ASSESSMENT OF LEARNING TECHNOLOGIES AND FRAMEWORKS FOR INTELLIGENT AND ETHICAL AI

D2.5 Initial best practices blueprint for ethical AI and AutoML Integration

Deliverable No:	D2.5
Deliverable Name:	Initial best practices blueprint for ethical AI and AutoML Integration
Work Package:	WP2
Task:	T2.5, T2.6
Dissemination Level:	PU
Deliverable Type:	R
Lead Organization:	DBC
Submission month:	M16
Date:	30 January 2026

Funded by the European Union

Funded by the European Union under Grant Agreement 101177912. Views and opinions expressed are however those of the author(s) only and do not necessarily reflect those of the European Union or European Research Executive Agency (REA). Neither the European Union nor the granting authority can be held responsible for them.

Project description

Acronym	**ALFIE**
Title	**Assessment of Learning technologies and Frameworks for Intelligent and Ethical AI**
Coordinator	CATALINK
Reference	101177912
Type	Research and Innovation Actions (RIA)
Programme	**Horizon Europe (HORIZON)**
Topic	HORIZON-CL2-2024-TRANSFORMATIONS-01-06 Beyond the horizon: A human-friendly deployment of artificial intelligence and related technologies
Start	01.10.2024
Duration	36 months
Website	alfie-project.eu
Consortium	**Catalink Limited** (CTL)**,** Cyprus, Coordinator **University of Brighton** (UoB), United Kingdom **Centre for Research & Technology** (CERTH)**,** Greece **Distributed Analytics Solutions LTD** (DAS), United Kingdom **Edge Hill University** (EHU), United Kingdom **Kempelen Institute of Intelligent Technologies** (KINIT), Slovakia **Universitat Autònoma de Barcelona** (UAB), Spain **Robert Bosch España SLU** (Bosch), Spain **Eindhoven University of Technology (**TU/e), Netherlands **Diadikasia Business Consulting** (DBC), Greece

Document history

Version	Date	Beneficiary	Description
0.1	17.10.2025	DBC	Initial ToC created
0.2	08.12.2025	CTL	Analysed insights from D2.1
0.3	09.01.2026	CERTH	Input added
0.4	12.01.2026	KINIT	Input added
1.0	16.01.2026	DBC	First draft ready
1.1	25.01.2026	BOSCH	Major revision
1.2	25.01.2026	BOSCH, KINIT	Peer review
1.3	27.01.2026	DBC	Peer review comments addressed; document finalised
1.4	28.01.2026	CTL	Final version

Authors

Partner	Name(s)
DBC	Stella Theologidou, Daphne Giakoumaki
KINIT	Matus Mesarcik
CERTH	Stamatis Samaras
CTL	Zenon Lamprou, Maria E. Skarkala
UAB	Sarah Anne McDonagh, Pilar Orero, Estela Oncins
BOSCH	Juan A. Recio

Contributors

Partner	Contribution type	Name
BOSCH	Review	Juan A. Recio
KINIT	Review	Adrian Gavornik

Glossary

Acronym	Definition
AI	Artificial intelligence
AI Act	Artificial Intelligence Act
ALTAI	Assessment List for Trustworthy AI
ARIA	Accessible Rich Internet Applications
AutoML	Automated Machine Learning
BEATS	Bias Evaluation and Assessment Test Suite
BPS	Business Partner Screening
CAVs	Connected Automated Vehicles
CV	Curriculum Vitae
DSM	Driver State Monitoring
ETD-Hub	EthiTech Dialogue Hub
EU	European Union
GA	Grant Agreement
GDPR	General Data Protection Regulation
GPAI	General-Purpose AI
HLEG	High-Level Expert Group (on AI)
IAM	Identity and Access Management
KG	Knowledge Graphs
LLMs	Large Language Models
NGOs	Non-Governmental Organisations
NLP	Natural Language Processing
PETs	Privacy-Enhancing Technologies
PII	Personally Identifiable Information
PMM	Post-Market Monitoring
POUR	Perceivability, Operability, Understandability and Robustness
PSI	Population Stability Index
R&D	Research and Development

SMEs	Small and Medium-sized Enterprises
STEP	Strategic Technologies for Europe Platform
WA	Web Accessibility
WCAG	Web Content Accessibility Guidelines
XAI	Explainable AI

Executive Summary

This deliverable presents the first version of a structured blueprint that identifies best practices for the responsible development, assessment and deployment of AI systems within the ALFIE project. The document focuses on how ethical principles and European policy requirements can be interpreted and applied in concrete AI application contexts, with particular attention to systems developed or supported through Automated Machine Learning (AutoML).

The blueprint is grounded in a desk-based analysis of the European regulatory and ethical landscape for AI, including current and emerging policy initiatives such as the AI Act and the General Data Protection Regulation (GDPR), as well as relevant academic and policy literature on trustworthy AI. Rather than treating these frameworks as abstract guidance, the deliverable examines how their key requirements relate to practical design and evaluation choices in AI systems. Furthermore, a central contribution of the document is the ethics risk assessment of three representative AI use cases used within ALFIE, covering driver state monitoring, automated assessment of web accessibility and corporate compliance analysis.

Overall, the deliverable contributes to a clearer understanding of how ethical and legal expectations for AI can be translated into actionable guidance grounded in real application scenarios, supporting the broader goal of developing AI systems that are aligned with European values and societal expectations.

Contents

1 Introduction

1.1 Purpose

The growing integration of Artificial Intelligence (AI) systems across public and private sectors has increased the urgency for robust approaches that ensure their ethical, transparent and trustworthy development and deployment. In response, the European Union (EU) has advanced a human-centric vision for AI by establishing an evolving regulatory and policy landscape aimed at aligning technological innovation with fundamental rights, democratic principles and broader societal values.

Within this context, ALFIE project brings together the EthiTech Dialogue Hub (ETD-Hub) and the AutoML platform to support ethical and trustworthy AI development. The ETD-Hub functions as a multidisciplinary discussion forum in which citizens, policymakers, AI professionals and legal experts exchange opinions about the ethical and legal dimensions of AI. In parallel, the AutoML platform enables users to build and evaluate AI models through configurable pipelines and interfaces that incorporate ethical and regulatory considerations. Through the interaction between these two components, ALFIE aims to promote AI systems that are both technically effective and aligned with European values.[1]

This deliverable, *D2.5 - Initial Best Practices Blueprint for Ethical AI and AutoML Integration*, provides a structured overview of key ethical considerations relevant to the design, assessment and deployment of AI systems. It reflects on the regulatory and ethical context, explores the evolving landscape of ethical AI and AutoML and considers how emerging requirements can inform responsible AI development. A core focus of the document is the risk assessment of ALFIE's three use cases, including the identification of key ethical and legal challenges and corresponding mitigation strategies. Based on these insights, the deliverable outlines initial best practices and recommendations to support the integration of ethical and legal considerations in AI systems. As the first iteration of this blueprint, the findings presented here will be further developed and refined in Deliverable 2.6, scheduled for submission at Month 30 of the project.

1.2 Structure of the deliverable

This deliverable is structured across six chapters. Following the present introductory chapter, Chapter 2 examines the broader regulatory and ethical landscape of AI, with an emphasis on European legal instruments, policy guidelines and emerging normative expectations relevant to trustworthy AI.

Chapter 3 operationalises these frameworks by identifying key design considerations that support the ethical development and deployment of AI systems, including principles such as fairness, transparency, accountability and data protection. Chapter 4 presents a risk assessment of ALFIE's three use cases, identifying specific ethical and legal challenges and proposing corresponding mitigation strategies tailored to each context.

[1]European Commission, "Grant Agreement No. 101177912 - ALFIE," European Commission, Brussels, Belgium, 2024.

Building on the preceding analysis, Chapter 5 outlines ALFIE's initial set of best practices for embedding ethical and legal considerations into AI systems' development and deployment, particularly those supported by the project's AutoML platform. Finally, chapter 6 synthesises the main findings of the deliverable and sets the direction for future work.

1.3 Relation to project objectives

This deliverable contributes directly to two core objectives of ALFIE. The first is to improve guidance and best practices for the ethical and responsible development and use of AI across different sectors. In support of this objective, the deliverable analyses the current European regulatory and policy environment, reviews key ethical challenges and proposes recommendations that address both technical and societal dimensions. These insights are intended to support AI developers, policymakers and other stakeholders in understanding how fundamental rights, democratic values and social expectations can be reflected in AI systems in practice.

The second objective to which this deliverable contributes is ALFIE's ambition to support the creation of AI systems that are trustworthy, inclusive and aligned with European legal and ethical principles. By conducting a structured risk assessment of the project's three use cases, the deliverable identifies context-specific ethical and legal risks and proposes mitigation strategies that can inform responsible decision-making during system design and deployment. The outcomes aim to promote greater transparency, fairness and accountability in the development of AI models and workflows, particularly within environments that rely on automated machine learning.

More broadly, the deliverable builds on ALFIE's interdisciplinary research efforts to bridge the gap between theory, policy commitments and real-world implementation. It translates abstract principles into practical guidance that can be applied in technical development processes and made accessible to both experts and non-experts. As the first version of ALFIE's ethical best practices blueprint, it sets the foundation for a more advanced framework to be presented in a follow-up report during the later stages of the project.

2 Regulatory and ethical landscape of AI

2.1 The need for ethical AI

The accelerated and widespread integration of Artificial Intelligence (AI) into critical sectors, including healthcare, financial services, autonomous vehicles and public governance, has necessitated a profound shift toward prioritising Ethical AI[2]. Ethical AI is defined as a systematic normative reflection, based on a holistic, comprehensive, multicultural and evolving framework of interdependent values and principles. This framework's ultimate purpose is to guide societies in dealing responsibly with the impacts of AI technologies.[3]

The urgency for establishing robust ethical guidelines stems directly from technology's profound societal impact and the inherent, often subtle, risks it introduces. Unlike traditional software errors, mistakes or biases embedded in AI systems can scale rapidly and affect millions of people, leading to significant real-world harm, the erosion of trust and the perpetuation of systemic social injustices. But without proactive ethical governance, the transformative potential of AI risks being overshadowed by its potential for misuse, discrimination, and the displacement of human control and moral judgement. Therefore, the concept of Ethical AI is not merely a philosophical concern but a pragmatic imperative for ensuring the sustainable and legitimate future of technology.

2.2 Challenges across AI lifecycle and data governance

The ethical complexities of AI are not confined to the final application but are embedded across the entire AI lifecycle, from initial data collection and model training to deployment and maintenance. These pervasive challenges primarily revolve around three interconnected failures of ethical governance: data integrity, system transparency and accountability[4].

2.2.1 Data Integrity: The Foundation of Algorithmic Bias

The challenge of Algorithmic Bias arises as a direct failure of data integrity at the earliest stages of the lifecycle, Data Collection and Model Training. Bias occurs when AI models are trained on historical or unrepresentative datasets that reflect and encode existing societal prejudices, such as racial, gender or socioeconomic biases. This structural flaw is then codified into the algorithm, leading to unfair or inequitable outcomes when the models are deployed. Correcting this requires continuous auditing of training data and the development of technical fairness metrics, acknowledging that the system's output can never exceed the integrity of its input.

---

[2]L. Floridi, J. Cowls, M. Beltrametti, *et al.*, "AI4People – An ethical framework for a good AI society: Opportunities, risks, principles, and recommendations," *Minds Mach.*, vol. 28, no. 4, pp. 689–707, 2018, doi: 10.1007/s11023-018-9482-5.

[3]UNESCO, "Recommendation on the ethics of artificial intelligence," General Conference, 41st session, SHS/BIO/REC-AIETHICS/2021, Nov. 2021. [Online]. Available: https://unesdoc.unesco.org/ark:/48223/pf0000380455.

[4]B. D. Mittelstadt, P. Allo, M. Taddeo, S. Wachter, and L. Floridi, "The ethics of algorithms: Mapping the debate," *Big Data Soc.*, vol. 3, no. 2, 2016, doi: 10.1177/2053951716679679.

2.2.2 System Transparency: The Black-Box and Explainability

Many modern, high-performing AI systems, particularly deep neural networks, operate as "black boxes", meaning their internal decision-making processes are too opaque for humans to fully understand or trace. This critical lack of explainability (also known as interpretability) poses a serious ethical hurdle, especially in high-stakes contexts like medical diagnostics or loan approval. The inability to explain why an AI arrived at a decision undermines public trust, prevents the effective identification and correction of errors and makes it virtually impossible to assign legal or moral responsibility when harm occurs. Consequently, the opacity of the model directly undermines the requirement for system transparency throughout the entire development process.

2.2.3 Accountability: Data Governance and Privacy Data Governance and Privacy

AI systems require vast quantities of data, raising fundamental issues concerning data governance, privacy and surveillance. Ethical AI mandates strict adherence to principles of informed consent, data minimisation and robust security protocols to protect sensitive personal information[5]. Regulations like the General Data Protection Regulation (GDPR) have established precedents, but the ethical challenge lies in ensuring that new AI techniques maintain these privacy standards while simultaneously enabling beneficial innovation.

2.3 EU regulatory frameworks, guidelines, and acts

The European Union has positioned itself as a global leader in the governance of AI, moving from a period of voluntary ethical self-regulation to a robust, legally binding regulatory architecture. This transition was prompted by a growing academic consensus that non-binding guidelines often lack the mechanisms to enforce their own normative claims, sometimes serving as a "decorative" alternative to strict legal oversight[6]. Additionally, legal vacuum is not suitable for development and deployment of novel AI solutions and may present an obstacle in innovations.

2.3.1 High-Level Expert Group (HLEG) Guidelines (The Foundation)[7]

The trajectory of EU regulation began with the **Ethics Guidelines for Trustworthy AI**, published in 2019 by the High-Level Expert Group on AI. These guidelines established the foundational premise that for AI to be considered trustworthy, it must be lawful, ethical, and robust throughout its entire lifecycle. To operationalise these values, the HLEG introduced seven key requirements that must be met:

i. **Human agency and oversight:** Ensuring systems support human autonomy and allow for human intervention.

---

[5]UNESCO, "Recommendation on the ethics of artificial intelligence," General Conference, 41st session, SHS/BIO/REC-AIETHICS/2021, Nov. 2021. [Online]. Available: https://unesdoc.unesco.org/ark:/48223/pf0000380455.

[6]T. Hagendorff, "The ethics of AI ethics: An evaluation of guidelines," *Minds Mach.*, vol. 30, no. 1, pp. 99–120, 2020, doi: 10.1007/s11023-020-09517-8.

[7]High-Level Expert Group on Artificial Intelligence, "Ethics guidelines for trustworthy AI," European Commission, 2019. [Online]. Available: https://digital-strategy.ec.europa.eu/en/library/ethics-guidelines-trustworthy-ai.

ii. **Technical robustness and safety:** Developing resilient systems with fallback plans to prevent harm.
iii. **Privacy and data governance:** Prioritising data integrity and security throughout the lifecycle.
iv. **Transparency:** Addressing the "black-box problem" through traceability and clear communication.
v. **Diversity, non-discrimination, and fairness:** Actively mitigating algorithmic bias to ensure equitable outcomes.
vi. **Societal and environmental well-being:** Ensuring AI development is sustainable and socially beneficial.
vii. **Accountability:** Establishing clear mechanisms for responsibility and auditability.

By articulating these specific pillars, the HLEG provided a concrete framework for assessing algorithmic impact. While initially non-binding, these principles served as the conceptual bridge, translating abstract ethical theory into the actionable requirements that define the subsequent EU AI Act.

Such approach is directly confirmed by the EU AI Act. Recital 27 reads: *"While the risk-based approach is the basis for a proportionate and effective set of binding rules, it is important to recall the 2019 Ethics guidelines for trustworthy AI developed by the independent AI HLEG appointed by the Commission. In those guidelines, the AI HLEG developed seven non-binding ethical principles for AI which are intended to help ensure that AI is trustworthy and ethically sound...The application of those principles should be translated, when possible, in the design and use of AI models."* This means that principles articulated by the AI HLEG should be implemented during the design and deployment phase of any AI system, regardless the classification based on risk as discussed below.

2.3.2 The EU AI Act (The Landmark Legislation)[8]

The centrepiece of the EU's strategy is the **Artificial Intelligence Act (AI Act)**, the world's first comprehensive horizontal legal framework for AI. The Act moves beyond broad ethical statements by adopting a strict risk-based approach, categorising AI systems based on the level of peril they pose to society. This regulatory logic ensures that the intensity of legal oversight is directly proportionate to the potential for harm to fundamental rights and safety.

The underlying philosophy of the AI Act can be summarised in two key ideas. First, AI is regulated as a product, not merely as software or abstract technology. Second, regulation is based on risk, not on the name, popularity or perceived sophistication of an AI system.

The EU AI Act forms part of the broader EU product safety framework. It extends established regulatory logic including conducting conformity assessment and implementing preventive risk management to AI systems.

[8] *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, 2024. [Online]. Available: http://data.europa.eu/eli/reg/2024/1689/oj.

The crucial change is the shift from ex post liability to ex ante risk prevention. Fundamental rights, safety, and trust are embedded directly into the design and deployment of high-risk AI systems.

This leads to the core design choice of the AI Act: the risk-based approach. Different regulatory obligations apply depending on the level of risk an AI system poses to health, safety, and fundamental rights.

Most AI systems fall into low or minimal risk categories and remain largely unregulated. Only a limited set of high-risk uses is subject to strict legal requirements.

Under this framework, AI systems are divided into four distinct categories:

[1] **Unacceptable Risk:** The Act explicitly bans AI practices considered a clear threat to people's safety or rights. This includes social scoring by governments and private parties, manipulative “dark patterns” that exploit vulnerable groups, predictive policing and restrictions for certain forms of real-time biometric identification in public spaces for law enforcement purposes.

[2] **High Risk:** Systems used in critical infrastructure, education, employment (e.g., CV-scanning software), and selected private and public services are classified as “High Risk”. These systems are subject to stringent compliance obligations, including mandatory risk management, high-quality training datasets to ensure data integrity, human oversight, transparency, cyber-security and detailed technical documentation to satisfy system transparency.

[3] **Limited Risk:** This category refers to AI systems with specific transparency obligations. For example, systems like chatbots or AI-generated content (including “deepfakes”) must be clearly labelled so that users are aware they are interacting with a machine. This ensures that the principle of accountability is upheld through user awareness.

[4] **Minimal or No Risk:** The vast majority of AI systems currently used in the EU (such as AI-enabled video games or spam filters) fall into this category. These systems are free to operate without additional legal obligations, although providers are encouraged to follow voluntary codes of conduct.

In response to the rise of expansive generative models, the Act also includes tailored requirements for **General-Purpose AI (GPAI)**. Unlike systems designed for specific purposes with predictable risk profiles, GPAI models possess a wide range of possible uses and can be integrated into a vast array of downstream applications. Because failure in a foundation model can have cascading effects, the Act mandates that GPAI providers maintain rigorous technical documentation and comply with EU copyright law. For the most powerful models, those classified as possessing "systemic risk", providers are further required to conduct adversarial testing and detailed risk assessments.

All providers and deployers of AI systems shall comply with the requirement of AI literacy. Article 4 of the EU AI Act introduces a general obligation to promote AI literacy among persons dealing with AI systems, ensuring they have the knowledge and skills necessary to use AI responsibly. Providers and deployers must take appropriate measures to ensure that staff and other relevant persons understand the functioning, limitations, risks, and appropriate use of AI

systems. The obligation is proportionate and context-dependent, taking into account the role of the AI system, the risks it poses, and the professional context of the users. AI literacy is framed not only as technical understanding, but also as awareness of legal, ethical, and fundamental-rights implications. The provision aims to reduce misuse, over-reliance, and harm by embedding human understanding as a core safeguard alongside technical and organisational measures.

The EU AI Act supports innovation through regulatory sandboxes that allow developers and public authorities to test AI systems under real-world conditions with regulatory guidance and reduced compliance uncertainty. It applies a proportionate, risk-based approach that limits strict obligations to high-risk AI, while leaving low-risk and general-purpose uses largely unrestricted. Additional support measures include guidance, standards, and tailored assistance for SMEs and start-ups to lower administrative and compliance burdens.

Enforcement is structured through a multi-level governance system combining national competent authorities with EU-level coordination bodies (primarily EU AI Office) to ensure consistent application across Member States. Clear allocation of responsibilities, powers to investigate, and access to documentation enable authorities to assess compliance effectively. The regulation is backed by harmonised penalties and corrective measures, ensuring credible deterrence while allowing proportionate responses based on the nature and gravity of infringements.

2.3.3 Interconnectivity with Data Governance and the GDPR[9]

The AI Act does not function in isolation, On the contrary, it is deeply interconnected with the **General Data Protection Regulation (GDPR)**, forming a comprehensive dual-layered regulatory framework. While the AI Act primarily focuses on the safety, technical robustness and logic of the algorithmic model, the GDPR remains the primary legal framework governing the processing of personal data used to train, test and validate those models.

This interconnectivity is most evident in the following three areas:

i. **Lawfulness and Data Integrity:** The AI Act mandates that providers of High-Risk systems utilise training datasets that are relevant, representative and sufficiently free of errors to mitigate algorithmic bias. However, the technical integrity of these datasets is secondary to their legal validity in case personal data is processed. The authority to collect and process such information is governed exclusively by Article 6 of the GDPR. Consequently, without a valid legal basis, such as informed consent or legitimate interest, the data integrity required by the AI Act cannot be achieved lawfully. This creates a regulatory dependency where technical quality is inseparable from legal compliance.
ii. **Purpose Limitation and Data Minimisation:** Under the GDPR, personal data must be collected for specified, explicit, and legitimate purposes, preventing any further

[9] *Regulation (EU) 2016/679 of the European Parliament and of the Council of 27 April 2016 on the protection of natural persons with regard to the processing of personal data (General Data Protection Regulation)*, OJ L, 2016. [Online]. Available: http://data.europa.eu/eli/reg/2016/679/oj.

processing that is incompatible with those original aims. The AI Act reinforces this principle by mandating strict data governance protocols across the entire AI lifecycle. By ensuring that the data used to refine an "algorithmic output" does not exceed the scope of the user's original agreement, the Act prevents the "mission creep" typically associated with large-scale AI training, thereby aligning technical development with the principle of data minimisation.

iii. **Automated Decision-Making and Rights of the Data Subject:** Article 22 of the GDPR provides individuals with the right not to be subject to a decision based solely on automated processing which produces legal effects. The AI Act's transparency and explainability requirements are designed to make this right meaningful. By forcing developers to address the "black-box problem", the AI Act provides the technical documentation necessary for a data subject to exercise their GDPR right to an explanation of the logic involved in an automated decision.

Ultimately, this synergy ensures that both data inputs and algorithmic outputs are subject to rigorous legal oversight, guaranteeing that technological innovation does not occur at the expense of fundamental privacy rights.

2.4 EU policy recommendations in ethical AI systems

Current EU policy is governed by the "European Approach to Excellence and Trust"[10], a dual-track strategy designed to stimulate innovation while maintaining the world's most stringent ethical safeguards. As the AI Act enters its mid-implementation phase in early 2026, policy recommendations have shifted focus towards sectoral adoption, the enforcement of transparency for GPAI and the mitigation of systemic risks.

Central to the current policy landscape is the **AI Innovation Package**[11] which seeks to secure Europe's technological sovereignty. This plan moves beyond mere regulation to provide the physical and financial infrastructure required for ethical AI development. A key component is the mobilisation of public and private investment through **InvestEU's** dedicated AI strands[12] and the **Strategic Technologies for Europe Platform (STEP)**, which provides a "Sovereignty Seal" to steer funding towards critical AI projects[13].

To ensure that investment aligns with Union values, current policy recommendations prioritise the establishment of a network of **"AI Factories"** co-located with European High-Performance Computing (HPC) supercomputers, building on the amendment to the EuroHPC Joint

[10]N. T. Nikolinakos, "A European approach to excellence and trust: The 2020 White Paper on Artificial Intelligence," in *EU Policy and Legal Framework for Artificial Intelligence, Robotics and Related Technologies: The AI Act*, vol. 53, Springer, 2023, pp. 101–125, doi: 10.1007/978-3-031-27953-9_5.

[11]European Commission, "AI package: Supporting AI startups and SMEs to develop trustworthy AI," European Commission - Digital Strategy, 2025.

[12]*Regulation (EU) 2021/523 of the European Parliament and of the Council of 24 March 2021 establishing the InvestEU Programme*, OJ L 107/30, 2021. [Online]. Available: http://data.europa.eu/eli/reg/2021/523/oj.

[13]*Regulation (EU) 2024/795 of the European Parliament and of the Council of 29 February 2024 establishing the Strategic Technologies for Europe Platform (STEP)*, OJ L, 2024. [Online]. Available: http://data.europa.eu/eli/reg/2024/795/oj.

Undertaking Regulation (EU) 2021/1173 introduced by Regulation (EU) 2021/1732[14], which formally adds the development and operation of AI Factories as an objective of the EuroHPC framework. These hubs serve as integrated ecosystems providing SMEs and researchers with the necessary computing power and curated **"Common European Data Spaces"** required to train models that are "ethical by design". Furthermore, recommendations focus on **Frontier Model Oversight**, which is operationalised through the **EU AI Office** (established by Commission Decision C/2024/1459)[15]. This oversight is facilitated through large-scale facilities dedicated to the development of "frontier" models, ensuring that the next generation of GPAI is developed under European democratic oversight and adheres to the GPAI Code of Practice facilitated by the AI Office.

While the AI Act provides the "rules of the road", the **Apply AI Strategy** serves as the primary vehicle for adoption through sectoral integration. As outlined in the **Commission's 2024 Innovation Package**, this policy recommends a "sector-first" approach, focusing on the integration of AI into strategic industries where the Union holds a competitive advantage, such as advanced manufacturing, biotechnology, and the energy transition[16]. Within these sectors, the Union recommends an **AI-First Policy** that encourages public and private entities to prioritise AI solutions demonstrating high levels of technical robustness, environmental sustainability and explainability. This transition is further supported by the expansion of EU-Level Regulatory Sandboxes. Pursuant to Article 57 of the AI Act, these sandboxes allow businesses, particularly SMEs, to test innovative AI systems in real-world conditions under the direct supervision of national competent authorities and the EU AI Office prior to a full market launch.

Finally, a cornerstone of current recommendations is the mandate for Human-Centricity and AI Literacy. Article 4 of the AI Act is now supported by the European AI Skills Academy, which provides fellowships and training programmes. Policy recommendations explicitly state that any organisation deploying AI must ensure that staff possess the "critical thinking skills" to interpret algorithmic outputs and exercise meaningful human oversight.

2.4.1 AutoML within the EU Ethical AI policy

ALFIE's AutoML platform aims to serve as a practical bridge between EU policy ideals and real-world technology. While the AI Act sets the legal boundaries, the platform shall provide the actual tools needed to build systems that respect those rules. By connecting the high-level legal and ethical debates from the ETD-Hub directly to a technical engine, the project aims to ensure that "ethical AI" is transformed from a theoretical concept into a functional requirement built into every model.

The platform plans to make AI accessible and understandable to everyone, rather than just technical experts. Its interaction layer shall allow users to create AI tools using natural language and speech, which aims to democratise the process by allowing citizens, lawyers

[14] *Regulation (EU) 2024/1732 of the European Parliament and of the Council of 17 June 2024 amending Regulation (EU) 2021/1173*, OJ L, 2024. [Online]. Available: http://data.europa.eu/eli/reg/2024/1732/oj.

[15] *Commission Decision of 24.1.2024 establishing the European AI Office*, OJ C, 2024. [Online]. Available: http://data.europa.eu/eli/dec/2024/1459/oj.

[16] European Commission, "Communication from the Commission on boosting startups and innovation in trustworthy artificial intelligence," COM/2024/28 final, 2024. [Online]. Available: https://eur-lex.europa.eu/legal-content/EN/TXT/?uri=CELEX:52024DC0028.

and small business owners to participate in shaping technology. By supporting multiple languages, the platform will also ensure that these tools feel local and relevant to people across all Member States, fostering the digital literacy and human oversight that European policy requires.

Ultimately, the ALFIE ecosystem will create a continuous conversation between policy and practice. The feedback loop between the ETD-Hub and the AutoML platform shall allow for real-time adjustments; as new ethical challenges emerge, they will be discussed in the hub and then tested on the platform. This aims to create a "living" system that evolves alongside society and technology, ensuring that European AI will remain competitive while staying true to its human-centric soul.

2.5 Insights from D2.1

The aim of D2.1, "Initial modern AI deployments and ethical gaps" (submitted in M12), was to provide a comprehensive report on contemporary AI technologies and ethical disparities. These insights are intended for subsequent integration into the ETD-Hub. This subsection offers a concise summary of the deliverable's primary insights.

As already noted above, modern AI deployments reveal a transformative technological landscape that is reshaping key sectors, including medicine, education, finance, logistics, and more. This technological shift is far outpacing the establishment of ethical standards and legal frameworks, creating a significant ethical gap. Fostering responsible AI requires a shared governance model that integrates transparency, accountability, and inclusivity throughout all stages of design, deployment, and oversight. Economic pressures, though, are prioritising rapid adoption over these crucial investments. When AI is aligned with principles of human rights, fairness, and transparency, it can drive innovation that is both effective and socially trustworthy. Instead of viewing AI ethics as a constraint, it should be recognised as the foundation for sustainable technological progress.

Ethical AI is a collective societal responsibility. The deployment of advanced AI systems, particularly in industries, occurs within a complex and rapidly changing governance environment, which is highly fragmented across different jurisdictions and sectors, preventing consistent ethical enforcement. This requires moving beyond traditional government regulation to a broader concept of governance that includes many stakeholders, such as international and regional organisations, national governments, businesses and Non-Governmental Organisations (NGOs). Because AI brings social and economic uncertainties, regulations should mainly aim to prevent misuse. Policy development should use a collaborative approach that relies on trust, negotiation, and shared values among all involved parties.

A significant discovery observed across all key sectors is the widespread challenge of algorithmic bias and discrimination. Algorithmic bias is mainly caused by training data that doesn't properly represent different groups, often due to historical or societal imbalances. In areas like Connected Automated Vehicles (CAVs), models designed for functions such as driver drowsiness detection demonstrate performance inconsistencies across diverse demographic groups, revealing that reliance on non-representative training data results in unequal safety outcomes. Similarly, in systems intended to promote digital inclusion, such as web accessibility checkers, the current research focus disproportionately favours visual impairments, thereby reinforcing existing technological disadvantages for users with auditory, cognitive, or motor impairments. In industrial applications, particularly compliance screening,

historical training data frequently perpetuates previous discriminatory practices and introduces cultural bias when systems are applied across varied global jurisdictions. The fundamental technical difficulty of simultaneously achieving different established fairness criteria, such as demographic parity and equal opportunity, further compounds the challenge of designing unbiased industrial systems.

Ethical challenges revolve around ensuring transparency and explainability, fairness and non-discrimination, privacy protection and clear accountability. Many ethical issues stem from the “black box” nature of machine learning models, making it difficult to understand how decisions are made. Without mechanisms for explainable AI, it’s nearly impossible to audit decisions, identify biases, or hold parties accountable when problems occur. This lack of clear internal logic limits meaningful human oversight, preventing affected individuals from contesting automated decisions, and undermines public trust.

While Explainable AI (XAI) technologies aim to provide explanations, they introduce new risks, including the potential for cultural bias in the explanations themselves, the risk that malicious actors could exploit these explanations, the tendency to shift ethical responsibility onto AI developers, and the possibility of giving users a false sense of confidence if explanations are oversimplified. Adding to this, accountability is often weakened because AI decision-making is so complex that it's hard to find clear and legal ways to hold anyone responsible when the system causes harm or discrimination.

Major challenges, including systemic bias, the lack of transparency in complex AI and crucial privacy issues are particularly magnified in advanced systems like Large Language Models (LLMs), Knowledge Graphs (KGs), Natural Language Processing (NLP), speech recognition and automated code generation (AutoML).

The need to collect and combine large amounts of data greatly increases privacy concerns across all modern AI systems. Privacy risks are magnified because AI requires collecting huge amounts of data, often sensitive information, and because the technology can be used for harmful purposes (dual use). NLP systems created for positive purposes can be repurposed for harmful ends (dual use), including the generation of misinformation, deepfakes, and automated social engineering campaigns. In CAVs, the continuous monitoring of biometric and behavioural data raises serious questions about intrusive surveillance and the absence of adequate consent mechanisms. Compliance screening also handles a lot of sensitive business information. Even in code generation, using huge datasets from the internet risks models remembering and leaking private or sensitive information. This calls for quick development of privacy-preserving learning techniques. Large Language Models (LLMs) require vast amounts of personal data, raising serious privacy concerns. There is a risk that they have memorised and could accidentally reveal sensitive data, such as private keys, from their massive training sets. Voice assistants and other natural language interfaces risk users unintentionally sharing sensitive information, and voice data can be used for advanced fraud through voice cloning.

NLP models, trained on huge datasets, often reflect and amplify societal biases (e.g., related to gender and race), leading to unfair results. Bias gets worse in automated AI development (AutoML). Similarly, Code Generation AI can spread bad or vulnerable code practices if its training data is flawed. Similarly, Knowledge Graphs (KGs) are essential for organising large amounts of data and helping AI reason, but ethical issues arise from how they are built. Knowledge Graphs are often created from web-scraped data that may be unrepresentative and skew towards Western perspectives, marginalising other cultures. When data for specific

domains or demographics is scarce, the KG reinforces existing disparities. The processes used to extract relationships from data can be unclear. This makes it hard to identify where bias is introduced, which is critical when KGs are used in high-stakes decisions. Proactive strategies are crucial, demanding rigorous bias detection, data diversification to ensure representative knowledge structures, and the adoption of Human-in-the-Loop approaches to validate complex relationships.

Beyond privacy, the environmental impact of the computational resources needed for AI is also a concern for sustainability. The immense computational costs and energy footprint associated with training and deploying cutting-edge LLMs introduce crucial ethical concerns regarding both environmental sustainability and global economic inequality, as access to these resources remains concentrated among a small number of large technological corporations.

In response to these ethical and technical challenges, regulatory bodies, most notably the EU, have introduced foundational frameworks such as the AI Act. To succeed, the Act requires clear guidelines, risk classification, regulation of prohibited systems, and harmonised technical standards across the EU, emphasising risk management and setting standards from the start, rather than reacting after issues arise.

The AI Act represents a crucial attempt to balance the promotion of innovation and economic competitiveness with the imperative to protect fundamental human rights and freedoms. However, the success of this regulatory intervention critically depends on the implementation details, requiring complex multi-agency cooperation between the European AI Office, national authorities, and industry stakeholders. Critics warn that the Act's complexity might deter investment compared to non-EU countries. Implementation is further complicated by the practical uncertainties of policy processes, where local interpretations of standards can lead to substantial variations in regulatory practice across member states, a pattern previously observed with the GDPR. However, the legislation's complexity and tight timeline for setting technical standards pose significant challenges. This complexity might discourage investment in the EU compared to other regions.

Ultimately, overcoming the ethical and societal misalignment of modern AI deployments requires a comprehensive, multi-layered strategy. Technically, this requires continuous development of advanced privacy-enhancing mechanisms, such as federated learning architectures and on-board data processing to protect sensitive information, alongside the creation of culturally sensitive and audience-specific explainability frameworks. Institutionally, ethics must be embedded throughout the entire AI lifecycle. This requires organisations to adopt governance mechanisms, risk assessments, and interdisciplinary review boards to ensure proactive ethical compliance rather than reactive damage control. Operationally, the shift must be toward designing AI with an "ethics-by-design" and "accessibility-by-design" mindset, promoting better human-AI collaboration. All these aspects are reflected in the projects understanding and treatment of AI systems as sociotechnical systems embedded in broader context.

In conclusion, modern AI is at a critical crossroads. If the current fast-paced progress and fragmented governance continue, there is a risk of deepening social inequalities and reducing public trust. It's crucial to recognise the high stakes involved in closing the ethical gap. The path forward demands that all stakeholders, developers, users, and regulators work together, as none currently have a complete understanding of AI's societal implications, to ensure the power of AI is governed responsibly. The ethical future of AI will ultimately depend not on the

complexity of its algorithms but on the deliberate human decisions made today to address existing ethical gaps, guiding technology toward outcomes that are fair, transparent, and serve the public interest.

A precautionary approach is crucial, establishing safeguards and ethical frameworks proactively before harm arises. The swift advancement of AI has surpassed current ethical and regulatory systems. However, society holds the collective responsibility and resources to guide AI toward the common good. Its future relies less on algorithms and more on the moral and political decisions made by those who create, deploy, and oversee AI.

It's important to take precautions early by setting up safety measures and ethical rules before problems occur. AI is advancing quickly and has outpaced existing rules and guidelines. Still, society as a whole has the responsibility and resources to steer AI in the right direction. The future of AI depends more on the moral and political choices of the people who develop, use, and manage it than on the technology itself. With continued awareness, collaboration and integrity, the ethical challenges identified can be addressed, ensuring AI becomes a foundation for a fair, transparent and human-centered digital future.

3 Designing ethical and trustworthy AI

3.1 Translating EU recommendations into design requirements for AI and AutoML

The design process for the ALFIE project shall begin with a systematic translation of high-level EU ethics guidelines, such as those found in the AI Act and the Assessment List for Trustworthy AI (ALTAI), into concrete technical specifications. This section aims to ensure that "excellence and trust" are not merely policy slogans but are measurable features within the software. To achieve this, the AutoML platform shall integrate automated checks that map Article 9's risk management requirements directly onto the development workflow. This shall involve creating "guardrails" within the AutoML engine that prevent the generation of models failing to meet pre-defined safety or fairness thresholds.

The platform will focus on translating technical robustness requirements into formal restrictions. These restrictions are transferred from the ETD-Hub to the AutoML in the form of knowledge graphs, which then serve to guide the model generation pipeline. By doing so, the system aims to provide a "safety-first" environment for all users, where the complex legal language of the Official Journal of the European Union is replaced by clear technical benchmarks. Furthermore, the interaction layer will aim to guide users through a series of prompts that help define the ethical boundaries of their specific use case, ensuring that design requirements are established before the first line of code is even generated.

3.2 Bias identification and mitigation strategies

It is a well-known fact that the identification and mitigation of bias is a mission-critical process designed to prevent the reinforcement of existing societal inequalities through automated systems. Within the European regulatory framework, specifically regarding the data governance requirements set out in Article 10 of the AI Act, AI systems must identify potential discrimination throughout their entire lifecycle. This process is structured across three critical stages: the data collection phase, the model training phase and the post-deployment phase. By treating bias mitigation as a continuous loop rather than a one-time checklist, organisations aim to protect themselves against significant risks of litigation and reputational damage that arise when AI produces "wrongheaded conclusions" or inaccurate predictions for protected classes[17].

During the initial data stage, designers shall focus on the "garbage in, garbage out" problem by ensuring the foundation of the AI is inherently equitable. This involves scanning datasets for historical imbalances in representation to ensure no specific demographic is unfairly excluded. For example, if a training dataset for a recruitment tool predominantly features historically over-represented groups, the resulting model will likely mirror those prejudices. To counteract this, the design process aims to proactively curate datasets that are representative of the diverse European population. This stage is not merely about volume but about the quality

[17]L. Short, "Ethics in AI: Why it matters," *Harvard Professional & Executive Development Blog*, 2025. [Online]. Available: https://professional.dce.harvard.edu/blog/ethics-in-ai-why-it-matters/.

and provenance of data, ensuring that the information used to train models is free from the systemic biases that have historically marginalised certain communities.

A good tactic for reducing bias during the training phase is the adoption of a model-selection strategy that prioritises fairness alongside technical performance. In traditional development, a team might automatically select the single most accurate model. However, this model often carries the highest risk of encoded bias. Instead, the design aims to generate a diverse set of candidate models that achieve similar levels of predictive power. From this high-performing pool, developers shall deliberately select the model that demonstrates the lowest level of disparate impact on protected groups[18]. This “fairness-first” selection tactic demonstrates that high accuracy and social equity are not mutually exclusive but can be achieved simultaneously through thoughtful engineering. By choosing the most equitable model among the top performers, the system provides a robust defence against “algorithmic discrimination”.

Following the training phase, post-mitigation strategies must be applied to adjust final outputs, ensuring they remain balanced when faced with real-world complexities. This final stage involves continuous monitoring after the tool has been deployed to catch “model drift”, where an AI begins to develop new biases based on the fresh data it encounters in the wild[19]. By embedding these comprehensive strategies into the design, AI systems aim to foster greater social justice and prevent the accidental reinforcement of historical prejudices.

3.3 Transparency, explainability and accountability

Ethical AI design moves away from the “black box” logic that was previously described by making transparency a fundamental technical feature. Developers ensure that algorithmic decisions are accompanied by a clear rationale that a human can realistically interpret. This level of clarity is vital in high-stakes sectors like finance or healthcare, where the design must allow humans to trace the reasoning behind every critical decision, such as the denial of a loan or a life-altering medical flag[20]. By implementing Explainable AI (XAI) techniques, the system aims to translate complex internal mathematical weights into intuitive, natural-language explanations. This ensures that the “right to an explanation” is not just a legal theory but a functional reality for any user affected by an automated output.

Accountability within the system is maintained through a rigorous framework of traceability and documentation. Every version of a model, its specific training data and the hyper-parameters chosen are recorded in a tamper-proof audit trail. This serves a dual purpose: it allows developers to refine the system based on performance and ensures that, if a system behaves unexpectedly or causes harm, a clear chain of responsibility is established. For a system to be truly accountable, it must provide a meaningful opportunity for recourse, allowing users to

[18]J. Kleinberg, S. Mullainathan, and M. Raghavan, "A simple tactic that could help reduce bias in AI," *Harvard Business Review*, Nov. 19, 2020. [Online]. Available: https://hbr.org/2020/11/a-simple-tactic-that-could-help-reduce-bias-in-ai.

[19]T. A. Kustitskaya, R. V. Esin, and M. V. Noskov, "Model drift in deployed machine learning models for predicting learning success," *Computers*, vol. 14, no. 9, p. 351, 2025, doi: 10.3390/computers14090351.

[20]L. Short, "Ethics in AI: Why it matters," *Harvard Professional & Executive Development Blog*, 2025. [Online]. Available: https://professional.dce.harvard.edu/blog/ethics-in-ai-why-it-matters/.

challenge a decision and have it reviewed by a human operator who has access to the AI's underlying logic.

Furthermore, the design includes the creation of standardised “Model Cards”[21] and “Data Statements”[22]. These documents act as an “ingredients label” for AI, transparently listing intended uses, known limitations and potential risks. By making this information public, organisations aim to bridge the gap between abstract philosophical principles and the tactical deployment of AI in everyday operations. This transparency ensures that deployers are fully aware of the technology's boundaries before it is put into service, fostering a culture of honest disclosure that is essential for long-term public trust in digital ecosystems.

Transparency, accountability and explainability constitute core pillars of ethical and trustworthy Artificial Intelligence, particularly in systems that rely on automated model generation and optimisation, such as AutoML. In the context of ALFIE, where models are trained on heterogeneous tabular datasets and deployed in high-impact decision-making scenarios, these principles are not abstract ideals but operational requirements that must be embedded directly into the technical architecture. The XAI component developed within ALFIE is designed to operationalise these principles by systematically analysing both input data and trained models to identify, quantify, and communicate potential sources of bias.

Transparency through data and model introspection: Transparency refers to the ability of stakeholders, developers, deployers, regulators and affected users, to understand how an AI system is constructed, how it behaves, and what factors influence its outputs. In AutoML settings, transparency is especially challenging because model selection, feature engineering, and hyperparameter optimization are largely automated, often obscuring critical design choices.

The ALFIE XAI component addresses this challenge by introducing structured transparency mechanisms at two complementary levels: data transparency and model transparency. At the data level, the component inspects tabular datasets prior to and after model training to identify imbalances, under-representation, proxy variables, and correlations that may lead to discriminatory outcomes. Statistical analyses such as feature distribution comparisons across sensitive or protected attributes, missing-value patterns, and target-label imbalance metrics are used to reveal hidden biases embedded in the data. By explicitly surfacing these characteristics, the system enables users to understand not only what data are used, but also how their structure may shape downstream decisions.

At the model level, transparency is achieved by analysing how trained models utilise input features to produce predictions. Feature attribution methods suitable for tabular data, such as additive explanation techniques, are employed to reveal which variables exert the greatest influence on model outputs, both globally and for individual predictions. This dual-level introspection ensures that transparency is not limited to model performance metrics but extends to the underlying causal and statistical relationships that drive automated decisions.

---

[21] M. Mitchell *et al.*, "Model cards for model reporting," in *Proc. Conf. Fairness Accountability Transp. (FAccT '19)*, 2019, pp. 220–229, doi: 10.1145/3287560.3287596.

[22] T. Gebru *et al.*, "Datasheets for datasets," *arXiv preprint arXiv:1803.09010*, 2018. [Online]. Available: https://arxiv.org/abs/1803.09010.

Explainability as a bridge between technical systems and human oversight: Explainability is closely related to transparency but focuses specifically on making complex model behaviour interpretable and meaningful to human stakeholders. In high-risk AI applications, explainability is a prerequisite for meaningful human oversight, contestability, and trust. Without clear explanations, users cannot assess whether a model behaves reasonably, regulators cannot audit compliance, and affected individuals cannot challenge automated decisions.

The ALFIE XAI component is explicitly designed to support post-hoc explainability for AutoML-generated models operating on tabular data. Given that AutoML may select opaque models (e.g., ensemble methods or deep neural networks), explainability is provided independently of the internal model architecture. This design choice ensures that explainability remains available regardless of the specific model selected by the AutoML pipeline.

Explanations are generated at multiple levels of abstraction. Global explanations describe overall model behaviour, highlighting dominant features, interactions, and systematic patterns that influence predictions across the dataset. These explanations are critical for identifying structural model bias, such as undue reliance on sensitive attributes or spurious correlations. Local explanations, by contrast, focus on individual predictions, enabling users to understand why a specific data instance was classified or scored in a particular way. This is especially important in contexts where individual-level decisions have legal or social consequences.

Crucially, the explainability outputs are designed to be interpretable by non-technical stakeholders. Rather than exposing raw mathematical artefacts, the system translates explanation results into structured summaries, visual indicators, and qualitative descriptions that support informed decision-making. This aligns with the EU requirement that transparency obligations be meaningful, rather than purely technical or symbolic.

Accountability through bias detection and traceability: Accountability in AI systems requires that responsibility for outcomes can be assigned, audited, and, where necessary, contested. In automated pipelines, accountability is often weakened by the diffusion of decision-making across data, algorithms, and optimisation processes. The ALFIE XAI component strengthens accountability by enabling traceability of bias from outcomes back to their potential sources.

By jointly analysing data characteristics and model behaviour, the component supports the identification of two distinct but interrelated forms of bias: data bias and model bias. Data bias arises when training data reflect historical inequalities, sampling imbalances, or measurement errors, while model bias occurs when the learning process amplifies or introduces unfair treatment beyond what is present in the data. Distinguishing between these forms is critical for accountability, as mitigation strategies differ depending on the source of the problem.

The XAI component provides structured evidence that links biased outcomes to specific features, data segments, or model behaviours. This evidence can be logged and stored as part of the system's technical documentation, supporting internal audits and external regulatory inspections. In line with the EU AI Act's requirements for high-risk systems, such documentation enables organisations to demonstrate due diligence, risk awareness, and proactive mitigation efforts.

Furthermore, accountability is reinforced by supporting human-in-the-loop workflows. The explanations and bias indicators generated by the system are intended to inform human

judgement rather than replace it. Decision-makers can review, override, or refine model outputs based on explainability insights, ensuring that ultimate responsibility remains with human actors rather than being implicitly delegated to automated systems.

Managing risks associated with explainability: While explainability is essential, it also introduces new ethical and technical risks that must be managed carefully. Over-simplified explanations may provide a false sense of confidence, while overly detailed technical explanations may overwhelm users or expose sensitive information. There is also a risk that explanations themselves may encode cultural or contextual biases, particularly when presented without sufficient contextualisation.

The ALFIE approach mitigates these risks by adopting a context-sensitive explainability strategy. Explanations are tailored to the intended audience and use case, balancing completeness with interpretability. The system avoids presenting explanations as definitive causal truths, instead framing them as evidence-based insights subject to uncertainty and human interpretation. This design choice helps prevent the misinterpretation or over-reliance on explainability outputs.

Additionally, safeguards are implemented to prevent the misuse of explainability for adversarial purposes, such as reverse-engineering models or exploiting decision boundaries. Access to detailed explanation outputs can be restricted based on user roles, ensuring that transparency and accountability are achieved without compromising security or system integrity.

Contribution to ethical and trustworthy AutoML: By embedding transparency, explainability, and accountability directly into the AutoML workflow, the ALFIE XAI component transforms ethical requirements into concrete technical capabilities. Rather than treating ethics as an external compliance layer, the system integrates ethical analysis into the core model development and evaluation process. This enables continuous monitoring of bias and interpretability throughout the AI lifecycle, from data ingestion to deployment and maintenance.

In doing so, the component supports the broader ALFIE objective of developing AI systems that are not only performant, but also aligned with European values of fairness, human oversight, and fundamental rights protection. Transparency allows stakeholders to see how decisions are made, explainability allows them to understand why those decisions occur and accountability ensures that responsibility for outcomes remains clear and enforceable. Together, these principles form the foundation for ethical, trustworthy and socially sustainable AutoML systems.

3.4 Privacy and data protection measures

Privacy must be integrated into AI systems from the very first design choice, following the “privacy by design” principle. Systems shall practice strict data minimisation, ensuring that only the information absolutely necessary for a specific task is ever processed. To protect individual identities and prevent fraud or identity theft, developers utilise Privacy-Enhancing Technologies (PETs) like differential privacy and strong encryption for data both at rest and in

transit[23]. These measures aim to ensure that technological innovation never comes at the cost of personal fundamental rights or the security of Personally Identifiable Information (PII).

In addition to technical safeguards, AI architecture incorporates strict Identity and Access Management (IAM) policies. This “security-first’ approach limits data access only to authorised personnel, ensuring that privacy remains a technical reality rather than a mere policy promise. By aligning with European data protection standards and using anonymisation as a standard procedure, these systems provide a safe environment for handling sensitive information. This proactive stance on security effectively mitigates the legal and financial liabilities associated with inappropriate data usage or accidental leaks, which are major concerns for modern organisations.

By respecting these rights, the ecosystem aims to empower European citizens, ensuring they remain the ultimate owners of their personal data while benefiting from the efficiencies and insights that artificial intelligence provides. The design also facilitates “data subject rights” management, ensuring that individuals can exercise their right to access, rectify, or delete their information, even within a complex AI environment. This involves creating interfaces that are clear for users to navigate, allowing them to manage their digital footprint with confidence. Ultimately, these measures serve to build a resilient framework where data utility and individual privacy exist in a balanced, mutually reinforcing relationship.

3.5 Real-world examples of ethical AI implementation

The practical value of ethical AI is best demonstrated through its application in high-stakes sectors where the impact on human rights and economic opportunity is direct. As the EU moves toward the full implementation of the AI Act, several pioneering organisations have already begun integrating ethical-by-design principles into their core operations. In the healthcare sector, for example, the German-headquartered Siemens Healthineers has implemented XAI within its diagnostic imaging platforms[24]. By providing “confidence scores” and clear rationale summaries, these systems allow radiologists to understand the underlying logic of a diagnostic suggestion rather than blindly following an automated output. This human-in-the-loop approach has been successfully deployed in clinical settings across Germany and France, where AI-powered screening tools assist in the early detection of conditions like lung cancer while ensuring that final medical decisions remain under the control of qualified professionals.

In the financial sector, European institutions are increasingly utilising AI to reform credit scoring and fraud detection in alignment with GDPR and the AI Act. Large banking groups, such as the French BNP Paribas, have integrated data management and AI transparency into their strategic plans, putting more than 800 AI use cases into production as of late 2025[25]. These

[23]L. Short, "Ethics in AI: Why it matters," *Harvard Professional & Executive Development Blog*, 2025. [Online]. Available: https://professional.dce.harvard.edu/blog/ethics-in-ai-why-it-matters/.

[24]Siemens Healthineers, "From innovation to impact: How AI elevates patient care," Oct. 24, 2025. [Online]. Available: https://www.siemens-healthineers.com/perspectives/from-innovation-to-impact-AI-in-healthcare.

[25]BNP Paribas, "Data & artificial intelligence: Mastering and implementing AI at the heart of the bank," 2025. [Online]. Available: https://group.bnpparibas/en/our-commitments/innovation/data-artificial-intelligence.

systems aim to avoid “negative social impacts” by specifically training data scientists to identify and address sexist or cultural biases in credit risk models. By moving away from opaque “black box” algorithms, these banks ensure that their automated decisions are reproducible and auditable, which is essential for maintaining regulatory compliance and long-term consumer trust within the Eurozone.

The recruitment sector also provides a strong blueprint for ethical implementation through European-born platforms like the UK-based MeVitae. This platform integrates with major HR systems like SAP SuccessFactors to anonymise candidate parameters, redacting over 30 personal identifiers such as gender, ethnicity and university names.[26] This “skills-first” evaluation ensures that hiring decisions are based purely on merit and role alignment rather than biased historical data. By using this “fairness-first” model selection tactic, European companies are finding they can unearth a wider pool of high-potential candidates while actively promoting diversity in the workplace. These examples demonstrate that when AI is designed with ethics at its core, it results in more robust, reliable and socially responsible products that benefit a wider segment of the European population.

Furthermore, the EIT Health ecosystem has supported numerous startups, such as the Netherlands-based ScreenPoint Medical, which uses AI to predict breast cancer risk years in advance[27]. These tools are subject to rigorous clinical validation and are designed to provide clear user information, fulfilling the transparency requirements for high-risk AI systems. These pan-European initiatives prove that a commitment to ethical design shall remain the primary standard for the future of EU technology. By documenting these successes through “Model Cards” and public audits, these organisations provide a clear pathway for SMEs to innovate within a framework that respects both technological excellence and human-centric values.

3.6 Addressing technical challenges in implementing ethical AI

One of the most significant hurdles in this field is the “accuracy-fairness trade-off.” Frequently, the most accurate models mathematically are also the most biased because they over-optimise for patterns found in flawed historical data. In traditional machine learning, the goal is often the absolute minimisation of error rates. However, when the training data reflects societal prejudices, the model learns to replicate those prejudices with high precision. Design teams address this by developing multi-objective optimisation strategies that treat fairness and explainability as primary performance metrics rather than secondary constraints[28]. This technical shift demonstrates that an explainable and fair model is ultimately more valuable for society than one that prioritises raw accuracy alone, as the former is less likely to fail in unpredictable or discriminatory ways during real-world deployment. By mathematically

[26]Oxiway Limited, "MeVitae ethical talent screening: Anonymised recruitment in SAP SuccessFactors," *SAP Store*, 2025. [Online]. Available: https://www.sap.com/products/hcm/partners/oxiway-limited-mevitae-ethical-talent-screening.html.

[27]EIT Health, "Scaling AI in European healthcare: Success stories from the EIT Health accelerator," *EIT Health e.V.*, 2025. [Online]. Available: https://eithealth.eu/news-article/scaling-ai-in-european-healthcare/.

[28]J. Kleinberg, S. Mullainathan, and M. Raghavan, "A simple tactic that could help reduce bias in AI," *Harvard Business Review*, Nov. 19, 2020. [Online]. Available: https://hbr.org/2020/11/a-simple-tactic-that-could-help-reduce-bias-in-ai.

weighting fairness as an equal objective to accuracy, developers can find an optimal solution where the model remains highly effective without compromising its ethical obligations.

Another technical challenge involves the computational cost of running continuous ethical audits and explainability features. These processes require significant processing power and can potentially slow down system performance, which is a major concern for real-time applications. Generating local explanations for a complex neural network's decisions can be as resource-intensive as the initial prediction itself[29]. To overcome this, developers leverage more efficient algorithms and High-Performance Computing (HPC) infrastructure as recommended in the European AI Innovation Package. By optimising how ethical checks are performed, such as using lightweight proxy models for explanations, the goal is to provide "real-time" ethics, where a model is audited and explained without causing a noticeable delay for the end-user. This ensures that safety and speed can coexist in the competitive digital market.

In addition, developers must address the challenge of "model drift", where an AI's performance degrades over time as it encounters new data. A model that is fair upon deployment may become biased as it adapts to changing trends or shifts in the input data it receives. To maintain ethical standards, systems must include automated monitoring tools that flag whenever a model starts to deviate from its fairness benchmarks. This requires a sophisticated technical architecture that can perform "gut checks" and institutionalised reviews automatically[30]. By building these monitoring layers into the system's core, developers can ensure that the AI remains self-correcting. Solving these technical challenges allows for the creation of a sustainable framework for AI that remains both highly effective and deeply rooted in human-centric values over the long term, preventing the accidental reinforcement of historical prejudices.

From an XAI perspective, implementing ethical AI in real-world AutoML pipelines faces a set of practical technical challenges that extend beyond "adding explanations" after model training.

A first challenge is **explanation validity and faithfulness**: post-hoc methods applied to complex models (e.g., ensembles or deep architectures) can yield explanations that are plausible but not necessarily aligned with the model's true decision logic. This risk is amplified in AutoML, where model structures change frequently across optimisation cycles. To address this, the XAI layer should include sanity checks and stability testing (e.g., explanation consistency across seeds, folds, and small perturbations) and should explicitly report uncertainty or variance in feature attributions rather than presenting a single deterministic "reason". Furthermore, for these insights to be actionable, the platform must support the personalisation of explanations. To ensure relevance across the project, the XAI output should be tailored to the specific user, providing granular, technical metrics for developers while offering high-level, intuitive justifications for non-technical end-users or regulators.

[29]L. Short, "Ethics in AI: Why it matters," *Harvard Professional & Executive Development Blog*, 2025. [Online]. Available: https://professional.dce.harvard.edu/blog/ethics-in-ai-why-it-matters/.

[30]L. Short, "Ethics in AI: Why it matters," *Harvard Professional & Executive Development Blog*, 2025. [Online]. Available: https://professional.dce.harvard.edu/blog/ethics-in-ai-why-it-matters/.

A second challenge is **bias detection under imperfect data conditions**: tabular datasets often contain missing values, proxies for protected attributes, imbalanced labels, and distribution shift across subgroups, conditions that can cause explainers to highlight artefacts rather than meaningful signals. The XAI component therefore needs to pair model explainability with data diagnostics (subgroup coverage, target imbalance metrics, proxy correlation analysis, and drift checks) so that model-level explanations are interpreted in light of data limitations. A third challenge is operational scalability: explanation methods can be computationally expensive, especially for instance-level analysis at deployment time. Practical implementations must support tiered explainability, where lightweight global explanations and bias indicators run by default, and more expensive local explanations are triggered on-demand (e.g., for contested cases, audits, or high-impact decisions).

Fourth, there is a **security transparency trade-off**: detailed explanations can leak sensitive information about training data or model behaviour, enabling inference attacks or manipulation. Ethical XAI therefore requires access control, logging, and redaction policies (role-based explanation views, aggregation of sensitive features, and limits on per-instance disclosure).

Finally, a major socio-technical challenge is **actionability**: explanations that do not connect to mitigation steps can create “transparency theatre.” For ethical AI, XAI outputs should be coupled with concrete remediation pathways, e.g., recommending data rebalancing, feature review, fairness constraints, or human override thresholds, so that transparency translates into accountable governance and measurable risk reduction.

3.7 Summary of consolidated best practices

Moving from ethical theory to the reality of trustworthy AI requires a practical approach that supports the people building the technology. To ensure innovation remains rooted in European values, these practices bridge the gap between high-level ideals and operational reality by integrating responsibility into the development process from the outset.

3.7.1 Maintaining Human Oversight and Multi-Disciplinary Design

Trustworthy AI is fundamentally human-centric, requiring the integration of "human-in-the-loop" mechanisms throughout the entire system lifecycle. Rather than a reactive final audit, human oversight must be a deliberate design choice from the conceptual phase through to deployment. By assembling multi-disciplinary teams, comprising legal, sociological and ethical experts alongside data scientists, organisations can identify and mitigate cultural and ethical biases before they become embedded in the technical architecture. Engaging stakeholders and end-users further ensures that practical risks are identified and that the resulting technology remains inclusive and responsive to the needs of a diverse society.

3.7.2 Cultivating a Culture of Transparency and Public Accountability

Technical measures are only truly effective when underpinned by a genuine institutional commitment to accountability. Organisations should treat transparency as a strategic asset rather than a compliance burden. By adopting standardised tools such as Model Cards and Data Statements, developers can provide a clear “ingredients label” for AI systems. These documents must detail the model’s intended purpose, its technical limitations and any identified risks, enabling both users and regulators to make informed, evidence-based decisions.

Furthermore, a commitment to regular, independent third-party audits ensures that ethical claims are verified through external validation. Publicly disclosing how these standards are met in practice is essential for building and maintaining long-term trust with European citizens and consumers.

3.7.3 Institutionalising Ethical Governance

To ensure ethics is a shared responsibility, it must be embedded directly within the organisational management structure. This integration ensures that ethical considerations are not confined to a single department. Instead, they actively inform and guide business decisions at every level of the organisation. Establishing a formal AI Ethics Board facilitates the rigorous assessment of high-risk use cases; these bodies serve to evaluate the fundamental viability of projects, determining whether a development should proceed based on its alignment with core values, regardless of technical feasibility.

Governance frameworks should be structured so that engineers, product managers, and executives all maintain a stake in the ethical outcome. This institutional approach ensures that abstract philosophical principles are translated into clear, actionable operational processes across the entire organisation.

3.7.4 Continuous Adaptation and Long-term Commitment

As technology and societal expectations evolve, maintaining ethical AI requires a continuous commitment rather than a static achievement. A sustainable ecosystem depends on ongoing learning and proactive adaptation. To this end, technical frameworks should include automated monitoring for “model drift”, ensuring that fairness benchmarks are upheld as the AI encounters new data. Additionally, organisations should engage in broader community dialogues to share insights and lessons learned. By adopting these measures, companies can comply with the world’s most rigorous safeguards while remaining at the forefront of human-centric design.

The consolidated best practices discussed in this section are summarised in Appendix A (Table of Consolidated Best Practices).

4 Ethical AI requirements in ALFIE Use Cases

4.1 Overview of methodology

The assessment of ethical AI requirements in the ALFIE use cases is based on a common methodological approach applied across all three scenarios. The aim is to ensure that ethical considerations are examined in a consistent manner, while taking into account the specific technical and operational characteristics of each use case.

As a first step, each use case is described in terms of its context, the role of the AI system, and the way in which it is intended to be used in practice. This initial analysis makes it possible to identify the points at which the use of AI may raise ethical or legal concerns. On this basis, **the methodology proceeds to a preliminary risk assessment and classification of each use case in accordance with the risk-tiered framework established by the AI Act** (e.g., Minimal, Limited or High Risk as described in Section 2.3.2 of the current Deliverable).

On that basis, relevant ethical requirements are identified, drawing on the applicable EU legal framework and widely recognised principles for the responsible use of AI, such as transparency, human oversight, fairness and accountability.

The next step consists in identifying potential risks associated with the use of AI, including risks related to data processing, decision-making, bias, transparency and security. For each risk, appropriate mitigation measures are then defined, combining technical and organisational safeguards where necessary.

Disclaimer on Regulatory Compliance: It should be noted that the ALFIE project is currently in a **research and development (R&D) phase**. Under Article 2(8) of the EU AI Act, the regulation does not apply to AI systems or models specifically developed and put into service for the sole purpose of scientific research and development, nor to R&D activities prior to an AI system being placed on the market or put into service. Consequently, formal compliance with the Act is not yet compulsory for the current scope of these use cases. However, the ALFIE project proactively adopts these standards as a “best practice” framework to ensure future-proofed, ethical and trustworthy innovation.

4.2 Use Case 1- CTL

4.2.1 Description of the use case

The IRIS use case is centered around technologies that evaluate a driver’s alertness in real time within **Connected Automated Vehicles (CAVs)**. As these vehicles function across different levels of automation, especially Levels 2 and 3, where drivers must be ready to regain control, systems that can detect fatigue are crucial for vehicle safety. Driver State Monitoring (DSM) technologies are designed to spot early signs of drowsiness or decreased vigilance and alert the driver before these conditions affect road performance. To accomplish this, DSM technologies utilise a mix of visual, physiological and behavioral indicators. Visual methods include in-cabin cameras that monitor aspects like eye closure, blinking rate, yawning or head movements. Physiological techniques use sensors to assess signals such as EEG, ECG, or breathing patterns, providing deeper insights into the driver’s cognitive state. Behavioral monitoring looks at driving patterns, such as steering consistency, lane-keeping or pedal usage, to identify subtle changes that may indicate developing fatigue.

The primary aim of the IRIS use case is to enhance the precision and reliability of drowsiness detection among varied driver demographics, primarily through visual sensors, supplemented by some physiological data like heart rate. Current systems may show inconsistent performance due to differences in facial features, skin tones, and fatigue indicators across racial groups, impacting the interpretation of visual and physiological signals. Consequently, the IRIS use case focuses on developing models, with the help of AutoML, and monitoring techniques that function reliably for all drivers, aiming to reduce performance disparities and address bias in detection. The integration of multimodal data, merging visual cues and physiological metrics, supports this goal by facilitating more holistic evaluations that do not overly rely on one single signal type.

Operating continuously and in real-time, IRIS utilises a mobile application that uses the front camera to capture images of the driver and predict drowsiness in real time. The app is used as a test medium to load all the models produced from the AutoML platform and test them in both test and real-life environments and assess their performance. The IRIS app issues alerts via audio signals and visual indications when the current model predicts that the driver is drowsy. Collectively, the use case aims to illustrate how advanced monitoring techniques can promote safer vehicle operations while ensuring consistent performance across a diverse user base, by eliminating any facial and racial biases.

The IRIS use case is considered a **High Risk** system under the EU AI Act. Numerous ethical challenges and dilemmas emerge, encompassing issues such as privacy concerns, data security, informed consent and potential biases. These complex challenges will be thoroughly examined and discussed in the following sections to highlight their implications and mitigation measures.

4.2.2 Ethical requirements for AI use

In accordance with the AI Act, the DSM system must be governed by ethical requirements that ensure the technology will remain safe, inclusive and subservient to human agency. As a high-risk system within the context of CAVs, the following principles must be integrated into the design to address the safety-critical nature of Levels 2 and 3 vehicle automation.

i. Equitable Safety through Technical Redundancy

The foundational mandate must be the provision of equitable safety, ensuring that the system will identify drowsiness with consistent precision across all driver demographics. To address historical performance gaps in visual-only sensors, where error rates should be monitored for disparities regarding skin tone, this use case must adopt a multimodal data architecture. By synthesising visual indicators with physiological data, such as heart rate variability, the system will ensure technical robustness. This redundancy will prevent the technology from becoming discriminatory, maintaining a high safety standard regardless of a driver's physical features or environmental lighting.

ii. Data Integrity and Representative Governance

To meet the stringent data governance standards for high-risk AI, the project must be committed to representativeness and resilience. This will involve the use of training and validation datasets that reflect almost the full diversity of the European population, including varied facial architectures and age groups. By initialising these parameters during the development phase via AutoML, the system must ensure that detection models will be

resilient against “noise” and will remain free from “hard-coded” algorithmic bias. Data integrity must be treated as a core ethical responsibility that will guide all technical operations.

iii. Preserving Human Agency and Autonomy

The “human-in-the-loop” principle must be applied to protect human agency, ensuring the AI will function as a supportive layer rather than a replacement for driver responsibility. Ethical implementation requires that oversight must be both meaningful and non-intrusive. Alerts must be designed to be clear and actionable to prevent “automation bias” where a driver might become over-reliant on the technology. Ultimately, the system must provide timely audio and visual indications that will empower the driver to regain control, fulfilling the ethical obligation to enhance road safety while respecting the user's ultimate authority over the vehicle.

4.2.3 Identified ethical risks and mitigation measures

The deployment of DSM systems within CAVs presents specific technical and ethical challenges that must be addressed to ensure safety and compliance. Under the framework of the AI Act, the following risks have been identified, alongside their corresponding mitigation measures.

i. Mitigation of Demographic and Racial Bias

A primary challenge in visual monitoring is the potential for algorithmic bias, where differences in skin tone or facial structure lead to inconsistent performance. In Level 2 and 3 automation, failing to detect drowsiness in a specific demographic is a critical safety failure. To address this, the project uses AutoML to test and refine models against diverse datasets. This ensures that detection algorithms are validated across a wide spectrum of users, removing performance gaps and ensuring that safety benefits are shared equitably.

ii. Addressing Environmental Resilience and Sensor Limitations

Relying solely on in-cabin cameras introduces the risk of system failure in poor lighting or when a driver wears sunglasses. To build a more robust system, multimodal data should be integrated. By combining visual cues with physiological indicators like heart rate, the system will create a necessary safety buffer. This approach ensures that the monitoring remains effective even if one signal is obstructed or degraded, avoiding a dangerous over-reliance on a single type of sensor.

iii. Managing Real-Time Model Drift

The shift from controlled testing to real-world driving can cause “model drift”, where an AI’s accuracy drops as it encounters new data patterns. This risk can be managed through a continuous monitoring framework supported by a dedicated mobile testing application. By using this app as a real-time medium, developers can load and assess models in both simulated and live environments. This will enable the immediate identification of performance issues and the deployment of the iterative updates required to maintain system reliability throughout its lifecycle.

iv. Data Sensitivity and Privacy Governance

Collecting real-time physiological data and in-cabin video naturally raises privacy concerns. These can be addressed by applying strict data minimisation principles. Personal data are

processed in real-time to trigger safety alerts, with clear documentation provided via Model Cards and Data Statements. This “ingredients label” approach ensures that users and regulators understand exactly how data are used and protected, which is essential for maintaining public trust.

4.3 Use Case 2 - UAB

4.3.1 Description of the use case

The UAB use case is centred around the right to access information. It involves using the ALFIE AutoML platform to process and analyse web content for compliance with Web Content Accessibility Guidelines[31] **(WCAG**), targeting improvements in accessibility to cater for the needs to access information for a diverse range of people, especially for blind and visually impaired users. The system ingests HTML files and performs automated checks across multiple WCAG criteria, assigning accessibility scores and identifying violations.

By focusing on the four core principles of **Perceivability, Operability, Understandability and Robustness (POUR)**, the platform identifies structural and semantic barriers within websites. This automated evaluation ensures that the digital infrastructure is optimised for users who rely on assistive technologies, such as screen readers.

The technical scope of this use case is defined by the rigorous assessment of reported several critical success criteria, at level A and AA which are the minimum requirements in the current European accessibility legislation to ensure access to information and democratic participation of all citizens. The platform evaluates **Non-Text Content (Success Criterion 1.1.1 – level A)** by analysing the presence and contextual quality of attributes for non-text content, ensuring that no essential information is hidden from visually impaired users, such as alt text for images. Furthermore, the system conducts a deep analysis of **Info and Relationships (Success Criterion 1.3.1 – level A)**, where the AI validates the Document Object Model (DOM) to ensure that heading hierarchies and form associations are identified. To address **Meaningful Sequence (Success Criterion 1.3.2 – level A)**, the AutoML models verify that the reading sequence is logical when content is presented. Visual barriers are further mitigated through the automated calculation of colour **Contrast (Success Criterion 1.4.3 – level AA)**, ensuring a minimum ratio of 4.5:1, and the verification of **Keyboard (Success Criterion 2.1.1 – level A)**, which guarantees that all site functionality is reachable without the use of a mouse.

This use case can be classified as **Limited Risk**. The ALFIE AutoML platform functions as a technical diagnostic tool designed to enhance digital accessibility by identifying non-compliance with WCAG 2.2 standards. As the system focuses on the structural and semantic analysis of HTML and DOM elements, rather than making autonomous decisions regarding natural persons or critical infrastructure, it does not meet the criteria for “High-Risk” AI. Its primary purpose is to provide objective, automated feedback to ensure that websites are perceivable, operable, understandable and robust for users of assistive technology. In accordance with Article 50 of the AI Act, the system is designed to meet **transparency obligations** by ensuring that users (auditors and developers) are made aware that they are

[31] World Wide Web Consortium, "Web Content Accessibility Guidelines (WCAG) 2.2," 2024. [Online]. Available: Web Content Accessibility Guidelines (WCAG) 2.2 .

interacting with AI-generated assessments, thereby supporting fundamental rights to accessibility without introducing algorithmic bias or safety concerns.

4.3.2 Ethical requirements for AI use

The ALFIE AutoML platform is designed to be governed by ethical requirements so to ensure the technology remains both safe and accessible. A foundational mandate for this use case is the provision of equitable access. Thus, the system must identify barriers with consistent precision across diverse web architectures. By synthesising structural HTML data with ARIA attributes, the platform aims to ensure technical robustness and avoid providing a false sense of compliance.

i. Data Integrity and Inclusive Governance

To meet stringent data governance standards, the project is committed to the representativeness of its underlying models. This involve using training and validation datasets that reflect the full diversity of EU web development standards. Because the training data are sourced from a wide spectrum of digital environments, the system must ensure that detection models remain resilient against "code noise". Furthermore, the pipeline should be free from algorithmic bias that might otherwise penalise non-standard but accessible coding practices. Data integrity is treated as a core ethical responsibility guiding all operations within the AutoML pipeline, ensuring the data are handled with the highest degree of accuracy.

ii. Preserving Human Oversight and Control

The "human-in-the-loop" principle is applied to protect human agency, ensuring the AI functions as a supportive layer for accessibility auditors rather than a total replacement for expert judgment. Ethical implementation requires that oversight remains meaningful, particularly when the system assigns accessibility scores. To fulfil transparency requirements, the platform explicitly discloses the use of AI to the auditor, ensuring they remain aware that the findings are machine-generated. To empower developers to make informed remediation decisions, the platform provides clear justifications for any flagged violations. Ultimately, the system aims to provide actionable insights that will allow the user to enhance digital accessibility, fulfilling the ethical obligation to respect the user's authority over the final compliance status of a website.

4.3.3 Identified ethical risks and mitigation measures

The deployment of the ALFIE platform presents specific technical and regulatory challenges that must be managed to ensure compliance with both the GDPR and the AI Act. Under these frameworks, the following risks have been identified, alongside the measures taken to address them.

i. "False Compliance" and Qualitative Errors

A primary risk in automated auditing is the potential for "false compliance", where the AI may overlook qualitative failures. For example, a system might flag an image as compliant because alternative text is present, even if that text is contextually irrelevant or poor in quality. To mitigate this, the project will use AutoML to refine models against "gold

standard"[32] accessibility examples. This process will ensure that algorithms are validated against real-world user experiences, moving beyond a simple "binary" check to a more nuanced understanding of accessibility. Additionally, a clear transparency notice will be integrated into the user interface to mitigate automation bias, reminding the user that AI outputs are recommendations that require final human verification.

ii. Model Drift and Evolving Standards

The rapid evolution of web technologies and standards introduces the risk of "model drift", where an AI's accuracy declines as new coding patterns emerge. This risk can be managed through a continuous monitoring framework where the data are periodically reviewed and updated to reflect the latest WCAG success criteria. This will ensure that the platform remains a reliable tool for long-term compliance in a shifting digital landscape.

iii. Data Sensitivity and Privacy Governance

While the system primarily processes public-facing HTML, the ingestion of web assets requires careful attention to data sensitivity. These concerns can be addressed by applying strict data minimisation principles. Personal data will be excluded from the processing scope, and all technical metadata will be documented via Model Cards and Data Statements. Thus, all stakeholders will be able to understand exactly how the data are utilised and protected, ensuring full compliance with the documentation requirements for Limited-Risk systems, which is essential for maintaining trust in the ALFIE ecosystem.

4.4 Use Case 3 - BOSCH

4.4.1 Description of the use case

The Bosch use case represents a relevant application of AutoML technology in the domain of compliance monitoring and corporative process optimisation. At the heart of this use case lies the challenge of managing **Business Partner Screening (BPS)** alerts that are continuously generated to monitor third-party relationships for potential violations of the Bosch Code of Business Conduct. Compliance officers at Robert Bosch currently receive a substantial volume of these alerts, each requiring manual evaluation to determine whether the flagged third-party relationship warrants further investigation or corrective action.

The existing manual process for handling these alerts presents significant operational challenges. Compliance officers must individually review each alert, assess the severity and relevance of the potential violation, and make prioritisation decisions based on their experience and available information. This approach creates bottlenecks in the compliance workflow and can lead to inconsistencies in how similar alerts are prioritised across different officers or time periods. Furthermore, the manual nature of the process means that critical alerts requiring immediate attention may not always be identified and escalated with the urgency they deserve.

[32] In regulatory and technical terms, this is often referred to as **"Ground Truth"** or **"Reference Data".** Within the EU AI Act (Article 10), it is specifically mentioned as the requirement for "high-quality training, validation, and testing datasets" that are relevant, representative and free of errors. *EU AI Act (Article 10): Requirements for High-Quality Training, Validation, and Testing Datasets*, 2024. [Note: Specifically referring to "Ground Truth" or "Reference Data" terminology].

The proposed AutoML solution aims to transform this compliance monitoring process by leveraging machine learning capabilities to automatically learn from the existing historical database of previously classified alerts. The system will analyze patterns in how alerts have been categorised and prioritised in the past, identifying key indicators that distinguish high-priority alerts requiring "further investigation needed" status from those that can be resolved through standard procedures. This approach not only promises to improve the efficiency of the compliance process but also ensures more consistent and objective prioritisation decisions.

This use case is classified as **High Risk**. The Bosch BPS AutoML platform is deployed within the domain of compliance monitoring and legal risk assessment, where it assists in the evaluation and prioritisation of potential violations of the Code of Business Conduct. Because the system's outputs directly influence the investigation and treatment of third-party business partners, it falls under the category of AI systems used for the “access to and enjoyment of essential private and public services” and legal scrutiny. Consequently, the ethical issues, as well as the necessary mitigation and compliance measures, will be analysed in the next sub-sections to ensure that automated prioritisation does not lead to inconsistent or discriminatory outcomes for the entities being screened.

4.4.2 Ethical requirements for AI use

In accordance with the AI Act, the BPS screening system must be governed by ethical requirements that ensure the technology remains a tool for human empowerment and does not introduce systemic bias into corporate governance.

i. Algorithmic Fairness and Non-Discrimination

The foundational ethical requirement for this use case is the prevention of discriminatory outcomes. Because the system analyses third-party relationships, it is vital that the AI does not inadvertently penalise business partners based on sensitive attributes such as geographic location or corporate age, unless these are legitimate risk factors documented in the Code of Conduct. The training must be audited to ensure that the machine learning models do not codify historical biases present in previous manual classifications. By maintaining a focus on objective risk indicators, the system will ensure that all business partners are evaluated through a fair and equitable lens.

ii. Data Integrity and Governance of Sensitive Information

To meet the high-risk data governance standards required for compliance monitoring, the project is committed to the absolute integrity of the underlying datasets. Since the screening process involves commercially sensitive and potentially personal information, all processing must strictly adhere to the GDPR. The training data must be handled using robust encryption and anonymisation techniques where appropriate. Ensuring that the data are representative of the current global regulatory environment must also be treated as a core ethical responsibility, preventing the models from becoming obsolete or inaccurate as international compliance standards evolve.

iii. Human Agency and the “Human-in-the-Loop” Mandate

The principle of human agency is central to this use case, ensuring that the AI functions as a prioritisation assistant rather than an autonomous decision-maker. Ethical implementation requires that the final determination for any “further investigation needed” status remains with a qualified compliance officer. The system is designed to provide XAI

outputs, offering clear justifications for why a specific alert was prioritised. This transparency will ensure that human oversight is meaningful and that officers can override AI suggestions based on nuanced contextual information that the model may not capture.

4.4.3 Identified ethical risks and mitigation measures

The deployment of AutoML for BPS presents specific technical and ethical challenges that must be addressed to maintain regulatory compliance and public trust.

i. Automation Bias and Over-Reliance

A significant risk in prioritisation systems is “automation bias”, where compliance officers may become overly reliant on the AI’s high-priority flags and neglect alerts categorised as low-risk. To mitigate this, the project shall implement a periodic random audit of low-priority alerts by senior compliance experts to avoid false negatives in the classification output. This will ensure the model’s accuracy is constantly verified against human standards and will prevent critical violations from slipping through the system due to an undue trust in automated outputs.

ii. Model Drift in Global Compliance

Global compliance standards and third-party risk profiles are subject to frequent change. There is a risk of “model drift”, where an AI trained on 2024 data becomes less effective as new types of corporate violations emerge in 2025 and 2026. To address this, the project shall adopt a continuous learning framework where the data are reviewed periodically. The AutoML platform allows for the rapid retraining of models to incorporate new EU regulatory requirements, ensuring the system remains resilient and effective throughout its lifecycle.

iii. Data Privacy and Minimisation

Processing BPS alerts involves handling data that could impact the reputation of third-party entities. The risk of data leakage or improper processing can and should be mitigated through strict data minimisation principles. Only the specific data points required for risk assessment must be ingested by the ALFIE platform. Furthermore, the use of Model Cards can provide a transparent “ingredients label” for the AI, documenting how the data are used, which is essential for maintaining the legal standing of the project and fulfilling ethical obligations regarding corporate transparency.

4.5 Cross-case assessment based on best practices

The evaluation of the ALFIE AutoML platform across the aforementioned three diverse cases (Automotive (DSM), Web Accessibility (WA) and Corporate Compliance (BPS)) reveals a unified set of best practices for ethical AI deployment. While the technical inputs range from physiological signals to structural HTML and corporate records, the underlying requirements for safety and fairness remain closely aligned. This section brings these shared strategies together to show how they form a unified approach to building and maintaining trustworthy AI.

i. Building Reliability through Data Diversity and Integrity

A clear finding across the use cases is that high-risk systems require tailored strategies to ensure reliability. While many scenarios benefit from technical redundancy, it is recognised that a multi-source approach is not always feasible. In instances where the architecture necessitates a single data stream, the platform focuses instead on ensuring single-source

integrity. By synthesising diverse information where possible and enforcing rigorous validation and precision protocols where it is not, the platform ensures that the resulting models remain robust and accurate across all operational constraints.

ii. Ensuring Data Integrity and Inclusive Representation

Across the board, the use cases demonstrate that data integrity is a core ethical responsibility rather than a secondary task. By adopting a strategy of “representative governance”, the project aims to ensure that training data are sourced from a wide spectrum of environments, including diverse facial architectures, varied web development standards and global regulatory profiles. Using AutoML to stress-test these models against “gold standard” examples has proven vital in removing performance gaps. This prevents the system from repeating historical biases and ensures a fair and equitable approach for all users.

iii. Respecting Human Judgment and Final Authority

The “human-in-the-loop” principle acts as a common protective layer across these different applications. In every scenario, the AI is positioned as a supportive partner rather than a replacement for human judgment. In the BPS and accessibility cases, the system provides clear justifications for its findings, allowing experts to override suggestions based on context. Similarly, in the automotive use case, the system focuses on empowering the driver to regain control. This common thread ensures that, regardless of the application, the user always retains ultimate authority over the final outcome.

iv. Managing Continuous Validation and System Growth

Finally, all three use cases show that ethical AI requires a long-term commitment to learning. Moving from a controlled lab to a real-world setting introduces “model drift” as driving habits, web standards or global laws change. The assessment confirms that a continuous monitoring framework is necessary to keep the AI aligned with its original goals. By periodically reviewing how the data are behaving in live environments, the platform can be updated to remain reliable and safe throughout its entire operational lifecycle.

5 ALFIE's Best practices Blueprint for ethical AI and AutoML integration

5.1 Insights and recommendations from ALFIE use cases

The following standards have been formalised based on the empirical results of the ALFIE use cases. These requirements translate specific technical observations into a set of platform-wide operational mandates, aligning with the latest research on high-stakes AI resilience.

i. Strengthening Reliability in AI Applications with Ethical Risks

Insights from the project confirm that reliability strategies must be tailored to the specific nature of each AI application. For systems involving significant ethical or safety risks, the platform prioritises robustness through either technical redundancy or single-source integrity, depending on the technical constraints of the environment. In scenarios where a multimodal approach is feasible, independent data streams are fused to create a vital safety layer that remains resilient even if one modality is degraded. Conversely, in applications where a multi-source architecture is not applicable, the platform ensures maximum accuracy and ethical alignment through rigorous data validation, semantic pairing and high-precision governance to ensure the primary input remains dependable across all operational contexts.

ii. Systematic Bias Auditing and Representative Governance

The risk of algorithmic bias necessitates a rigorous data governance protocol. The platform prioritises the representativeness of training data to prevent performance disparities that could lead to discriminatory outcomes. This is particularly critical in complex use cases, such as the BPS use case, where datasets may harbour subtle biases, including the under-representation of specific geographical regions or company types. To mitigate these risks, the platform mandates the use of the AutoML validation module to audit datasets against "gold standard" benchmarks, such as the Bias Evaluation and Assessment Test Suite (BEATS)[33]. This process requires formal verification that training data are demonstrably representative of the diverse demographics and technical standards of the EU, fulfilling the latest mandates for equitable AI governance.

iii. Meaningful Human Oversight through Decision Justification

Across all use cases, human agency was identified as the final safeguard against automation bias. To ensure that human oversight is meaningful, the platform integrates XAI as a foundational feature. The platform is designed to provide granular, human-readable justifications, which are documented through automated Model Cards, a format now essential for regulatory compliance checks in the EU[34]. This transparency empowers

[33] A. Abhishek, L. Erickson, and T. Bandopadhyay, "Data and AI governance: Promoting equity, ethics, and fairness in large language models," *MIT Sci. Policy Rev.*, vol. 6, pp. 139–146, 2025, doi: 10.38105/spr.1sn574k4lp.

[34] A. Mehraj *et al.*, "AI model cards: State of the art and the path to automated regulatory compliance checks," *Tampere Univ. Press (Trepo)*, 2025. [Online]. Available:

expert auditors to challenge or override AI suggestions, fostering a collaborative human-AI framework where the technology remains subservient to professional judgement.

iv. Dynamic Compliance and the Monitoring Lifecycle

The transition to real-world deployment introduces the risk of "model drift", where accuracy degrades over time. ALFIE implements a continuous improvement protocol that goes beyond the initial deployment. Deployed models are subject to periodic performance reviews to ensure that underlying data are updated to reflect evolving regulatory criteria, such as new WCAG versions. Crucially, this protocol incorporates the continuous checking of biases to detect and remediate any discriminatory patterns that may emerge as the system interacts with live data. This lifecycle-based approach guarantees that AI systems remain safe and ethically aligned as global standards and user behaviours continue to shift.

5.2 Integrating EU policy recommendations in ALFIE's AutoML platform

The technical architecture of the ALFIE platform is specifically engineered to automate the regulatory requirements of the AI Act, ensuring that compliance is an inherent part of the development workflow. This integration is achieved through a suite of modular tools designed to handle the complexities of high-risk AI governance. At the data level, the platform incorporates a dedicated Data Profiling Module that addresses the stringent mandates of Article 10 of the AI Act. This module performs automated audits to verify that the data are relevant, representative, and accurate. By executing statistical checks for demographic parity and identifying gaps in training sets before a model is finalised, ALFIE will effectively eliminate performance disparities that could lead to discriminatory outcomes. Furthermore, every dataset processed within the AutoDW (Data Warehouse) is assigned a unique identifier and a machine-readable Data Statement, ensuring full traceability throughout the system lifecycle.

Transparency and information provision, as required by Article 13 of the AI Act, are managed through the automated generation of standardised Model Cards[35]. These digital documents act as a comprehensive "ingredients label" for the AI, detailing its intended purpose, performance benchmarks and any known limitations. Because these cards are produced in a JSON-based format, they provide a consistent and verifiable record that enables humans to interpret the system's outputs correctly. This is complemented by the platform's automated logging feature, which satisfies the record-keeping requirements of Articles 11 and 12. By maintaining a tamper-resistant "paper trail" of every hyperparameter adjustment and validation result, ALFIE provides regulators with the evidence needed to verify that the system was developed according to strict ethical standards. Ultimately, the integration of natural language explanations within the XAI component ensures that the technology remains subservient to human agency, fulfilling the mandate for human oversight as defined in Article 14 of the Regulation.

[35] https://researchportal.tuni.fi/en/publications/ai-model-cards-state-of-the-art-and-path-to-automated-use/

5.3 Continuous improvement

To support the long-term reliability and ethical alignment of its AI systems, ALFIE aims to establish a systematic process for continuous refinement. This process is designed to provide a methodology for evaluating model performance within real-world environments, with the objective of maintaining compliance with evolving technical and legal standards.

As already noted, managing the risk of “model drift” is a primary technical objective of this refinement cycle. Should real-world conditions shift, whether through changes in driver behaviour or evolving web standards, the relationship between input data and predicted outcomes may weaken. To address this, ALFIE could employ specific statistical triggers, (such as the Population Stability Index (PSI)), to flag significant divergences between live data and the original training baseline. If a deviation were detected, the platform would ideally issue an automated alert and initiate a formalised retraining pipeline. Crucially, this process is intended to include the continuous checking of biases. Such a mechanism would allow for any discriminatory patterns emerging from new data environments to be detected and remediated, thereby supporting the model's precision and ethical integrity over time.

A central element of this lifecycle would be the bi-directional feedback loop established between the AutoML platform and the ETD-Hub. This connection is intended to allow qualitative insights from legal experts, policymakers and citizens to inform the technical evolution of the models. By integrating this human-centric feedback, the framework’s goal is to move beyond static compliance to a model of “human-on-the-loop” refinement. In this way, the technology can evolve in response to shifts in the legal and social context of the EU, without displacing the ultimate authority of human judgment.

6 Conclusions and next steps

6.1 Lessons learned

The process of developing this initial best practices blueprint for ethical AI and AutoML integration has revealed important lessons about how ethical and legal values can and must be translated into the development of real AI systems.

One of the clearest lessons is that ethical risks are not abstract; they are already embedded in the design of AI systems at early stages, long before deployment. Our analysis of ALFIE's three use cases showed how ethical risks manifest differently depending on the application context: biometric bias in driver monitoring, interpretability gaps in corporate compliance models, or limitations in automated accessibility checks. These risks are not theoretical; they have practical consequences for fairness, inclusion, accountability and safety.

Another important lesson is that automated machine learning workflows introduce new layers of ethical complexity. AutoML environments enable broader participation in AI development, but they also risk abstracting away key decisions related to fairness, explainability, and data governance. In such systems, ethical safeguards must be embedded by design, rather than left to individual judgment or post-hoc review. This challenges traditional notions of responsibility and requires new methods for supporting ethical decision-making within the tool itself.

A third lesson is the difficulty of operationalising regulatory frameworks. While instruments such as the AI Act and the GDPR provide essential principles, applying them meaningfully at system level remains a complex task. Developers often lack the interpretive tools needed to connect legal obligations to concrete design decisions. Our work demonstrated the need for structured, use-case-aware methodologies that can bridge this policy-practice gap in a consistent and scalable way.

Finally, the collaborative nature of this work reinforced the importance of interdisciplinary and participatory approaches. The integration of legal, technical, ethical and domain-specific knowledge was crucial in identifying nuanced risks and developing actionable recommendations. This experience affirmed ALFIE's core hypothesis: that ethics in AI cannot be treated as an external layer but must be co-developed through shared dialogue and joint design.

6.2 Summary of insights

Building on the lessons outlined above, this deliverable has produced a number of concrete insights that are relevant both for the ongoing development of ALFIE and for broader efforts to operationalise ethics in AI systems.

First, the analysis confirmed that ethics in AI must begin with context. Across the three use cases examined, driver monitoring, corporate compliance, and digital accessibility, it became clear that ethical risks are different depending on data types, system purposes, user interactions and societal impact. While principles such as fairness, transparency and accountability are foundational, they do not apply uniformly. What fairness means in the context of visual monitoring differs significantly from what it means in language-based

compliance systems. This reinforces the need for use-case-driven approaches when developing ethical frameworks.

In addition, the work demonstrated that ethical reflection can be translated into concrete design considerations, even at an early stage. The risk assessments conducted for each use case revealed points of tension between functional goals and ethical concerns, such as balancing model accuracy with explainability, or expanding automation without undermining user agency. These tensions are not always resolvable in absolute terms, but recognising and documenting them is a necessary step toward more responsible system development.

Furthermore, the development of best practices showed that ethical and legal principles can inform system design, even before final technical specifications are in place. This deliverable sets the direction for how that integration will occur, by identifying areas where fairness aware modelling, human oversight, or transparency enhancing measures should be implemented. This forms the groundwork for the next phase of the project, where these considerations will be translated into operational features.

Finally, the deliverable reinforced the value of linking policy frameworks to technical development through intermediate layers of guidance. Documents like the AI Act and GDPR set expectations, but developers need support in understanding how to translate those expectations into technical choices. The current blueprint represents an important step: it contextualises legal and ethical obligations within specific application scenarios and begins to map those to design implications.

In summary, the insights gained through this work strengthen ALFIE's foundational claim that trustworthy AI cannot be achieved through either regulation or design alone. It requires tools, methods, and shared frameworks that support both technical and societal alignment. This deliverable provides an initial version of such a framework, grounded in real use cases and informed by ongoing regulatory evolution.

6.3 Next steps

This deliverable presents the first version of ALFIE's best ethical practices blueprint. In the next phase of the project, this work will be further refined through additional analysis, stakeholder input and alignment with the evolving European regulatory landscape.

The recommendations developed here serve as a guide for designing an ethical AI framework and seamlessly integrating EU policy recommendations into the AutoML platform. Special attention will be given to ensuring that these recommendations are accessible to end users and relevant to the technical functionalities being developed.

An updated version of the blueprint will be presented later in the project and will incorporate additional feedback from use cases, participatory dialogue activities and ongoing policy developments. This will ensure the framework remains relevant, applicable and aligned with ALFIE's aim to support trustworthy and human-centric AI.

References

1. A. Abhishek, L. Erickson, and T. Bandopadhyay, "Data and AI governance: Promoting equity, ethics, and fairness in large language models," *MIT Sci. Policy Rev.*, vol. 6, pp. 139–146, 2025, doi: 10.38105/spr.1sn574k4lp.
2. A. Mehraj *et al.*, "AI model cards: State of the art and the path to automated regulatory compliance checks," *Tampere Univ. Press (Trepo)*, 2025. [Online]. Available: https://researchportal.tuni.fi/en/publications/ai-model-cards-state-of-the-art-and-path-to-automated-use/.
3. B. D. Mittelstadt, P. Allo, M. Taddeo, S. Wachter, and L. Floridi, "The ethics of algorithms: Mapping the debate," *Big Data Soc.*, vol. 3, no. 2, 2016, doi: 10.1177/2053951716679679.
4. BNP Paribas, "Data & artificial intelligence: Mastering and implementing AI at the heart of the bank," 2025. [Online]. Available: https://group.bnpparibas/en/our-commitments/innovation/data-artificial-intelligence.
5. *Commission Decision of 24.1.2024 establishing the European AI Office*, OJ C, 2024. [Online]. Available: http://data.europa.eu/eli/dec/2024/1459/oj.
6. EIT Health, "Scaling AI in European healthcare: Success stories from the EIT Health accelerator," *EIT Health e.V.*, 2025. [Online]. Available: https://eithealth.eu/news-article/scaling-ai-in-european-healthcare/.
7. *EU AI Act (Article 10): Requirements for High-Quality Training, Validation, and Testing Datasets*, 2024. [Note: Specifically referring to "Ground Truth" or "Reference Data" terminology].
8. European Commission, "AI package: Supporting AI startups and SMEs to develop trustworthy AI," European Commission - Digital Strategy, 2025.
9. European Commission, "Communication from the Commission on boosting startups and innovation in trustworthy artificial intelligence," COM/2024/28 final, 2024. [Online]. Available: https://eur-lex.europa.eu/legal-content/EN/TXT/?uri=CELEX:52024DC0028.
10. European Commission, “Grant Agreement No. 101177912 - ALFIE,” European Commission, Brussels, Belgium, 2024.
11. High-Level Expert Group on Artificial Intelligence, "Ethics guidelines for trustworthy AI," European Commission, 2019. [Online]. Available: https://digital-strategy.ec.europa.eu/en/library/ethics-guidelines-trustworthy-ai.
12. J. Kleinberg, S. Mullainathan, and M. Raghavan, "A simple tactic that could help reduce bias in AI," *Harvard Business Review*, Nov. 19, 2020. [Online]. Available: https://hbr.org/2020/11/a-simple-tactic-that-could-help-reduce-bias-in-ai.
13. L. Floridi, J. Cowls, M. Beltrametti, *et al.*, "AI4People – An ethical framework for a good AI society: Opportunities, risks, principles, and recommendations," *Minds Mach.*, vol. 28, no. 4, pp. 689–707, 2018, doi: 10.1007/s11023-018-9482-5.
14. L. Short, "Ethics in AI: Why it matters," *Harvard Professional & Executive Development Blog*, 2025. [Online]. Available: https://professional.dce.harvard.edu/blog/ethics-in-ai-why-it-matters/.

15. M. Mitchell *et al.*, "Model cards for model reporting," in *Proc. Conf. Fairness Accountability Transp. (FAccT '19)*, 2019, pp. 220–229, doi: 10.1145/3287560.3287596.
16. N. T. Nikolinakos, "A European approach to excellence and trust: The 2020 White Paper on Artificial Intelligence," in *EU Policy and Legal Framework for Artificial Intelligence, Robotics and Related Technologies: The AI Act*, vol. 53, Springer, 2023, pp. 101–125, doi: 10.1007/978-3-031-27953-9_5.
17. Oxiway Limited, "MeVitae ethical talent screening: Anonymised recruitment in SAP SuccessFactors," *SAP Store*, 2025. [Online]. Available: https://www.sap.com/products/hcm/partners/oxiway-limited-mevitae-ethical-talent-screening.html.
18. *Regulation (EU) 2016/679 of the European Parliament and of the Council of 27 April 2016 on the protection of natural persons with regard to the processing of personal data (General Data Protection Regulation)*, OJ L, 2016. [Online]. Available: http://data.europa.eu/eli/reg/2016/679/oj.
19. *Regulation (EU) 2021/523 of the European Parliament and of the Council of 24 March 2021 establishing the InvestEU Programme*, OJ L 107/30, 2021. [Online]. Available: http://data.europa.eu/eli/reg/2021/523/oj.
20. *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, 2024. [Online]. Available: http://data.europa.eu/eli/reg/2024/1689/oj.
21. *Regulation (EU) 2024/1732 of the European Parliament and of the Council of 17 June 2024 amending Regulation (EU) 2021/1173*, OJ L, 2024. [Online]. Available: http://data.europa.eu/eli/reg/2024/1732/oj.
22. *Regulation (EU) 2024/795 of the European Parliament and of the Council of 29 February 2024 establishing the Strategic Technologies for Europe Platform (STEP)*, OJ L, 2024. [Online]. Available: http://data.europa.eu/eli/reg/2024/795/oj.
23. Siemens Healthineers, "From innovation to impact: How AI elevates patient care," Oct. 24, 2025. [Online]. Available: https://www.siemens-healthineers.com/perspectives/from-innovation-to-impact-AI-in-healthcare.
24. T. A. Kustitskaya, R. V. Esin, and M. V. Noskov, "Model drift in deployed machine learning models for predicting learning success," *Computers*, vol. 14, no. 9, p. 351, 2025, doi: 10.3390/computers14090351.
25. T. Gebru *et al.*, "Datasheets for datasets," *arXiv preprint arXiv:1803.09010*, 2018. [Online]. Available: https://arxiv.org/abs/1803.09010.
26. T. Hagendorff, "The ethics of AI ethics: An evaluation of guidelines," *Minds Mach.*, vol. 30, no. 1, pp. 99–120, 2020, doi: 10.1007/s11023-020-09517-8.
27. UNESCO, "Recommendation on the ethics of artificial intelligence," General Conference, 41st session, SHS/BIO/REC-AIETHICS/2021, Nov. 2021. [Online]. Available: https://unesdoc.unesco.org/ark:/48223/pf0000380455.
28. World Wide Web Consortium, "Web Content Accessibility Guidelines (WCAG) 2.2," 2024. [Online]. Available: Web Content Accessibility Guidelines (WCAG) 2.2 .

Appendices

Appendix A: Table of consolidated best practices

Category	Core Principle	Operational Actions
Human Oversight & Inclusive Design	Establishing human-centric control as a fundamental design requirement rather than a final audit.	▪ **Integrate** "human-in-the-loop" checkpoints throughout the system lifecycle. ▪ **Appoint** multi-disciplinary teams (legal, ethical, and sociological experts) to development cycles. ▪ **Facilitate** stakeholder engagement to identify practical risks to diverse user groups.
Transparency & Accountability	Treating transparency as a strategic asset to enable informed decision-making by users and regulators.	▪ **Standardise** documentation through Model Cards and Data Statements. ▪ **Implement** regular, independent third-party audits to verify ethical performance. ▪ **Disclose** technical limitations and intended use cases to build public trust.
Institutional Governance	Embedding ethical responsibility into the management structure to ensure collective accountability.	▪ **Empower** a formal AI Ethics Board to oversee high-risk use cases. ▪ **Ensure** executives and product managers share a stake in ethical outcomes. ▪ **Prioritise** viability based on value-alignment, regardless of technical feasibility.
Continuous Adaptation	Transitioning from a static compliance mindset to a model of long-term learning and monitoring.	▪ **Deploy** automated tools to monitor "model drift" and fairness benchmarks in real-time. ▪ **Participate** in broader community dialogues to exchange best practices. ▪ **Commit** to ongoing system updates based on evolving societal and legal expectations.