for being easily applicable on AI models to create an explainable approximation of their outputs, these methods often produce plausible-looking but unfaithful explanations lacking accuracy, stability, and reliability.

Furthermore, explanations generated by different XAI methods vary greatly, which makes it difficult not only to interpret them but also to assess their quality in order to rely on explanations for critical decision-making in the real world.[68]

Incorrect or incomplete explanations of AI predictions pose a severe risk, leading to incorrect understanding and flawed reasoning, thereby putting high-stakes decisions and entire projects at risk. Users blindly relying on such false explanations can also lead to them neglecting their critical scrutiny, which worsens the overall human-AI collaboration.[69]

Apart from incorrect explanations, incorrect interpretations of explanations by users pose another important limitation of XAI methods. Although an explanation of an AI model may be factually correct, it is important to ensure a certain level of intelligibility and clarity of produced explanations for decision-makers with different backgrounds.[70] This shows another challenge of XAI, as solutions have to be heavily individualised to their respective use cases and target audiences in order to guarantee a positive value.[71]

[66] IBM, *Global AI Adoption Index 2023*
[67] National Institute of Standards and Technology (NIST), Artificial Intelligence Risk Management Framework: Generative Artificial Intelligence Profile (NIST AI 600-1), July 2024; NIST, Artificial Intelligence Risk Management Framework: Generative Artificial Intelligence Profile (NIST AI 600-1), pp. 1–3
[68] Hooshyar, D., & Yang, Y. (2024). Problems with SHAP and LIME in interpretable AI for education: A comparative study of post-hoc explanations and neural-symbolic rule extraction. IEEE Access, 12, 137472-137490.
[69] Bauer, K., Von Zahn, M., & Hinz, O. (2023). Expl (AI) ned: The impact of explainable artificial intelligence on users' information processing. Information systems research, 34(4), 1582-1602.
[70] Aysel, H. I., Cai, X., & Prugel-Bennett, A. (2025). Explainable artificial intelligence: advancements and limitations. *Applied Sciences*, *15*(13), 7261.
[71] Hooshyar, D., & Yang, Y. (2024). Problems with SHAP and LIME in interpretable AI for education: A comparative study of post-hoc explanations and neural-symbolic rule extraction. *IEEE Access*, *12*, 137472-137490.

It is therefore of great importance to evaluate the intelligibility and overall quality of explanations to ensure a profitable implementation of XAI practices.[72]

Recommendations for the Implementation across the AI Lifecycle

However, implementing the right explainability measures remains a greater challenge than it may initially appear. Deploying XAI features does not automatically produce better understanding, more calibrated trust, or safer decisions. Research shows that users frequently overestimate how well they understand a system's behaviour when presented with explanations, a phenomenon known as the illusion of explanatory depth[73]. Even technically trained practitioners are not immune: data scientists have been shown to routinely over-trust and misuse tools such as SHAP, often unable to accurately interpret the visualisations they produce[74]. Overreliance is a further structural risk, as users of AI-assisted decision support tools frequently accept AI suggestions even when those suggestions are wrong, and adding explanations does not reliably reduce this tendency. In some cases, explanations that carry no meaningful information about a model's actual behaviour can generate levels of user trust comparable to genuine ones[75][76]. Organisations should therefore not treat the deployment of an explanation interface as sufficient evidence of compliance or meaningful oversight. Explanation quality must be validated empirically with actual target audiences, and governance processes should include mechanisms to detect and correct cases where explanations are misleading or misunderstood. Therefore, explanations stemming from *interpretability* research and understanding model behaviour can guide so-called *post-hoc* explanations, communicating model results to different user groups. However, the effectiveness and usefulness of explanations remain heavily use case and target group specific, therefore robust validation is necessary, when integrating XAI features into AI systems.

Particular notions of AI model explanations appear throughout the AI lifecycle.[77] Whereas concerns regarding data quality and debugging dominate the early stages, in later stages, explanations geared toward business actionability become more important. The uses for XAI and recommendations for incorporating XAI throughout the AI lifecycle can be detailed along four stages: (1) data, (2) models and model training, (3) deployment, and (4) monitoring.

[72] Aysel, H. I., Cai, X., & Prugel-Bennett, A. (2025). Explainable artificial intelligence: advancements and limitations. *Applied Sciences*, *15*(13), 7261.
[73] Chromik, M., Eiband, M., Buchner, F., Krüger, A., and Butz, A. (2021). I Think I Get Your Point, AI! The Illusion of Explanatory Depth in Explainable AI. In Proceedings of the 26th International Conference on Intelligent User Interfaces (IUI '21), pp. 307–317.
[74] Schmitt, M., et al. (2024). The Role of Explainability in Collaborative Human-AI Disinformation Detection. In Proceedings of the 2024 ACM Conference on Fairness, Accountability, and Transparency (FAccT '24). Association for Computing Machinery, New York, NY, USA, 2157–2174. https://doi.org/10.1145/3630106.3659031.
[75] Schmitt, Vera, Kruse, N., Sahitaj, P., and Schöning, J. (2026). Transparency as Architecture: Structural Compliance Gaps in EU AI Act Article 50 II. *arXiv preprint arXiv:2603.26983* (.
[76] Vasconcelos, H., Jörke, M., Grunde-McLaughlin, M., Gerstenberg, T., Bernstein, M.S., and Krishna, R. (2023). Explanations Can Reduce Overreliance on AI Systems During Decision-Making. Proceedings of the ACM on Human-Computer Interaction, 7(CSCW1), Article 129. https://doi.org/10.1145/3579605.
[77] Dhanorkar, S., Wolf, C. T., Qian, K., Xu, A., Popa, L., & Li, Y. (2021, June). Who needs to know what, when?: Broadening the Explainable AI (XAI) Design Space by Looking at Explanations Across the AI Lifecycle. In Proceedings of the 2021 ACM Designing Interactive Systems Conference (pp. 1591-1602).

With regard to the data a model is trained on, the most important interaction is between data scientists and domain experts. Domain experts can play an essential role in data selection, data understanding, and data quality evaluation (e. g., identification of data biases, incorrect labels, and the required amount and variety of data).[78] Even at such an early stage, XAI methods may be deployed: For instance, methods to identify influential data points can be used to address challenges related to data quality and collection (e.g., Shapley values). From an ethical perspective, privacy requirements related to the data used need to be considered at this stage of the lifecycle.[79]

For models and model training, developers, once again, depend on collaboration with domain experts. Effective modelling depends on domain knowledge for tasks such as feature selection, understanding a model's limitations, and identifying spurious correlations.[80] A spurious correlation exists between two variables when their relationship is not causal, but coincidental or due to confounding with a third variable.[81] For instance, a model designed to prioritize care for pneumonia patients may learn that asthma is a negative predictor for readmission.[82] Evidently, this is not due to asthma reducing the risk of pneumonia. Instead, it reflects biased training data: The correlation occurred, because patients with asthma received greater care in the first place. Similarly, language models may encode and reproduce biases which are part of their training data, for instance, referring to *women doctors* as though *doctor* itself entails *not-woman.*[83] XAI can assist in identifying these spurious correlations and mitigating the biases and performance losses with which they are associated with.[84] XAI can, further, help to foster a shared understanding of a model between developers and domain experts. In this regard, XAI can support the understanding of risks and potential failures associated with AI, as well as in identifying weak points.[85] For this to be the case, however, explanations need to come in semantics familiar to domain experts. Further, the fidelity and reliability of generated explanations need to be ensured, especially for debugging and evaluation purposes.

In the deployment phase, the consideration of end users' concerns becomes crucial. When XAI is employed with the end user in mind, the types of explanations which are required differ from those which, for instance, aid developers in debugging a model. Designing XAI interfaces in a suitable and accessible way ensures that AI systems no

[78] Decker, T., Gross, R., Koebler, A., Lebacher, M., Schnitzer, R., & Weber, S. H. (2023, July). The thousand faces of explainable AI along the machine learning life cycle: industrial reality and current state of research. In *International Conference on Human-Computer Interaction* (pp. 184-208). Cham: Springer Nature Switzerland.
[79] Dhanorkar, S., Wolf, C. T., Qian, K., Xu, A., Popa, L., & Li, Y. (2021, June). Who needs to know what, when?: Broadening the Explainable AI (XAI) Design Space by Looking at Explanations Across the AI Lifecycle. In *Proceedings of the 2021 ACM Designing Interactive Systems Conference* (pp. 1591-1602).
[80] Decker, T., Gross, R., Koebler, A., Lebacher, M., Schnitzer, R., & Weber, S. H. (2023, July). The thousand faces of explainable AI along the machine learning life cycle: industrial reality and current state of research. In *International Conference on Human-Computer Interaction* (pp. 184-208). Cham: Springer Nature Switzerland.
[81] Ye, Wenqian, et al. »The clever Hans mirage: A comprehensive survey on spurious correlations in machine learning.« *arXiv preprint arXiv:2402.12715* (2024).
[82] Caruana, Rich, et al. »Intelligible models for healthcare: Predicting pneumonia risk and hospital 30-day readmission.« *Proceedings of the 21th ACM SIGKDD international conference on knowledge discovery and data mining*. 2015.
[83] Bender, Emily M., et al. »On the dangers of stochastic parrots: Can language models be too big?« *Proceedings of the 2021 ACM conference on fairness, accountability, and transparency*. 2021.
[84] Ross, Andrew Slavin, Michael C. Hughes, and Finale Doshi-Velez. »Right for the right reasons: Training differentiable models by constraining their explanations.« *arXiv preprint arXiv:1703.03717* (2017).
[85] Decker, T., Gross, R., Koebler, A., Lebacher, M., Schnitzer, R., & Weber, S. H. (2023, July). The thousand faces of explainable AI along the machine learning life cycle: industrial reality and current state of research. In *International Conference on Human-Computer Interaction* (pp. 184-208). Cham: Springer Nature Switzerland.

longer appear as opaque black boxes to end users.[86] In this regard, visualisation can play a key role, enabling an effective communication of model behaviour and supporting a user's understanding of a model and its potential limitations. In a business context, explanations should be geared toward actionability: The insights they provide should be tied to the business workflows within which they are integrated.[87]

Finally, monitoring and maintenance requires a variety of roles (e. g., application and automation engineers, service technicians, operators, AI engineers and data scientists).[88] Once again, XAI can support this phase in a variety of ways, including the communication with non-domain experts and incident analysis. For the latter task, the analysis of feature importance provided by some XAI approaches can play a crucial role.

Technical

It is important to ensure that the choice of XAI methods is appropriate for the corresponding use case.[89] In general, the required level of explainability should also depend on the risk level and the impact of the system. There is a plethora of XAI methods and tools available in this regard.

As concerns the problem for which XAI should be used, both the task to be explained and the solution used for the task must be clear.[90] A specific XAI method usually only applies to specific task types (e. g., unsupervised clustering, regression, classification, segmentation) and to specific input data types (e. g., symbolic data such as numerical data or natural language, or non-symbolic data such as images or audio). Models also differ in terms of their level of interpretability.[91] At one end of the spectrum, are intrinsically or inherently interpretable models (e. g., decision trees or linear and logistic regression). At the other end of the spectrum, are post-hoc methods which use a helper model to derive an explanation. This helper model tries to imitate the behaviour of the model to be explained. Where the use case allows, simpler, intrinsically interpretable models should be preferred, as a model which can be fully inspected is more defensible than a more complex one. Where more complex models are necessary, SHAP can serve as a starting point, as it can be applied to any ML model.[92] SHAP scores indicate how each feature contributes to a model's output, but they should be validated with domain experts.

[86] Decker, T., Gross, R., Koebler, A., Lebacher, M., Schnitzer, R., & Weber, S. H. (2023, July). The thousand faces of explainable AI along the machine learning life cycle: industrial reality and current state of research. In *International Conference on Human-Computer Interaction* (pp. 184-208). Cham: Springer Nature Switzerland.
[87] Dhanorkar, S., Wolf, C. T., Qian, K., Xu, A., Popa, L., & Li, Y. (2021, June). Who needs to know what, when?: Broadening the Explainable AI (XAI) Design Space by Looking at Explanations Across the AI Lifecycle. In *Proceedings of the 2021 ACM Designing Interactive Systems Conference* (pp. 1591-1602).
[88] Decker, T., Gross, R., Koebler, A., Lebacher, M., Schnitzer, R., & Weber, S. H. (2023, July). The thousand faces of explainable AI along the machine learning life cycle: industrial reality and current state of research. In International Conference on Human-Computer Interaction (pp. 184-208). Cham: Springer Nature Switzerland.
[89] Schwalbe, Gesina, and Bettina Finzel. »A comprehensive taxonomy for explainable artificial intelligence: a systematic survey of surveys on methods and concepts.« *arXiv preprint arXiv:2105.07190* (2021).
[90] Schwalbe, Gesina, and Bettina Finzel. »A comprehensive taxonomy for explainable artificial intelligence: a systematic survey of surveys on methods and concepts.« *arXiv preprint arXiv:2105.07190* (2021).
[91] Schwalbe, Gesina, and Bettina Finzel. »A comprehensive taxonomy for explainable artificial intelligence: a systematic survey of surveys on methods and concepts.« *arXiv preprint arXiv:2105.07190* (2021).
[92] Lundberg, Scott M., and Su-In Lee. »A unified approach to interpreting model predictions.« *Advances in neural information processing systems* 30 (2017).

The explanation system component can be thought of as a function which outputs an explanation.[93] For post-hoc explanations, a central question is the extent to which the applied methods are portable to other types of models to be explained (i.e. to other input types): Model-agnostic (or black-box) methods, such as SHAP or LIME, only require access to the inputs and outputs of the model to be explained, whereas model-specific (or white-box) models, such as Layer-wise Relevance Propagation[94], require access to the internal workings of the model to be explained.[95] Another core distinction relates to whether the input to the explanation system component represents local or global behaviour: In the case of local behaviour, the provided explanation output is only valid for a given input example, describing why a particular decision was made.[96] Examples are counterfactual explanations or instance-level feature attributions. In the case of global behaviour, the provided explanation is valid in the complete input space, describing how a decision was made. This may, for example, involve decision-rule extraction. Global explanations can help to gain a high-level understanding of a system.[97] In general, the type of output may convey different types of information, ranging from feature importance to example instances to rule-based types such as decision trees.[98]

When deploying these methods, some of their limitations must be borne in mind. Otherwise, there is a risk of over-trusting explanations.[99] For instance, SHAP and LIME are model-dependent: For the same task and the same dataset, the importance assigned to features will differ depending on the ML model used for the task.[100] Further, since SHAP assumes features to be independent and LIME converts any model into a local linear model, collinearity may lead to important features being assigned a low score or weight.[101] A lack of robustness can lead to a similar misidentification of feature importance. Ideally, methods providing local explanations should be robust to input perturbations, producing similar explanations for similar inputs. However, many methods fail to meet the test.[102] Crucially, the aforementioned benefit of identifying biases may also not be realized. Both feature attribution methods and counterfactual explanations have been shown to be susceptible to adversarial attacks, leading to a

[93] Schwalbe, Gesina, and Bettina Finzel. »A comprehensive taxonomy for explainable artificial intelligence: a systematic survey of surveys on methods and concepts.« *arXiv preprint arXiv:2105.07190* (2021).
[94] Montavon, Grégoire, et al. »Layer-wise relevance propagation: an overview.« *Explainable AI: interpreting, explaining and visualizing deep learning* (2019): 193-209.
[95] Schwalbe, Gesina, and Bettina Finzel. »A comprehensive taxonomy for explainable artificial intelligence: a systematic survey of surveys on methods and concepts.« *arXiv preprint arXiv:2105.07190* (2021).
[96] Schwalbe, Gesina, and Bettina Finzel. »A comprehensive taxonomy for explainable artificial intelligence: a systematic survey of surveys on methods and concepts.« *arXiv preprint arXiv:2105.07190* (2021).
[97] Dhanorkar, S., Wolf, C. T., Qian, K., Xu, A., Popa, L., & Li, Y. (2021, June). Who needs to know what, when?: Broadening the Explainable AI (XAI) Design Space by Looking at Explanations Across the AI Lifecycle. In *Proceedings of the 2021 ACM Designing Interactive Systems Conference* (pp. 1591-1602).
[98] Schwalbe, Gesina, and Bettina Finzel. »A comprehensive taxonomy for explainable artificial intelligence: a systematic survey of surveys on methods and concepts.« *arXiv preprint arXiv:2105.07190* (2021).
[99] Kaur, Harmanpreet, et al. »Interpreting interpretability: understanding data scientists' use of interpretability tools for machine learning.« *Proceedings of the 2020 CHI conference on human factors in computing systems*. 2020.
[100] Salih, Ahmed M., et al. »A perspective on explainable artificial intelligence methods: SHAP and LIME.« *Advanced Intelligent Systems* 7.1 (2025): 2400304.
[101] Salih, Ahmed M., et al. »A perspective on explainable artificial intelligence methods: SHAP and LIME.« *Advanced Intelligent Systems* 7.1 (2025): 2400304.
[102] Alvarez-Melis, David, and Tommi S. Jaakkola. »On the robustness of interpretability methods.« *arXiv preprint arXiv:1806.08049* (2018).

biased model appearing as unbiased in an explanation.[103] Formal methods may alleviate some of these issues, but they may not be scalable.[104]

In terms of evaluating an explanation system component, different kinds of evaluations and associated metrics may be distinguished based on their level of human involvement. Application-grounded evaluation involves human experiments within the target application context.[105] It seeks to assess the quality of an explanation in the context of its end-task. Example metrics measure end-user satisfaction or the improvement of human judgment.[106] Human-grounded evaluation involves simpler human-subject experiments, potentially with lay persons. Their goal is to assess more general notions of the quality of an explanation.[107] Example metrics, such as interpretability or comprehensibility, seek to measure the quality of the mental model of the system to be explained, which is developed by the receiver of the explanation.[108] Finally, functionally-grounded evaluation requires no human experiments.[109] They measure formal properties of the explanation system component, such as its fidelity, robustness, expressiveness, or consistency.[110]

In terms of documentation, it is of great importance to trace all relevant information on AI-systems, their actions, and the outcome of applied XAI methods to have a reliable information base for transparency and evaluation, such as debugging or error analysis. In this regard, Data Cards can provide a human-centered documentation of datasets,[111] and the use of model cards has become an established practice.[112]

Organisational

There are various approaches to integrate explainability into working businesses on an organisational level.

Regarding roles and responsibilities when deploying XAI methods, it is important to consider the target audience, as different external stakeholders need different levels of explanation. Based on the AI knowledge external stakeholders possess and the demanded informational outcome, the level of appropriate technical detail of the explanation should be adapted. Additionally, the timing of an explanation in the AI lifecycle of a stakeholder should be considered as well, when designing explanations.

[103] Slack, Dylan, et al. »Feature attributions and counterfactual explanations can be manipulated.« *arXiv preprint arXiv:2106.12563* (2021).
[104] Marques-Silva, Joao, and Alexey Ignatiev. »Delivering trustworthy AI through formal XAI.« Proceedings of the AAAI Conference on Artificial Intelligence. Vol. 36. No. 11. 2022.
[105] Doshi-Velez, Finale, and Been Kim. »Towards a rigorous science of interpretable machine learning.« *arXiv preprint arXiv:1702.08608* (2017).
[106] Schwalbe, Gesina, and Bettina Finzel. »A comprehensive taxonomy for explainable artificial intelligence: a systematic survey of surveys on methods and concepts.« *arXiv preprint arXiv:2105.07190* (2021).
[107] Doshi-Velez, Finale, and Been Kim. »Towards a rigorous science of interpretable machine learning.« *arXiv preprint arXiv:1702.08608* (2017).
[108] Schwalbe, Gesina, and Bettina Finzel. »A comprehensive taxonomy for explainable artificial intelligence: a systematic survey of surveys on methods and concepts.« *arXiv preprint arXiv:2105.07190* (2021).
[109] Doshi-Velez, Finale, and Been Kim. »Towards a rigorous science of interpretable machine learning.« *arXiv preprint arXiv:1702.08608* (2017).
[110] Schwalbe, Gesina, and Bettina Finzel. »A comprehensive taxonomy for explainable artificial intelligence: a systematic survey of surveys on methods and concepts.« *arXiv preprint arXiv:2105.07190* (2021).
[111] Pushkarna, Mahima, Andrew Zaldivar, and Oddur Kjartansson. »Data cards: Purposeful and transparent dataset documentation for responsible ai.« *Proceedings of the 2022 ACM conference on fairness, accountability, and transparency*. 2022.
[112] Liang, Weixin, et al. »Systematic analysis of 32,111 AI model cards characterizes documentation practice in AI.« *Nature Machine Intelligence* 6.7 (2024): 744-753.

End users, for instance, seek easily understandable outcome-oriented explanations, while development experts may require deeper technical information. Therefore, it is essential to analyse stakeholders and their needs to support different perspectives on explainability.[113]

In terms of governance structures, managers and similar stakeholders should ensure that the deployment of AI and its advancements benefit both employees and the organisation as a whole, through not only deploying efficient and profitable models, but also ensuring safety and explainability.[114] This can be achieved through defining clear goals and policies for the optimal deployment and design of explanations. Furthermore, compliance with legal requirements, such as the GDPR, should be ensured. Additionally, it is of great importance to trace all relevant information on AI systems, their actions, and the outcome of applied XAI methods to have a reliable information base for transparency and evaluation like debugging or error analysis. When deploying agent-based AI technologies, it is crucial to regulate and restrict an agent's access to available information and its autonomous actions. The design of such networks should always include a human-in-the-loop design, to always secure a complete overview and possible interventions.

To integrate XAI in ongoing processes like developing and deploying AI-based systems, it is essential to establish human-centred development frameworks to continuously include stakeholder concerns regarding mistrust and explainability of an AI-based system. This includes understanding the target stakeholders and their needs, defining clear goals, proposing concrete actions, and validating them empirically.[115] Furthermore, collaborative decision-making processes and transparent trust dynamics should be included in development and evaluation processes, as they are essential for establishing trust in AI interactions through XAI and fostering effective collaboration between humans and AI.[116]

[113] Dhanorkar, S., Wolf, C. T., Qian, K., Xu, A., Popa, L., & Li, Y. (2021, June). Who needs to know what, when?: Broadening the Explainable AI (XAI) Design Space by Looking at Explanations Across the AI Lifecycle. In *Proceedings of the 2021 ACM Designing Interactive Systems Conference* (pp. 1591-1602).
[114] Molina, D. A., Kharlov, V., & Chen, J. S. (2024, August). Towards effective human-AI collaboration in decision-making: A comprehensive review and conceptual framework. In 2024 Portland international conference on management of engineering and technology (PICMET) (pp. 1-6). IEEE.
[115] Zhukova, A., von Sperl, L., Matt, C. E., & Gipp, B. (2024). Generative user-experience research for developing domain-specific natural language processing applications. Knowledge and Information Systems, 66(12), 7859-7889.
[116] Molina, D. A., Kharlov, V., & Chen, J. S. (2024, August). Towards effective human-AI collaboration in decision-making: A comprehensive review and conceptual framework. In 2024 Portland international conference on management of engineering and technology (PICMET) (pp. 1-6). IEEE.

6 Conclusion and Strategic Outlook

Trust & Its Limitations: Who Do We Trust?

In the context of modern AI systems, Trustworthy AI Systems cannot be assigned to a single entity. Their creation is inherently distributed across multiple layers, including providers, systems, organisations, and human actors, and each of these layers introduces its own uncertainties and limitations.

Providers of AI models and infrastructure often operate as partial black boxes to each other, especially in API-based deployments, requiring organisations to rely on external security practices, update cycles, and data handling policies without full visibility. At the same time, AI systems themselves are probabilistic and adaptive, meaning their behaviour cannot be fully predicted or validated once and for all. With the advent of more autonomous AI agents, these interdependencies become even more complex.

Even within the same organisation, Trustworthy AI Systems cannot be guaranteed. Human interaction with AI introduces additional risks, such as overreliance, misinterpretation, or misuse of outputs. Therefore, a reasonable level of mistrust in AI systems across all levels of the value chain increases the likelihood of effectively establishing Trustworthy AI Systems.

As a result, trust in AI must be reframed. It is neither a property of a model nor a feature that can be certified once at deployment. Instead, it emerges from the continuous management of risks across all involved actors and system components.

Trust is not a static attribute; it is a continuously managed outcome.

Limitations of Existing Regulations

Regulatory frameworks such as the EU AI Act play a critical role in establishing minimum standards for safety, transparency, and accountability. However, they are not sufficient to create trust on their own.

A key limitation lies in the mismatch between static regulation and dynamic systems. While regulations define structured requirements and compliance checkpoints, AI systems evolve continuously through data updates, retraining, integration changes, and shifting usage contexts.

Furthermore, compliance does not equate to trustworthy behaviour in practice. Meeting formal requirements, such as documentation, classification, or transparency obligations, does not guarantee that a system behaves correctly, fairly, or safely under real-world conditions.

In addition, regulations primarily define *what* must be achieved but provide limited guidance on *how* organisations should operationalize these requirements within complex, interconnected environments. This is further complicated by the need to align multiple regulatory regimes, creating fragmentation and increasing implementation complexity.

Regulation therefore establishes necessary guardrails, but trust must be built beyond compliance.

From Trustworthy AI to AI Governance

As AI systems become deeply embedded in business processes and societal infrastructure, the focus is shifting from evaluating isolated system properties toward managing AI as a continuous, system-level responsibility.

The concept of »Trustworthy AI« captures desired system characteristics, but it assumes a degree of stability that does not reflect real-world conditions. In practice, AI systems evolve, interact, and operate within dynamic environments. This requires a broader perspective.

AI Governance addresses this shift by extending the focus from individual systems to the full lifecycle and ecosystem in which AI operates.

In practical terms, this means moving:

- from point-in-time validation to continuous monitoring and control
- from isolated technical safeguards to coordinated organisational processes
- from siloed responsibilities to cross-functional accountability

The goal of AI Governance is to manage, avoid, and mitigate risks across all lifecycle phases, ensuring that AI systems remain aligned with organisational objectives, regulatory requirements, and societal expectations over time.

Internal Guidelines That Build Trust

Trust is not created through high-level principles alone. It is built through policies that are operational, enforceable, and embedded into daily practice.

Effective guidelines share a common characteristic: they **translate abstract requirements into concrete actions across the lifecycle**. This includes:

- integrating controls across data, model development, deployment, and operations
- establishing clear accountability through defined roles and responsibilities
- linking governance to measurable controls, such as monitoring, testing, and escalation mechanisms

Equally important is the ability to adapt. Given the evolving nature of AI systems and their risk landscape, policies must not remain static. They require regular review,

continuous feedback loops, and flexibility to respond to new risks, technologies, and use cases.

In practice, trust emerges where organisations succeed in aligning:

- technical robustness with organisational discipline, and
- governance structures with real operational behaviour

Strategic Outlook

AI Governance is evolving into a core organisational capability. Similar to cybersecurity or financial governance, it will become a permanent management function rather than a project-based activity.

Organisations that succeed in this environment will not rely on the assumption that AI systems are inherently trustworthy. Instead, they will treat AI as dynamic, risk-bearing infrastructure that requires continuous oversight.

This implies a shift from compliance-driven thinking toward risk-driven and lifecycle-oriented management. Trust, in this context, is no longer a prerequisite for using AI, it is the result of effective governance.

Annex

Glossary

For further background, Bitkom publications ↗»KI & Informationssicherheit: Ein Überblick zu Informationssicherheit von und durch KI« and ↗»Security of AI Agents« serve as overarching reference documents.

Agentic AI
AI systems designed to operate with high autonomy, able to interpret goals, generate plans, and execute actions with limited human involvement. Agentic AI emphasizes initiative, adaptiveness, and context-aware decision-making, forming the conceptual foundation for AI Agents.

AI Agent
An autonomous, goal-driven software entity that uses AI reasoning to perform tasks, make decisions, and interact with digital or physical environments. AI Agents represent the operational embodiment of agentic AI capabilities.

Trustworthy AI
AI systems that uphold fundamental rights, transparency, human oversight, technical robustness, safety, accountability, and fairness as required under the EU AI Act. Trustworthy AI ensures that systems are reliable, predictable, and aligned with societal values.

Responsible AI
Practices, processes, and governance mechanisms ensuring that AI technologies are developed and deployed in ways that are ethical, compliant, and socially accountable. Responsible AI emphasizes risk management, explainability, data governance, inclusivity, and continuous monitoring across the AI lifecycle.

Adversarial Attack
The use of subtly manipulated inputs to intentionally cause an AI model to produce inaccurate or harmful outputs, often without being detectable by humans.

Data Poisoning
The deliberate insertion of corrupted or misleading data into training datasets to compromise a model's integrity, reliability, or safety.

Explainability
The property of an AI system that enables humans to understand, interpret, and verify how and why it produces specific decisions or outputs.

Fairness
Ensures that systems treat individuals and groups equitably and do not produce unjustified bias or discrimination. It requires identifying and mitigating biases in data, design, and deployment to avoid systematically skewed outcomes.

Generative AI (GenAI)
AI models capable of producing original content, such as text, images, audio, or code, based on learned patterns.

Hallucination
The generation of incorrect, fabricated, or unsupported outputs by an AI system that appear coherent or plausible.

Least-Privilege Principle
A governance and security guideline ensuring that AI systems, agents, or users receive only the minimum access rights necessary to perform their tasks, reducing misuse and unintended impact.

Prompt Injection
A manipulation technique in which crafted prompts are used to override or bypass an AI model's intended constraints, influencing behaviour beyond its authorized scope.

Retrieval-Augmented Generation (RAG)
An approach that enhances generative AI by retrieving factual information from external knowledge sources before generating outputs, improving reliability and trustworthiness.

Bitkom represents more than 2,300 member companies from the digital economy. In Germany, they generate over 200 billion euros in turnover through digital technologies and solutions and employ more than 2 million people. Its members include more than 1,000 SMEs, over 700 start-ups and virtually all global players. They offer software, IT services, telecommunications or internet services, manufacture devices and components, operate in the digital media sector, create content, provide platforms or are otherwise part of the digital economy. 82 per cent of the companies involved in Bitkom have their headquarters in Germany, a further 8 per cent come from the rest of Europe and 7 per cent from the USA. 3 per cent come from other regions of the world. Bitkom promotes and drives the digital transformation of the German economy and advocates for broad social participation in digital developments. The aim is to make Germany a high-performing and sovereign digital hub.

Publisher

Bitkom e.V.

Albrechtstr. 10 | 10117 Berlin

Contact

Lucy Czachowski | Head of AI & Cloud – Resilience & Infrastructure

T +49 30 27576-320 | l.czachowski@bitkom.org

Responsible Bitkom Working Group

WG Artificial Intelligence

Authors

Andrea Martin (IBM), **Benjamin Herd** (Fraunhofer IKS), **Catharina Kreiling** (BCG), **Danilo Brajovic** (IPA Fraunhofer), **Jens Beier** (divis), **Kirsten Rulf** (BCG), **Lena Münstermann** (KPMG), **Lucy Czachowski** (Bitkom), **Manoj Kahdan** (RWTH Aachen), **Mario Köpke** (HPE), **Maximilian Eder** (Bauhaus-Universität Weimar), **Michael Hoche** (Airbus Defence & Space), **Michael Krah** (OFFIS – Institut für Informatik), **Paul Zenker** (KPMG), **Philippe Krajsic** (Cyberagentur), **Rebekka Görge** (IAIS Fraunhofer), **Reinhard Stolle** (IKS Fraunhofer), **Sönke Erdmann** (University of Potsdam), **Sofia Trojanowska** (Infosys), **Syrko Kulas** (Cyberagentur), **Sven Trendow** (AI Grid), **Teresa Kutzner** (Hochschule der Medien), **Usani Lingamoorthy** (KPMG), **Dr. Michael A. Hedderich**, **Dr. Vera Schmitt** (TU Berlin), **Vivek Chavan** (TU Berlin & Fraunhofer IPK)