ASSESSMENT OF LEARNING TECHNOLOGIES AND FRAMEWORKS FOR INTELLIGENT AND ETHICAL AI

D4.5 Initial use case definition and AutoML validation

Deliverable No:	D4.5
Deliverable Name:	Initial use case definition and AutoML validation
Work Package:	WP4
Task:	T4.5, T4.6
Dissemination Level:	PU
Deliverable Type:	R
Lead Organization:	Bosch
Submission month:	M10
Date:	31 July 2025

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 Espana SLU** (Bosch), Spain **Eindhoven University of Technology (**TU/e), Netherlands **Diadikasia Business Consulting** (DBC), Greece

Document history

Version	Date	Beneficiary	Description
0.1	15.05.2025	CTL	Structure proposal and initial draft.
0.2	06.06.2025	BOSCH	Use Cases
0.3	23.06.2025	UAB	User Centred Design, Inclusive Design
0.4	24.06.2025	UAB	UAB use case description
0.5	26.06.2025	UAB	Types of users, user requirement extraction section
0.6	09.07.2025	BOSCH	First complete draft
0.7	10.07.2025	UAB	Draft revision
0.8	15.07.2025	CTL	Reviewed the document and proposed amendments
0.9	16.07.2025	UAB	Amended Sections
1.0	17.06.2025	BOSCH	Revised document according to proposed amendments
1.1	20.06.2025	UoB	Revision
1.2	21.06.2025	BOSCH	Revised document according to proposed amendments
1.3	25.06.2025	CTL	Final version

Authors

Partner	Name(s)
Bosch	Leonardo Martín Abad, Juan A. Recio García
UAB	Pilar Orero, Sarah McDonagh, Chiara Gunella
CTL	Zenon Lamprou, Maria E. Skarkala

Contributors

Partner	Contribution type	Name
UoB	Review	Nikolaos Polatidis
EHU	Review	Yannis Korkontzelos

Glossary

Acronym	Definition
AI	Artificial Intelligence

ALFIE	Assessment of Learning technologies and Frameworks for Intelligent and Ethical AI
API	Application Programming Interface
ARIA	Accessible Rich Internet Applications
AutoML	Automated Machine Learning
BPS	Business Partner Screening
CAVs	Connected Autonomous Vehicles
CI/CD	Continuous Integration/Continuous Deployment
DBC	Diadikasia Business Consulting
DSM	Driver State Monitoring
ETD-Hub	EthiTech Dialogue Hub
EU	European Union
F1	F1-score (harmonic mean of precision and recall)
GDPR	General Data Protection Regulation
HTML	HyperText Markup Language
ID	Identifier
IRIS	Driver State Monitoring mobile application
ISO	International Organization for Standardization
IT	Information Technology
KPIs	Key Performance Indicators
mAP	mean Average Precision
ML	Machine Learning
MoSCoW	Must have, Should have, Could have, Won't have (prioritization method)

MVP	Minimum Viable Product
RBAC	Role-Based Access Control
R&D	Research and Development
RIA	Research and Innovation Actions
SDO	Sustainable Development Objectives
SUS	System Usability Scale
TUCs	Trial Use Cases
UCD	User Centred Design
UI	User Interface
URL	Uniform Resource Locator
WCAG	Web Content Accessibility Guidelines
XAI	Explainable Artificial Intelligence

Executive Summary

The ALFIE (Assessment of Learning technologies and Frameworks for Intelligent and Ethical AI) project represents a groundbreaking initiative under the EU Horizon Europe programme aimed at democratizing access to ethical AI development through Automated Machine Learning (AutoML) technologies. This deliverable presents the initial definition of three trial use cases, comprehensive user requirements analysis, and the foundational validation framework for the ALFIE platform.

The first pilot use case, led by Bosch, focuses on compliance alerts classification to automate the prioritization of Business Partner Screening alerts for regulatory compliance monitoring. The second pilot use case, coordinated by UAB, addresses website accessibility evaluation through the development of AI tools that automatically assess and improve web content accessibility for users with disabilities. The third pilot use case, managed by CTL, concentrates on ethical driver monitoring by creating unbiased emotion and drowsiness detection systems for autonomous vehicles. This use case addresses the critical challenge of mitigating demographic biases in facial recognition models to ensure equitable performance across diverse populations.

Through a comprehensive User-Centred Design approach, the project systematically identified and prioritized requirements across two distinct user categories. The analysis revealed that a significant portion of end-user requirements focused on accessibility features, bias detection capabilities, and multilingual support functionality. These requirements consistently focused on transparency, explainability, and ethical AI behaviour as fundamental expectations for platform acceptance. The examination of the designer user requirements showed many requirements emphasized on data management capabilities, comprehensive bias detection and mitigation tools, and flexible model deployment options. Designer user requirements encompassed technical specifications for code generation, local development environments, and seamless integration with existing organizational systems, reflecting the more technically sophisticated needs of users responsible for creating AI solutions rather than consuming them.

The ALFIE project implements a comprehensive three-phase validation framework designed to ensure systematic evaluation of platform capabilities, effectiveness, and trustworthiness across diverse application domains. This approach allows for continuous improvement while maintaining rigorous standards for technical performance, user acceptance, and ethical compliance. The initial validation phase focuses on establishing robust foundations through thorough data preparation, quality assessment, and baseline performance measurements using existing state-of-the-art solutions, demonstrating core AutoML capabilities. The second phase emphasizes real-world integration and user-centered evaluation through comprehensive prototype testing with diverse participant groups representing different backgrounds, technical expertise levels, and accessibility needs. Iterative refinement processes ensure that user feedback directly informs platform improvements, with particular attention to usability, accessibility, and trustworthiness factors. The final phase provides comprehensive evaluation under realistic operational conditions through full-scale demonstrations across all three trial scenarios. Performance validation employs both quantitative metrics aligned with established KPIs and qualitative assessments capturing user satisfaction, trust, and perceived value, while cross-scenario analysis identifies transferable insights for broader platform optimization.

Contents

Figures

Tables

1 Introduction

1.1 Overview

The ALFIE project represents a significant advancement in democratizing access to ethical Artificial Intelligence (AI) development through automated machine learning technologies. This deliverable presents the initial definition of three trial use cases that will serve as comprehensive validation scenarios for the ALFIE AutoML platform, demonstrating its capacity to generate trustworthy AI solutions across diverse application domains while maintaining strict adherence to ethical principles and accessibility standards.

The primary scope of this deliverable encompasses the detailed specification of trial use cases that collectively address critical challenges in contemporary AI development, including bias mitigation, accessibility enhancement, and regulatory compliance. Each use case has been carefully selected to represent distinct application domains that require AI capabilities while presenting unique ethical and technical challenges. Through systematic evaluation of these scenarios, the project aims to demonstrate the AutoML platform's effectiveness in enabling non-expert users to develop advanced AI solutions that meet high standards for fairness, transparency, and inclusivity.

The validation framework presented here establishes comprehensive methodologies for assessing platform performance across multiple dimensions, including technical accuracy, user experience, ethical compliance, and real-world impact. This multi-faceted approach ensures that the AutoML platform demonstrates not merely technical competence but genuine utility for diverse user communities, particularly those who have traditionally faced barriers to accessing advanced AI technologies.

1.2 Use cases overview

The three trial use cases represent selected scenarios that collectively validate the AutoML platform capabilities across critical areas of AI application. The **compliance alerts classification** scenario addresses industrial needs for automated decision support in regulatory environments, focusing on improving efficiency while maintaining accuracy in business partner screening processes. The **website accessibility evaluation** use case demonstrates the platform's capacity to generate tools that promote digital inclusion, specifically targeting the creation of accessible interfaces for users with diverse abilities. The **driver monitoring system** scenario tackles the development of unbiased emotion and drowsiness detection models, addressing critical safety applications while ensuring fairness across different demographic groups.

These use cases span different complexity levels and application domains, providing comprehensive coverage of the challenges that AutoML platforms must address to achieve widespread adoption. Each scenario incorporates specific requirements for bias detection and mitigation, accessibility support, and explainable AI functionality, ensuring that the validation process addresses the full spectrum of ethical AI development considerations.

Compliance Alerts Classification (Bosch)

Objective: Automate the prioritization of Business Partner Screening alerts for regulatory compliance monitoring.
Challenge: Manual processing of large volumes of compliance alerts creates bottlenecks and inconsistencies.
Target KPIs: >95% classification accuracy, >5% improvement in resolution time, >4.5/5 user satisfaction rating.

Website Accessibility Evaluation (UAB)

Objective: Develop AI tools to automatically assess and improve web content accessibility for users with disabilities.
Challenge: Ensuring digital inclusion through automated detection of accessibility barriers and generation of remediation recommendations.
Target KPIs: >10 diverse end-users engaged, >4.5/5 satisfaction across user groups, >90% compliance with EU regulatory framework.

Ethical Driver Monitoring (CTL)

Objective: Create unbiased emotion and drowsiness detection systems for autonomous vehicles.
Challenge: Mitigating demographic biases in facial recognition models to ensure equitable performance across diverse populations.
Target KPIs: >70% mean Average Precision, <5% performance deviation across demographic groups, >4.5/5 user satisfaction.

1.3 Relation to other tasks and deliverables

This deliverable is related to the following other ALFIE tasks and deliverables:

Receives inputs from:

Table 1. D4.5 Input from other tasks and deliverables

Deliverable	Due Date	Input for D4.5
D2.3	28 Feb 2025	Stakeholder engagement strategy and initial policy recommendations that inform the trial scenario design and stakeholder involvement approach.
D6.2	31 Mar 2025	Instructions from the Ethics Advisor regarding the experiments with participants and the handling of personal data.

Provides outputs to:

Table 2. D4.5 Output from other tasks and deliverables

Deliverable	Due Date	Output from D4.5
D4.1	30 Sep 2025	Use case requirements and validation criteria that inform the AutoML platform co-designed architecture and functional requirements specification
D4.2	31 Jan 2027	Validated use case scenarios and evaluation methodology that guide the operational prototype development and testing procedures
D4.3	30 Apr 2026	Initial validation results and user feedback that inform the first prototype development of AutoML platform and frontend interfaces
D4.4	31 Jul 2027	Comprehensive validation framework and KPIs that guide the development of the second prototype and integration testing procedures
D4.6	31 May 2026	Trial use case evaluation results and new requirements for AutoML platform refinement and improvement
D4.7	30 Sep 2027	Final trial case evaluation framework and lessons learned that contribute to the comprehensive platform assessment and future recommendations
D3.1	31 Mar 2026	Data requirements and semantic knowledge specifications derived from use case analysis that inform the AutoML data warehouse development
D3.3	31 Mar 2026	Multilingual interaction requirements and user interface specifications that guide AI tools development for natural language processing
D3.5	31 Mar 2026	Explainable AI requirements and transparency specifications that inform XAI feature development and trustworthiness implementation

1.4 Structure of the deliverable

The deliverable is organized to provide coverage of both theoretical foundations and practical implementation strategies regarding the trial use cases. The document begins with an overview of user profiles within the ALFIE ecosystem, distinguishing between designer users who will create AI models and end users who will interact with the resulting systems. This foundation enables precise specification of requirements and evaluation criteria that address the needs of both user categories.

The core content presents detailed specifications for each of the three trial use cases, including descriptions of objectives, challenges, KPIs, demonstration approaches, and evaluation methodologies. Each use case section provides sufficient detail to guide implementation activities while maintaining focus on the validation objectives that will demonstrate platform effectiveness.

The user requirements sections present the results of stakeholder engagement activities, documenting both end user and designer user requirements that have been systematically gathered, categorized, and prioritized using an established methodology. The validation framework section synthesizes these inputs into evaluation criteria that will assess platform performance across technical, ethical, and usability dimensions.

The document concludes with detailed specifications for the first stakeholder workshop and the overall AutoML validation framework, providing concrete methodologies for ongoing stakeholder engagement and systematic platform evaluation throughout the project lifecycle.

2 User profiles in ALFIE

In the ALFIE project, we established two categories of users: the **designer user** and the **end user**, summarised in Table 3. The distinction between designer users and end users reflects the different roles these stakeholders play within the ALFIE ecosystem across diverse application domains.

Designer users represent professionals who will utilize the AutoML platform to create, configure, and deploy AI models and tools for specific organizational or societal purposes. These users possess varying levels of technical expertise but share the common goal of leveraging AutoML capabilities to develop solutions that address specific challenges across different domains.

End users represent individuals who will directly interact with AI models and tools created through the AutoML platform, relying on these solutions to meet their specific functional needs without necessarily being involved in the technical development process. Across the three trial use cases described in Section 3, end users include compliance officers in industrial settings, diverse stakeholders requiring accessible AI interfaces including vulnerable populations, and autonomous vehicle operators who benefit from unbiased emotion and drowsiness detection systems.

2.1 Designer users

A designer user is a professional involved in the creation and development of products, systems, or models. In the ALFIE project, a designer user refers to the user who wants to use the AutoML platform to create an ethical tool or model for a specific purpose. The designer user may be a designer, developer, or experienced user seeking to build a tool or model tailored to their specific needs or a particular purpose.

Example:

Name: Alex Moore.

Role: Web Designer and Accessibility Advocate.

Goal: To ensure that websites comply with accessibility standards, particularly for blind users, by developing a model that evaluates website accessibility.

Background

Alex is a professional web designer with a disability who strongly values inclusive design. Alex is blind and will interact with the AutoML platform using speech-to-text, a screen reader, and a keyboard. Over the years, they've noticed a gap in tools that provide a comprehensive analysis of how accessible a website is to people with disabilities. Many existing tools highlight issues without offering actionable insights. Alex envisions a model that not only detects accessibility issues but also recommends fixes, empowering designers and developers with varying levels of technical expertise to create more accessible websites.

2.2 End users

An end user is an individual who directly engages with a product, system, or service, relying on its functionality to meet their needs. In the ALFIE project, the end user is the person who interacts with the model or tool created by the designer user through the AutoML platform. The end user may be a customer, client, or member of the general public who relies on the model or tool for its intended functionality, benefiting from the AI solution without being directly involved in its development.

Example:

Name: Sam Fisher**.**

Role: Business owner.

Goal: Check the accessibility of his website for blind users.

Background

Sam Fisher, a business owner, wants to ensure his company's website is accessible to blind users. To achieve this, he uses Alex's accessibility checker, an AI-powered tool developed on the AutoML platform. He begins by inputting his website's URL or uploading relevant content, allowing the tool to scan for potential accessibility barriers such as missing alt text, poor colour contrast, and inadequate keyboard navigation. After the analysis, Sam receives a detailed report highlighting issues and providing recommendations for improvement. He then collaborates with his web development team to implement the suggested changes and re-tests the website to verify compliance with accessibility standards. As a result, Sam enhances his website's accessibility, creating a more inclusive experience for blind users while ensuring compliance with accessibility regulations.

Table 3. Types of Users in the ALFIE Project

User Type	**Name**	**Role**	**Goal**	**Background**
Designer User	Alex Moore	Web Designer and Accessibility Advocate	Develop a model to evaluate website accessibility for blind users.	Alex, a blind web designer, uses AutoML with speech-to-text and a screen reader. They aim to create a tool that detects and recommends fixes for accessibility issues, helping designers build more inclusive websites.

End User	Sam Fisher	Business Owner	Ensure his website is accessible to blind users.	Sam uses Alex's AI-powered accessibility checker to scan his website, receive recommendations, and improve accessibility compliance.

3 Trial Use Cases (TUCs)

This section describes the three Trial Use Cases designed to demonstrate the capabilities of the AutoML platform.

3.1 Resolution of IT incidents and the optimization of manufacturing processes

3.1.1 Description

The Bosch use case represents a relevant application of AutoML technology in the domain of compliance monitoring and industrial process optimization. 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 prioritization 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 prioritized 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.

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 categorized and prioritized 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 prioritization decisions.

This implementation builds strategically upon Bosch's existing participation in the EU Horizon 2020 SHOP4CF[1] (Smart Human Oriented Platform for Connected Factories) program, where the company has been exploring the optimal integration of automation technologies with human expertise in manufacturing environments. This use case extends this proven philosophy to the compliance domain, demonstrating how AutoML can serve as an intelligent assistant that enhances human decision-making capabilities rather than attempting to replace the critical thinking and contextual understanding that compliance professionals bring to their work.

[1] https://shop4cf.eu

3.1.2 Objectives, needs and challenges

Objectives

The primary objective of this use case centers on demonstrating how ALFIE's AutoML platform can enhance the capabilities of compliance officers and industrial process managers to optimize machine learning (ML) applications within their specific operational contexts. This enhancement focuses on intelligent augmentation of human expertise, ensuring that the integration of AI tools results in measurably improved outcomes while maintaining the essential human oversight that compliance processes require.

A fundamental objective involves ensuring that the deployment of AutoML for resolving IT incidents and optimizing manufacturing-related compliance processes maintains strict alignment with relevant **privacy regulations and data protection laws**. Given the sensitive nature of compliance data and the potential implications of third-party relationship assessments, the system must demonstrate robust privacy protection mechanisms while delivering enhanced analytical capabilities. This objective requires careful consideration of data handling protocols, access controls, and audit trails that meet both internal Bosch standards and external regulatory requirements.

The techniques for the **identification and mitigation of biases** present in ML applications generated by ALFIE represent another critical objective. Compliance decisions must be fair and unbiased, free from any discriminatory patterns that might unfairly impact certain types of third-party relationships or geographical regions. The AutoML system must include bias detection and mitigation mechanisms that can identify potential sources of unfairness in alert prioritization and provide corrective measures to ensure equitable treatment of all compliance cases.

Efficiency and effectiveness improvements in IT incident resolution constitute core objectives of the use case. The system aims to measurably reduce the time required to process compliance alerts from initial generation through final resolution, while improving the accuracy of prioritization decisions. This dual focus on speed and accuracy reflects that compliance processes must balance the need for rapid response with the requirement for thorough and accurate assessment of potential violations.

Key needs

The key needs driving this use case include the **development of an automated prioritization system** that can significantly reduce the manual review overhead currently required for BPS alerts. Compliance officers need access to intelligent recommendations that can help them focus their attention on the most critical alerts while ensuring that lower-priority items are not overlooked entirely. This need extends beyond simple automation to include decision support capabilities that can explain the reasoning behind prioritization recommendations.

Integration capabilities with existing Bosch IT infrastructure and compliance workflows represent essential needs that will determine the practical success of the AutoML implementation. The solution must connect with current systems and processes without disrupting established workflows or requiring extensive retraining of compliance staff. This integration needs encompasses both technical compatibility and organizational change management considerations.

The development of **user-friendly interfaces** that enable non-ML experts to interact effectively with AI-generated models addresses the critical need for democratizing access to analytical capabilities. Compliance officers should be able to understand, interpret, and act upon AutoML recommendations without requiring deep technical knowledge of ML algorithms or data science methodologies.

Critical challenges

Critical challenges facing the implementation include **ensuring data quality and consistency within the historic alert database** that will serve as the foundation for ML model training. The effectiveness of any machine learning solution depends fundamentally on the quality and representativeness of the training data, making it essential to assess and potentially improve the existing data repository before model development begins.

Regulatory compliance presents an ongoing challenge that extends beyond the technical implementation to encompass the evolving landscape of AI governance and compliance requirements. The AutoML solution must be designed with sufficient flexibility to adapt to new regulatory requirements, for example, any new national or European regulation, while maintaining consistent performance and reliability.

Change management challenges center on the need to manage the **transition from manual to AI-assisted compliance processes** while maintaining staff confidence and preserving institutional knowledge. This challenge **requires attention to training, communication, and gradual implementation strategies** that allow compliance officers to develop trust in the new AI tools.

Model interpretability represents a fundamental challenge in ensuring that ML-generated prioritization decisions can be adequately explained and justified to internal stakeholders, external auditors, and regulatory authorities. The AutoML platform must provide clear, understandable explanations for its recommendations that satisfy both technical and business requirements for transparency and accountability.

In summary, the critical challenges for the AutoML platform encompass the checking of the input data to find anomalous values, the incorporation of restrictions from existing regulations into the generated model, ensuring the privacy of institutional data, and providing explainability capabilities to enhance the users' confidence in the platform.

3.1.3 KPIs progress

The measurement of success for this use case relies on a set of Key Performance Indicators (KPIs) that capture both the technical effectiveness of the AutoML implementation and the practical acceptance and integration of the solution within Bosch's operational environment. These KPIs have been specifically chosen to provide measurable evidence of improvement while ensuring alignment with broader organizational objectives.

The incident resolution **time improvement target of greater than 5%** represents an operational enhancement that will be measured through before-and-after analysis of the complete compliance alert lifecycle. This measurement encompasses the time from initial alert generation through final resolution decision, including all intermediate steps such as data gathering, analysis, consultation, and documentation. As this is a semi-automatic process involving human interaction, improvements must be rooted in the instantaneous classification of alerts and the provision of explainability capabilities to speed up human validation. This 5%

improvement target, while seemingly modest, represents substantial time savings when applied across the volume of alerts processed annually by Bosch compliance teams.

Incident identification accuracy, measured through a precision for "Relevant" class **greater than 95%**, ensures that the AutoML system demonstrates genuine analytical capability rather than simply processing speed improvements. This metric specifically measures the system's ability to correctly identify alerts that truly require further investigation, distinguishing them from alerts that can be resolved through standard procedures.

Employee satisfaction measurement through **Likert scale assessment targeting a rating greater than 4.5 out of 5** addresses the critical human factors that determine the long-term success of AI implementation in industrial environments. This measurement captures satisfaction with the AI-assisted incident resolution process, including factors such as ease of use, reliability of recommendations, clarity of explanations, and overall impact on job satisfaction.

The **tool integration success metric**, requiring that AI tools be effectively **used by more than 80% of participating compliance officers**, ensures that the AutoML platform demonstrates practical utility across the breadth of compliance activities rather than succeeding only in narrow, specialized applications. This metric acknowledges that true democratization of AI capabilities requires successful adoption across diverse user groups.

3.1.4 Demonstration approach

Demonstration will follow a structured approach that showcases the AutoML platform capabilities in a real-world industrial compliance environment:

Phase 1: Data preparation and model development (M10-M15)

The initial phase of the demonstration focuses on data preparation and model development activities that establish the foundation for all subsequent validation activities. Historic BPS alert data will be collected and processed through preprocessing procedures that ensure data quality while maintaining strict privacy compliance protocols. This phase involves close collaboration with Bosch compliance teams to understand the nuances of alert categorization and the contextual factors that influence prioritization decisions in practice. The **integration into the ALFIE's AutoML data warehouse (AutoDW)** provides a secure, scalable environment for managing the compliance dataset while enabling sophisticated analytical capabilities and will serve to evaluate this complementary component of the ALFIE project.

The **automated model selection and training process** represents a key demonstration element that showcases the AutoML platform's ability to identify optimal classification algorithms for alert prioritization without requiring deep machine learning expertise from users. This demonstration will highlight how the platform evaluates multiple algorithmic approaches, automatically tunes hyperparameters, and selects models based on cross-validation procedures that ensure robust performance.

Phase 2: Prototype integration and testing (M16-M22)

The prototype integration and testing phase provides opportunities to demonstrate the practical integration of trained models with existing Bosch compliance workflows. This phase includes the **development and demonstration of user-friendly interfaces** that present AI-generated prioritization recommendations in formats that compliance officers find intuitive and actionable. The pilot testing with selected compliance officers provides authentic user feedback that validates both the technical functionality and the practical utility of the AutoML-enhanced system.

Phase 3: Full-scale demonstration and evaluation (M23-M30)

The full-scale demonstration and evaluation phase will provide evidence of the AutoML platform's capabilities under realistic operational conditions. This phase includes real-time demonstration of alert processing with AI-assisted prioritization, showing how incoming BPS alerts are automatically analyzed, prioritized, and presented to compliance officers with supporting explanations and confidence indicators. The demonstration will include examples of different alert types and complexity levels, showing how the system adapts its recommendations based on specific circumstances and available information. This latest stage will serve to perform a comprehensive evaluation against established KPIs and generate documentation of lessons learned and best practices.

3.1.5 Evaluation methodology

The evaluation methodology is designed as a framework that assesses both technical performance and organizational impact while ensuring that the evaluation process contributes valuable insights for future AutoML implementations in industrial compliance environments. It is organized according to several complementary dimensions:

Quantitative evaluation framework

The quantitative evaluation begins with a baseline establishment that documents current manual processing approaches across multiple dimensions, including processing times, accuracy rates, consistency measures, and resource allocation patterns for BPS alert handling.

Next, controlled comparison methodologies, including **A/B testing protocols**, will provide mechanisms for measuring improvement metrics while controlling for external factors that might influence compliance processing performance. These controlled comparisons involve parallel processing of similar alerts through both manual and AI-assisted approaches, with careful attention to ensuring that comparison groups are genuinely equivalent in terms of alert complexity, compliance officer experience, and operational context.

This action encompasses classification performance assessment using comprehensive metrics, including precision, recall, F1-score, and mean average. Additionally, the AutoML platform applies fairness metrics to identify and quantify any discriminatory patterns in alert prioritization, ensuring that the AI system maintains equitable treatment across different categories of third-party relationships, geographical regions, or other potentially sensitive classifications.

Qualitative evaluation framework

The qualitative evaluation framework provides insights into user experience and organizational impact that cannot be captured through purely technical metrics. Structured interview protocols with compliance officers provide understanding of how AI assistance influences their work processes, decision-making approaches, and professional satisfaction. These interviews explore both positive impacts and challenges encountered during the implementation, providing valuable feedback for system refinement and future development priorities.

Comprehensive usability testing employs standardized assessment protocols that evaluate interface effectiveness, information presentation quality, and overall user experience. These usability assessments include task completion rate measurements, error rate analysis, and subjective satisfaction assessments, which provide a detailed understanding of how well the AutoML platform succeeds in democratizing access to advanced analytical capabilities.

Stakeholder impact analysis

Focus group discussions provide opportunities to capture collective insights and identify common themes in user experiences that might not emerge through individual interviews or quantitative measurements. These discussions explore the broader organizational impact of AI implementation, including effects on team dynamics, knowledge sharing, and professional development opportunities.

Stakeholder feedback integration encompasses the systematic collection and analysis of perspectives from management teams, IT departments, and compliance organizations to ensure that the evaluation captures impacts across the full range of organizational constituencies affected by the AutoML implementation.

The designed timeline for the evaluation methodology incorporates both interim and final assessment phases that provide timely feedback for system refinement while ensuring comprehensive final evaluation against all established success criteria:

- **Interim evaluation** (M20): Mid-implementation assessment focusing on technical performance and early user feedback.
- **Final evaluation** (M30): Comprehensive assessment against all KPIs with detailed analysis of lessons learned.

3.2 Smart website accessibility analyzer

3.2.1 Description

This use case involves using the ALFIE AutoML platform to process and analyse web content for compliance with Web Content Accessibility Guidelines[2] **(WCAG)**, specifically targeting improvements in accessibility 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.

3.2.2 Objectives, needs and challenges

Objectives

The primary objective of this use case is to streamline the evaluation of web content accessibility, focusing particularly on the experiences of blind users. By leveraging the ALFIE AutoML platform, the goal is to automate the identification of WCAG violations that are often overlooked or time-intensive to detect manually. These violations include issues such as missing alternative text for images, absence of ARIA (Accessible Rich Internet Applications) roles, poor navigability using keyboards, and inadequate colour contrast. The system is designed not only to flag these problems, but also to guide developers toward creating content

[2] https://www.w3.org/TR/WCAG21/

that is more accessible and usable for those relying on screen readers and other assistive technologies. Ultimately, the tool supports a shift toward accessibility-conscious design becoming a built-in part of web development practices rather than an afterthought.

Key needs

This use case arises from a need for an efficient and scalable solution to test digital content for accessibility barriers. Development teams are often under time constraints and may lack specialized knowledge in accessibility standards. A tool that can automatically scan HTML files and return structured feedback allows them to focus their efforts where it matters most. There is a strong need for the system to produce results that are both technically precise and understandable to developers, designers, and accessibility reviewers. In addition, the ability to integrate this feedback into existing workflows, such as CI/CD pipelines or development environments, would support earlier detection and remediation of issues, reducing the cost and effort of retroactive fixes. A further need lies in helping teams prioritise fixes based on severity and user impact, allowing them to improve accessibility incrementally but meaningfully.

Critical challenges

Critical challenges facing the implementation include **ensuring data quality** that will serve as the foundation for ML model training. The effectiveness of any machine learning solution depends fundamentally on the quality and representativeness of the training data, making it essential to assess and potentially improve the existing data repository before model development begins.

Regulatory compliance presents an ongoing challenge that extends beyond the technical implementation to encompass the evolving landscape of AI governance and compliance requirements. It also has to fulfil other compliance, such as WCAG and General Data Protection Regulation (GDPR), when dealing with personal data of vulnerable groups. The AutoML solution must be designed with sufficient flexibility to adapt to new regulatory requirements while maintaining consistent performance and reliability.

Model interpretability represents a fundamental challenge in ensuring that ML-generated prioritization decisions can be adequately explained and justified to internal stakeholders, end users such as vulnerable groups, external auditors, and regulatory authorities. The AutoML platform must provide clear, understandable explanations for its recommendations that satisfy both technical and business requirements for transparency and accountability.

3.2.3 KPIs progress

To ensure the AutoML platform solutions are accessible, inclusive, and effective across diverse user groups, a set of KPIs has been defined. These KPIs help monitor progress in user engagement, satisfaction, platform growth, ethical compliance, tool adoption, and training outcomes. The targets are designed to reflect measurable improvements in usability and inclusion, especially for users with special skills and needs.

1. User engagement and inclusion

- Measure the level of engagement and interaction with the AI platform among users with special skills, targeting participation from more than 10 end-users in this category.
- Measure the diversity of users engaging with the platform to ensure representation across various special skills and domains, aiming to test users from at least 2 to 3 distinct skill groups.

2. User satisfaction and experience

- Measure the satisfaction of diverse user groups, including elders, disabled individuals, and children, with the generated interfaces. Target a satisfaction rating of greater than 4.5 on a Likert scale.
- Evaluate the ease of use and accessibility of the platform based on user feedback and usability testing, aiming for an average score of more than 4 out of 5 on the Likert scale.

3. Platform impact and growth

- Evaluate the impact of accessibility features on usability and inclusivity, aiming for a yearly increase of more than 10% in the number of people using the AutoML platform from the previous year.
- Gauge the scalability of the designed interfaces for deployment in different applications and industries, targeting a 10% increase in industry adoption each year.

4. Ethical compliance

- Verify adherence to ethical guidelines, ensuring transparency, fairness, and privacy in the generated interface designs. Target at least 90% compliance of the generated model with all the EU regulatory framework, with updates implemented in less than one month after regulatory changes are accepted.

5. Tool integration and use

- Assess the successful integration of AI tools into the workflows of users with special skills, aiming for at least 5 AI tools being actively used by more than 80% of these users.

6. Training and proficiency

- Monitor the effectiveness of training modules by evaluating user proficiency and skill enhancement. Target that more than 70% of users are able to use the platform effectively to perform their tasks.

3.2.4 Demonstration approach

Phase 1: Data preparation and model development (M10–M15)

The first phase of the demonstration focuses on the preparation of accessibility evaluation data and the development of models that enable automated detection and validation of web accessibility issues. Sample web content, including HTML pages and interface components, will be collected and processed using structured preprocessing routines that ensure data consistency while maintaining user privacy and regulatory compliance. This phase involves collaboration with accessibility specialists and interface developers to accurately label typical violations and understand design decisions that impact usability. The processed dataset will be integrated into the ALFIE AutoML data environment, which provides the infrastructure to support secure storage, efficient processing, and scalable analysis of accessibility-related features.

The demonstration will highlight the platform's core functionality of automated model selection and training. Through this process, the AutoML platform will explore a range of classification and evaluation models to detect issues such as missing alt text, low color contrast, structural inconsistencies, and lack of ARIA attributes. The platform will automatically tune relevant parameters, select the best-performing models based on cross-validation results, and ensure that each model is capable of generalizing across different interface types. This step illustrates how the platform enables non-expert users to carry out sophisticated accessibility assessments using AI-powered tools.

Phase 2: Prototype integration and testing (M16–M22)

During the second phase, the demonstration will focus on the integration of the trained models into a user-facing prototype that supports accessibility assessment workflows. The interface will allow users to upload web content and receive annotated accessibility evaluations along with structured recommendations for remediation. The system will present issues in a way that is understandable to developers, designers, and reviewers, including visual indicators of problem areas and suggested fixes.

Pilot testing will involve selected users who will interact with the prototype in a controlled setting. Their feedback will help validate both the accuracy of the platform's detection capabilities and the clarity of the guidance provided. This phase emphasizes the usability of the platform and its ability to produce results that are not only technically correct but also meaningful and actionable in practice. Feedback collected during this period will also inform improvements to the interface and help refine user support features.

Phase 3: Full-scale demonstration and evaluation (M23–M30)

The final phase will involve a full-scale demonstration of the AutoML platform's capabilities under realistic conditions, simulating how it would be used in actual development or audit workflows. This will include live assessments of diverse web pages, with real-time feedback on accessibility status and compatibility with screen readers. The demonstration will show how the platform identifies, categorizes, and prioritizes accessibility issues, providing contextual explanations and confidence scores for each assessment.

Different types of content and varying levels of complexity will be used to showcase the system's adaptability and its handling of nuanced design challenges. This phase will also include a formal evaluation based on previously defined KPIs, covering metrics such as issue detection accuracy, user satisfaction, and overall improvement in interface accessibility. The process will be documented in detail, with insights, lessons learned, and best practices compiled to guide further development and potential scaling across additional use cases.

3.2.5 Evaluation methodology

The evaluation of the use case will follow a structured approach that is fully aligned with the KPIs previously presented. It will focus on how the AutoML platform performs in terms of accessibility, usability, adoption, and compliance. The process will involve **more than 10 users**, selected to reflect a range of needs and user profiles, including elders, children, and individuals with disabilities. Evaluation activities will combine user testing, system tracking, surveys, and compliance checks.

1. Engagement and interaction

A minimum of 10 users will interact with the use case over the course of the evaluation. These users will engage with various parts of the system, such as uploading content, reviewing accessibility feedback, and applying suggested fixes. Their usage patterns will be monitored to verify that engagement is active and not limited to a single action or session.

How it will be measured: System usage logs (e.g., number of sessions, time spent, features used) and participation records

KPI addressed: More than 10 end-users actively engaging with the platform

2. User satisfaction

A diverse group of users, including elders, disabled individuals, and children, will test the use case during the evaluation. After using the platform, each participant will complete a satisfaction questionnaire using a 5-point Likert scale, where:

- 1 = Very dissatisfied
- 5 = Very satisfied

The questionnaire will cover aspects such as clarity of information, ease of navigation, comfort of use, and overall experience with the generated interfaces. The responses will be averaged to assess overall satisfaction across all groups.

How it will be measured: Post-use satisfaction questionnaire (5-point Likert scale)

KPI addressed: Average score greater than 4.5 across all user groups

3. Ease of use and accessibility

Users will perform key tasks on the use case, such as uploading content, receiving automated accessibility feedback, and interpreting the results. After completing these tasks, each user will fill out the System Usability Scale (SUS), a standardized questionnaire consisting of 10 statements, each rated on a 1 to 5 scale, where:

- 1 = Strongly disagree
- 5 = Strongly agree

The SUS responses will be used to calculate a mean usability score for each user and for the group overall. This will give a clear indication of how usable and accessible the platform feels from the user's perspective. The SUS is widely used in English-language evaluations and will be administered in English.

How it will be measured: Task-based usability testing and SUS questionnaire (1–5 scale)

KPI addressed: Usability rating above 4; SUS score target ≥ 75 (converted from mean SUS responses)

4. Impact on platform adoption

To evaluate whether the use case's accessibility features encourage broader adoption, usage will be tracked over time, comparing initial and follow-up engagement data. For example, if 50 users are registered in the first year, at least 55 should be active in the second year to meet the growth target.

How it will be measured: Growth in platform usage and registration records

KPI addressed: Over 10% increase in users each year

5. Ethical and regulatory compliance

The use case will be reviewed against current EU regulations on accessibility, data protection (e.g. GDPR), and ethical AI design. This includes checking if the system ensures transparency in its recommendations and protects user data appropriately. Any legal or policy updates will be tracked to ensure timely compliance.

How it will be measured: Internal compliance audits and documented update logs

KPI addressed: At least 90% compliance and updates implemented within 1 month of regulation changes

6. Scalability

The evaluation will track how the use case is used in different industries or sectors, such as education, public services, and private web development. The goal is to observe and document a growing presence in at least one new domain per year.

How it will be measured: Deployment tracking reports and usage breakdown by sector

KPI addressed: 10% annual increase in industry uptake

7. Integration of AI tools

The platform includes multiple AI-powered tools (e.g. image analysis, HTML structure checking, ARIA label detection). Usage of these tools will be logged. The evaluation will confirm that at least five distinct tools are used by the majority of users during the evaluation period.

How it will be measured: Tool activation logs and frequency of use reports

KPI addressed: At least 5 tools used by more than 80% of users

8. Training and proficiency

All users will receive a brief onboarding session or tutorial before beginning tasks. They will then complete a series of tasks on their own. Their ability to complete these tasks without further assistance will be assessed. For example, users might be asked to upload a file, interpret accessibility issues, and apply corrections independently.

How it will be measured: Pre- and post-training task completion rates and user self-assessment feedback

KPI addressed: More than 70% of users complete tasks without help after training

9. User diversity

The participant group will be intentionally selected to include individuals with varied needs and roles. For instance, the evaluation will include persons with cognitive accessibility needs, older adults who may have reduced digital literacy, and users who rely on assistive technologies. Their involvement ensures that the platform is assessed under different use conditions.

How it will be measured: Review of user profiles and group composition records

KPI addressed: At least 2–3 different user groups represented.

All these elements just described will be brought together in a structured user journey questionnaire, allowing for a unified assessment of how users interact with the platform, how accessible and usable it is in practice, and how well it meets the defined KPIs across different user groups.

3.3 Ethical AI tools for autonomous vehicles to ensure consistency across different drivers

3.3.1 Description

In the ongoing pursuit of enhanced road safety, Driver State Monitoring (DSM) technologies are undeniably pivotal, especially as the full realization of autonomous driving remains a distant aspiration [1]. These systems, designed to monitor a driver's condition and intervene when necessary, are crucial for mitigating risks associated with human error at the wheel.

A significant challenge confronting current DSM systems, however, lies in the inherent **limitations of their underlying facial landmark recognition models** [2]. These models, upon which drowsiness detection and other vital functions depend, frequently exhibit biases that compromise their accuracy when applied to individuals from diverse racial backgrounds or with varying facial characteristics. Such biases can manifest as a reduced ability to correctly identify key facial features or expressions across different demographics. This pervasive issue can lead to serious consequences, particularly in the context of drowsiness detection. If a DSM system fails to accurately detect signs of drowsiness in a driver due to model biases, it may produce "false negatives." These false negatives, where drowsiness is present but not identified, directly undermine the system's reliability and critically compromise its ability to prevent accidents.

To address these critical shortcomings and ensure the principles of fairness, inclusivity, and effectiveness are upheld, this specific use case, led by CTL, leverages the advanced capabilities of the ALFIE's AutoML platform. The primary objective is to **generate a diverse array of model variations specifically tailored for drowsiness detection**. These varied models will then undergo rigorous testing across a comprehensive range of scenarios. The ultimate goal of this extensive validation process is to unequivocally demonstrate a tangible increase in performance compared to state-of-the art drowsiness detection models [3][4][5]. This improved performance will be measured by the enhanced accuracy in identifying drowsiness in a broad spectrum of individuals, explicitly including those from different racial backgrounds and with diverse facial characteristics.

Furthermore, we intend to evaluate the efficacy of the AutoML generated models within a real-world setting, utilizing a mobile application called IRIS[3]. IRIS is a DSM application that employs a smartphone's front-facing camera to issue alerts to the user, thereby mitigating the risk of vehicular accidents caused by drowsiness. Our objective is to replace a third-party model presently integrated into the IRIS application with the newly developed models, thereby employing IRIS as a platform for real-life testing scenarios. By developing and deploying more inclusive and robust models, this initiative aims to significantly bolster the reliability and life-saving potential of DSM technologies.

3.3.2 Objectives, needs and challenges

Objectives

The primary objective of this use case is to enhance road safety by improving **the accuracy and fairness** of DSM technologies, particularly in drowsiness detection but more importantly **mitigating any biases and fairness issues** currently in the models used. This is achieved by leveraging the ALFIE platform to provide diverse model variations or provide suggestions that specifically address the biases present in existing facial landmark recognition models and facial classification models. The overarching goal is to achieve a tangible increase in performance compared to the state-of-the-art drowsiness detection models by ensuring enhanced accuracy across a broad spectrum of individuals, including individuals from multiple racial backgrounds and a variety of facial features. Ultimately, this use case aims to significantly bolster the reliability and life-saving potential of drowsiness detection systems.

Key needs

The key needs driving this use case are centered around overcoming the limitations of current DSM systems. One of the main needs is **bias mitigation in facial landmark recognition models** and facial classification models is crucial to ensure accurate drowsiness detection for all drivers, regardless of their racial background or facial characteristics. This also involves the generation of diverse model variations to specifically tailored models for improved drowsiness detection that are robust and inclusive. Furthermore, another need is to perform rigorous

[3] https://iris.catalink.eu/

testing across comprehensive scenarios is necessary to validate the effectiveness of the varied models in real-world conditions. Moreover, we need to increase accuracy in identifying drowsiness, thereby reducing false negatives and improving the reliability of DSM systems. Finally, the models should be able to run in real-time, providing predictions either with a fast connection to a server or through on-device processing and in the case of on-device processing, the models generated should also be able to run on mobile devices (Android).

Critical challenges

The main challenges regarding our use case are divided into two aspects. The first set of challenges is related to the AutoML platform. One of the main challenges is ensuring **high-quality and consistent data**, which is essential for effective ML model training, as it forms the foundation for reliable outcomes. Alongside this, adapting to evolving AI governance and regulatory compliance requirements is critical, all while maintaining the performance of the models. Moreover, incorporating the knowledge base from the ETD-Hub into the training process further enriches the model's capabilities and relevance. Additionally, providing **interpretability** and **clear explanations** of how the models function, demonstrating their fairness and regulatory adherence, enhances transparency, thereby fostering trust and deeper understanding.

The second set of challenges concerns the implementation of our use case. One of the main challenges is **addressing the inherent biases** in existing facial landmark recognition and classification models, a key step toward improving their accuracy across diverse demographic groups. To support this goal, it is essential to ensure that the training data is truly representative. This involves developing or utilizing publicly available datasets that encompass a wide range of facial characteristics and expressions. To format such a dataset creates its own challenges in recruiting individuals from different racial backgrounds and requires careful and strategic design of test scenarios to simulate real-world diversity effectively. Furthermore, another challenge regarding this use case is that once the models are developed, their performance must be rigorously validated across varied scenarios, including different driving conditions and driver states, to ensure robustness. Lastly, the final challenge identified is to build user acceptance and trust, as the success of such systems depends not only on technical performance but also on the confidence they inspire in drivers and stakeholders within the automotive industry.

3.3.3 KPIs progress

To effectively assess the success and impact of the CTL use case, a series of KPIs have been established. These KPIs are designed to measure various aspects of the developed models, including the accuracy in detecting drowsiness, the adaptability of AI models to cultural variations and diverse user demographics, their contribution to overall driving safety, and the satisfaction of users with the unbiased drowsiness detection system. By tracking these metrics, the use case aims to quantify the improvements achieved through the ALFIE platform and ensure the reliability and fairness of the DSM technologies.

Accuracy and consistency of emotion and drowsiness detection

A primary KPI is to ensure high accuracy and consistency in detecting drowsiness. The target is set at **>70% in terms of mean Average Precision (mAP)**. This metric indicates how well the AI models can correctly identify and localize instances of drowsiness across diverse drivers. Models that reach or exceed 70% mAP are typically considered state-of-the-art in object detection and classification benchmarks, suggesting they meet the performance standards expected in production or research environments [6]. The emphasis on "*drivers from different countries and demographic groups*" highlights the importance of the model's robustness and generalization capabilities, ensuring it performs reliably regardless of a driver's background.

Adaptability to cultural variations in expressions

Beyond raw accuracy, the system aims to demonstrate strong adaptability to cultural nuances in drowsiness indicators. The goal is to **achieve <5% deviation in mAP results amongst different target populations**. This KPI addresses a critical challenge in AI development which is preventing bias and ensuring equitable performance across diverse user groups. Cultural differences can significantly impact detection. By setting a low deviation target, the project aims to validate that the AI models are not culturally biased and can interpret a wide range of human expressions accurately.

Impact on overall driving safety and alertness

A crucial real-world impact KPI is to **reduce car accidents by 5%**. This KPI directly links the AI system's performance to an improvement in driving safety. The underlying assumption is that accurate and timely detection of emotion (e.g., anger, stress) and drowsiness will enable interventions or alerts that can prevent accidents by enhancing driver alertness and reducing risky behaviours. This metric serves as a high-level indicator of the system's effectiveness in achieving its ultimate purpose of creating safer driving environments.

User satisfaction with unbiased detection systems

Finally, user feedback is vital for assessing the system's success and ensuring its acceptance. The objective is to **achieve >4.5 on a Likert scale** regarding user satisfaction with the unbiased drowsiness detection systems. A mean score greater than 4.5 on a 5-point Likert scale typically reflects a high level of agreement or satisfaction, indicating that most respondents selected the most favourable responses [7] [8]. This qualitative measure captures the user experience and perception of the system's fairness and effectiveness. A high Likert score would indicate that drivers find the system helpful, reliable, and non-intrusive, and that they perceive it as truly unbiased, which is crucial for widespread adoption and trust in AI-driven safety features. This also implicitly validates the success of the adaptability objective from a user's perspective.

3.3.4 Demonstration approach

This section outlines a structured approach to demonstrate the successful integration and enhanced performance of drowsiness detection models generated by the ALFIE platform within a mobile application. The demonstration aims to showcase the real-time capabilities of the application and validate the success of the ALFIE’s AutoML produced models in mitigating biases and improving accuracy across diverse user demographics and characteristics.

Phase 1: Preparation and baseline establishment (M10–M15)

The initial phase focuses on setting up the environment and establishing a baseline for comparison between the current state-of-the art technologies used for drowsiness detection and the generated models and/or suggestions from the ALFIE platform. The first task of the demonstration phase will involve enhancing the IRIS mobile application, by incorporating a mechanism for loading custom models for drowsiness detection, facilitating the integration of models provided by AutoML. Furthermore, diverse visual and acoustic alert systems will be implemented to ensure accessibility. These enhancements will integrate into the app the current state-of-the art models for drowsiness detection and provide a mobile interface for real-time driver state monitoring, using the built-in camera, displaying indicators of alertness and drowsiness. Additionally, the application will allow the designer users to incorporate into the app various models for testing purposes. The data collection phase will focus on firstly identifying public datasets that contain different lighting conditions and head poses. Secondly, if the public datasets are not helpful for the newly generated models, we aim to generate synthetic data to help AutoML train unbiased models. Finally, if the two previous methods don't show significant improvement in terms of bias and fairness in drowsiness detection, we aim to collect a diverse dataset of driver facial characteristics and states (alert, mild drowsiness, moderate drowsiness), in a controlled environment such as workshops and/or recruiting beta testers of the application. For all the datasets we will try to include individuals from various racial backgrounds to underscore the limitations of current models. To assess each stage of the data collection phase, we will evaluate each sub-phase against our established baseline. This baseline will quantify the accuracy, precision, and bias within the identified state-of-the-art models from our initial research.

Phase 2: ALFIE model generation and integration (M16-M22)

This phase will demonstrate the core value proposition of the ALFIE platform by generating improved and bias-mitigated models.

Our core objective is to leverage the ALFIE platform for advancing drowsiness detection models. This will be achieved through a multi-faceted approach, beginning with the strategic utilization of diverse datasets. By feeding these varied datasets into the ALFIE system, we aim to train sophisticated new drowsiness detection models. A critical aspect of this initial phase will be the thorough identification and mitigation of potential biases embedded within the data and the resulting models and/or recommendations. This proactive approach ensures that the drowsiness detections models are not only accurate but also fair and equitable in their application.

Following this initial training and bias mitigation phase, we will transition into the model generation and selection stage. The ALFIE AutoML's capabilities will be fully showcased, by generating a multitude of model variations, each representing a unique permutation of algorithms and parameters. Simultaneously, ALFIE will provide a comprehensive suite of analytical tools designed to facilitate the selection process. These tools will enable us to rigorously evaluate each model variation against predefined fairness and accuracy metrics. This meticulous evaluation ensures that the chosen models are not only highly effective in detecting drowsiness but also demonstrate minimal bias, aligning with the ALFIE's commitment to responsible AI development.

Finally, the culmination of this process will be the seamless integration of ALFIE models into the IRIS application. This crucial step involves replacing the existing baseline drowsiness detection model within the mobile application with the newly generated and optimized models from the ALFIE AutoML. This integration will represent a significant upgrade, equipping the mobile application with a more accurate, reliable, and unbiased drowsiness detection capability, ultimately enhancing user safety and experience.

Phase 3: Live demonstration and performance validation (M23-M30)

The final phase will involve live demonstrations and rigorous evaluation to prove the success of the ALFIE's AutoML generated models. This will include a **real-time drowsiness detection demo**, where live demonstrations of the mobile application will be conducted with volunteers representing diverse facial characteristics. This will showcase the application's ability to accurately detect varying degrees of drowsiness in real-time.

For a **quantitative performance comparison**, a side-by-side comparison of the baseline model's performance versus the ALFIE's AutoML generated models will be presented using previously collected datasets. This will emphasize improvements in accuracy (mAP) by demonstrating a tangible increase in the mean average precision for drowsiness detection across diverse demographic or ethnic groups. Consistency (deviation in mAP) will be shown through a significant reduction in the deviation of mAP results among different target populations, proving enhanced fairness. Furthermore, false negative reduction will quantify the reduction in false negatives, particularly for underrepresented groups.

User experience feedback will be gathered qualitatively from participants regarding the application's performance and the perceived fairness and reliability of the drowsiness detection.

This multi-phase demonstration approach will provide compelling evidence of the ALFIE AutoML's ability to produce highly accurate, fair, and effective drowsiness detection models that are ready for real-world deployment in mobile applications, ultimately enhancing road safety.

3.3.5 Evaluation methodology

Phase 1: Offline baseline establishment and bias identification

The initial phase focuses on establishing a baseline from existing drowsiness detection models (i.e. Vision Transformer [4] and YoLo [4]). This involves a thorough offline evaluation of these pre-existing models using a comprehensive set of performance metrics. Standard measures such as accuracy, precision, and recall will be utilized to quantify the models' overall predictive power. However, a critical addition to this phase is the explicit identification and quantification of bias. This will be achieved by meticulously calculating the number of false negatives and false positives across different demographic groups, specifically focusing on people of varying racial backgrounds and those with diverse facial characteristics. This granular analysis is crucial for understanding whether the models exhibit preferential performance or systemic errors related to specific demographics. To facilitate this inclusivity, multiple datasets (at least three) will be integrated into the evaluation process. These datasets will be carefully selected and curated to ensure sufficient diversity and distinctiveness, thereby creating a robust testbed that allows for a comprehensive assessment of the models' fairness and generalizability. This phase is fundamental for setting a benchmark against which subsequent model improvements and the performance of AutoML generated models can be compared.

Phase 2: Real-time application testing and mobile compatibility assessment

Following the offline baseline establishment, the next phase involves loading these pre-existing models into IRIS application for real-time scenario testing. This transition from an offline, controlled environment to a dynamic, real-time setting is crucial for evaluating the models' practical viability. During this phase, specific metrics will be defined to measure the time taken for prediction, assessing the computational efficiency and responsiveness of the models within the application. Furthermore, a key aspect of this phase will be to determine if there is any drop in accuracy when converting the models to a mobile-compatible setting. This assessment is vital as the ultimate deployment of these models is envisioned within a mobile application, and any significant performance degradation due to optimization or conversion would need to be addressed. This phase provides valuable insights into the operational performance of the models in a simulated user environment, highlighting potential bottlenecks or areas for further optimization.

Phase 3: AutoML generated model evaluation (offline and online)

This phase mirrors the methodology of the previous phase but specifically focuses on the models generated from the AutoML platform. The AutoML generated models will first undergo a rigorous offline evaluation, like Phase 1, utilizing the same set of performance metrics and bias identification techniques. This allows for a direct comparison of their inherent accuracy and fairness characteristics against the established baselines. Subsequently, these AutoML models will be loaded into the IRIS application and tested in an online, real-time device setting, mirroring the evaluation conducted in Phase 2. The performance in terms of prediction time and accuracy in a mobile-compatible environment will be carefully measured. The core objective of this phase is to directly compare the performance of the AutoML generated models with the pre-existing models, both offline and online. This comparison will be instrumental in determining whether the AutoML platform indeed delivers on its promise of generating unbiased models without compromising overall accuracy. This assessment is crucial for validating the efficacy of the AutoML approach in producing high-performing and fair cognitive load and drowsiness detection solutions.

Phase 4: Live demonstration and user experience validation

The final and most crucial phase involves a live demonstration of the best-performing model. This demonstration will be conducted in the form of a workshop, involving at least 10 participants from as many as possible diverse racial backgrounds and with varying facial characteristics. This ensures that the model's effectiveness and fairness are rigorously tested in a real-world, inclusive setting. During the workshop, participants will be actively engaged and asked to answer a series of questions designed to gather comprehensive feedback on their experience. We will ask several questions to the users in the form of a questionnaire using the Likert scale with values ranging from 0 to 5. These questions will specifically aim to determine the perceived effectiveness of the model in detecting drowsiness, the level of comfort they felt while using the application, and their trust in the application with their safety. Furthermore, participants will be asked if they would be willing to integrate the application into their daily routine. This direct user feedback is invaluable, as it provides qualitative insights into the usability, reliability, and trustworthiness of the application, moving beyond purely quantitative performance metrics. This final phase serves as the ultimate validation step, ensuring that the chosen model not only performs accurately and fairly but also resonates positively with its intended users, ultimately contributing to its successful adoption and impact.

4 Quantification and assessment methodology

The ALFIE methodology is structured around principles of User-Centred Design (UCD) and Inclusive Design. This dual approach ensures that the needs, preferences, and constraints of diverse users are at the core of the design process for the AutoML platform. The following sections provide an overview of these two complementary methodologies, highlighting their distinct characteristics and shared commitment to prioritising end-user engagement. Section 5 will focus on their application in the ALFIE project.

4.1 Inclusive Design

Inclusive Design can be achieved by two different approaches: through the development of assistive technologies for a few or by “Universal Design" or "design for all". The Universal Design approach focuses on creating products and services that eliminate unnecessary barriers, ensuring usability for everyone [9]. This methodology aims to design products and services that are inherently adaptable to the needs of the widest possible user profiles, avoiding the reliance on costly, rigid modifications or "patchy" solutions targeted at specific groups (assistive technologies). A key aspect of Inclusive Design is enabling all individuals, especially those with disabilities, to use standard IT products and services. These solutions are typically more affordable, widely available, and frequently updated compared to assistive technologies - designed only for a few.

Universal Design fosters digital inclusion by considering the needs of all users from the design phase, thereby enhancing usability for a broad audience. For example, web accessibility guidelines, originally designed to remove barriers for users with disabilities, also benefit non-disabled individuals navigating challenging conditions, such as noisy or mobile environments. This demonstrates how Inclusive Design principles can simultaneously improve access for marginalised groups and create a better experience for the general population.

4.2 User Centred Design (UCD)

In ALFIE, we adopted a User Centred Design approach (UCD), a methodology emphasising the active involvement of end-users throughout the design and development process. Grounded in the principle that understanding users' needs, preferences, and limitations is essential to creating effective and inclusive solutions [10]. UCD prioritises iterative engagement with users to ensure that designs are both usable and meaningful. This approach integrates multidisciplinary expertise to foster creativity and collaboration, while ensuring the voice of the user remains central to the process. The UCD approach, also referred to as Human-Centered Design, aligns with the international standard ISO 9241-210:2019 [11], which outlines a structured four-phase process. These phases (Figure 1) include:

1. Understanding and defining the context of use;

2. Identifying user requirements;

3. Developing design solutions and

4. Evaluating the design against established requirements.

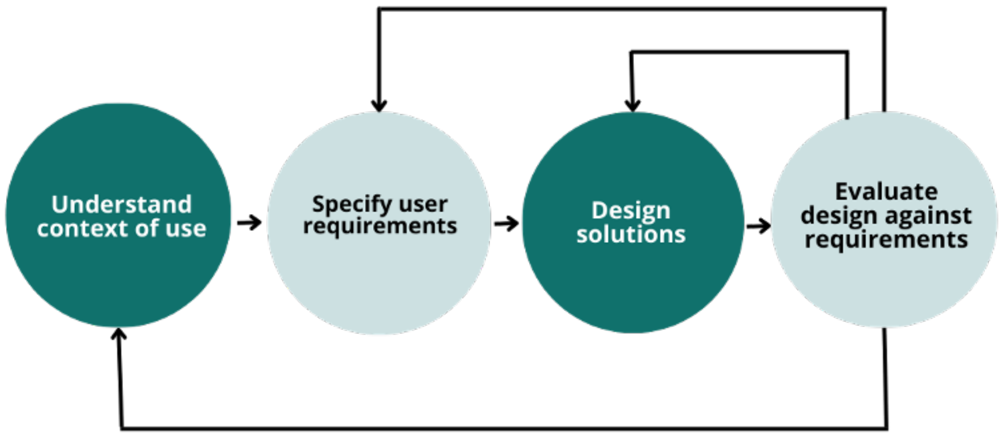

Figure 1. The four stages of user-centred design

The following sections will analyse each step of the User-Centric Design process.

4.2.1 Understanding and specifying the context of use

This phase focuses on understanding and specifying the context of use, ensuring that the system is designed to align with real-world user needs, environments, and interactions. To achieve this, a combination of qualitative and quantitative research methodologies can be employed, allowing for a well-rounded analysis of user behaviours and requirements. The selection of a specific methodology depends on factors such as the level of detail needed, the nature of user interactions, and practical constraints. Key approaches that contribute to this understanding include ethnographic studies, contextual inquiries, observational techniques, surveys, and focus groups. Each method provides unique insights and can be applied individually or in combination to develop a complete picture of how users engage with a system. A qualitative approach is ethnographic research, which involves observing users in their natural environments to understand their behaviours and contextual interactions [12]. For example, Sommerville et al. [13] conducted an ethnographic study of air traffic controllers to inform the design of an improved control system, uncovering key aspects of their workflows and decision-making processes. By immersing researchers in real user environments, ethnography provides rich insights into how systems are used in practice.

To further specify the context of use, contextual inquiry integrates observation with user interviews, allowing researchers to gather task-specific data while capturing the language and implicit knowledge that users rely on [14]. This method enables a deeper exploration of workflows, user interactions, and situational constraints, making it particularly effective for refining system requirements. While ethnographic research provides broad contextual insights, contextual inquiry focuses on specific processes and decision-making practices, making them highly complementary. To enhance this understanding with structured data, questionnaires and surveys can be used to collect quantifiable information on user preferences, experiences, and expectations. These tools help identify patterns and trends across a wider audience, offering measurable insights that complement qualitative findings. Additionally, focus groups allow for collective discussions among users, facilitating the identification of shared needs, challenges, and expectations. As Krueger and Casey [15] emphasize, group discussions can reveal insights that may not emerge through individual interviews, enriching the overall understanding of the context of use. By thoughtfully selecting and integrating these methodologies, researchers can tailor their approach to uncover critical details about how and where users interact with a system. This comprehensive understanding of the context of use is fundamental to User-Centric Design, ensuring that the developed system is aligned with real-world conditions and user needs.

4.2.2 Specifying user requirements

This phase focuses on defining user profiles and collecting and defining their requirements to ensure the system is developed in alignment with stakeholder needs and business objectives. The approach to gathering requirements is not a one-size-fits-all process. It depends on the context, including factors such as system complexity, the diversity of user groups, and available resources. Various techniques can be applied, and their selection should be adapted to effectively capture the necessary insights for a given project.

To give some examples, brainstorming sessions create an open environment where stakeholders can generate ideas, define key requirements, and establish priorities [16], [17]. This method is especially beneficial in the early stages when broad ideas need to be refined into actionable insights. Similarly, interviews and surveys help collect qualitative and quantitative data from various user groups. Interviews provide detailed insights into user concerns, while surveys help identify trends and common expectations across a larger audience [18], [19]. Another technique is defining user stories. User stories follow the format: *"As a [user], I want to [goal], so that [benefit]."* This ensures a user-focused approach to system design. Meanwhile, use cases outline specific interactions between users and the system, detailing steps, conditions, and expected outcomes [20]. These structured methods translate user requirements into precise design specifications. Additionally, prototyping and wireframing can be used in the early stages to create visual representations of the system, allowing stakeholders to refine requirements before full-scale development begins [21]. Observation and shadowing further enhance requirement gathering by enabling direct monitoring of user interactions with existing systems. This method is particularly useful when users find it difficult to articulate their needs, as it highlights pain points and workflow inefficiencies [19].

Once requirements are gathered, various techniques can be used to structure and prioritise them effectively, with the choice of method depending on factors such as project context, available resources, and stakeholder needs. To give you three examples, we have the MoSCoW prioritisation method, User Story Mapping, and the Kano Model.

The **MoSCoW** prioritisation method is a structured approach that categorises requirements into four levels: Must-have, Should-have, Could-have, and Won't-have. This method ensures that critical functionalities essential for the system's core operation are prioritised, while non-essential features remain flexible for future iterations. By applying MoSCoW, teams can make informed trade-offs, particularly in projects with strict time or budget constraints, ensuring that the most vital elements are delivered first while secondary features can be implemented later [22].

Another technique is **user story mapping**, which organises and prioritises requirements by focusing on the user's journey. This approach structures system functionality into high-level user activities and then breaks them down into specific tasks that users need to complete within each activity. By visualising these interactions, teams gain a clearer understanding of how features contribute to the overall user experience, allowing for better alignment between development efforts and user needs. Additionally, User Story Mapping helps in defining a logical, incremental development process, enabling teams to first deliver a minimum viable product (MVP) before iterating with further enhancements [23].

A different perspective is offered by the **Kano model**, which prioritises features based on customer satisfaction and their impact on user experience. This method categorises requirements into three groups: basic needs, performance needs, and delighters. Basic needs include essential functionalities that users expect, where their absence leads to dissatisfaction. Performance needs are those where improvements lead to greater user satisfaction in direct proportion to their implementation. Delighters, on the other hand, are unexpected features that provide a significant boost in customer satisfaction and can help differentiate a product in a competitive market. By using the Kano Model, teams can balance the delivery of essential features with those that enhance the overall user experience, ensuring a well-rounded product that meets both functional and emotional user expectations [24].

These three methods represent just a few of the many techniques available for structuring and prioritising requirements. The best choice depends on factors such as project goals, stakeholder expectations, and development constraints. Understanding the strengths and applications of different prioritisation frameworks allows teams to make strategic decisions that align with both technical feasibility and user value.

To ensure that requirements remain relevant and accurately reflect user and organisational needs, feedback loops should be incorporated throughout the development process. These follow-up techniques help validate assumptions, refine requirements, and maintain alignment with stakeholder expectations. Users review sessions allow key decision-makers to assess requirements iteratively and make adjustments as needed. For example, in agile development, stakeholders review and refine requirements at the end of each sprint to ensure they align with project goals. Follow-up emails and surveys serve as additional touchpoints, allowing users to clarify priorities, provide feedback, or raise concerns that may not have been addressed in earlier discussions. Finally, after gathering and refining all necessary inputs, the final requirement documentation and approval process ensures that all stakeholders are aligned before moving into the implementation phase. This step formalises the collected insights, creating a shared understanding of the system's scope, priorities, and expected outcomes [25]. By integrating a combination of collection, structuring, and feedback techniques, organisations can ensure a user-centric approach to system development that evolves in response to real user needs while remaining aligned with business goals.

4.2.3 Developing design solutions

Developing design solutions plays a critical role in creating systems that effectively meet user needs. This step involves translating identified user needs and contextual insights into concrete design ideas and prototypes. Initially, this includes generating early design concepts, such as sketches, wireframes, or low-fidelity prototypes, which serve as tangible interpretations of user and stakeholder requirements. As the design process progresses, these concepts are refined into more detailed and functional designs that take into account interaction flows, visual layout, accessibility considerations, and the broader user experience.

A key aspect of this stage is collaboration with users and stakeholders, often through participatory or co-design approaches, to ensure the design direction remains aligned with real-world needs. The focus is not only on usability but also on desirability, accessibility, and overall satisfaction. Additionally, the designs produced during this stage are used as the basis for ongoing evaluation and testing, feeding into the iterative nature of the UCD process. At this stage, the emphasis is on continuous refinement—solutions are evaluated, revised, and re-evaluated as understanding deepens or new requirements emerge.

In real-world applications, developing design solutions is highly relevant across a wide range of industries. In software and app development, UX teams often use tools like Figma or Sketch to create interactive prototypes that evolve through user feedback, as described later in Section 4.2.4). In product design, iterative prototyping helps ensure that interfaces are both intuitive and safe to use. Web designers focus on meeting accessibility standards such as the WCAG, ensuring that solutions accommodate diverse user capabilities. In service design, this process may involve mapping complete user journeys and touchpoints to build coherent and user-friendly service ecosystems. Ultimately, developing design solutions is about iteratively transforming user requirements into viable, usable, and effective systems. Through continuous testing and refinement, this step ensures that the final product or service not only functions well but is also aligned with the real-world context in which it will be used.

4.2.4 Evaluating the design against requirements

Evaluating the design against user needs and requirements is a systematic and iterative process aimed at ensuring that the solution is both functional and user-friendly. This involves continuously testing and refining the design based on real user interactions and feedback. As the fourth stage of the UCD process, this phase prioritises usability testing, structured feedback sessions, and real-world scenario simulations to identify strengths, weaknesses, and opportunities for improvement. A critical component of this evaluation is the use of continuous feedback mechanisms, which should be carefully selected based on the specific context, available resources, and project requirements. These mechanisms can include methods such as user interviews, focus groups, observational studies, and follow-up questionnaires, all of which help design teams detect usability issues early and iteratively refine the system. The evaluation process can incorporate various techniques depending on the research objectives.

For instance, ethnographic research, as emphasized by Spinuzzi [26], is a valuable method for observing users in their real-world environments. This approach helps uncover hidden or emerging user needs that may not be immediately evident during the initial design stages.

Similarly, stakeholder workshops offer a collaborative space where participants can assess design performance, provide insights, and co-create solutions to address identified challenges. Additionally, follow-up questionnaires can be distributed to users after testing sessions or system deployment to gather structured feedback over time. These surveys help track long-term user satisfaction and identify any persistent issues that might not have been immediately recognized. By adopting an iterative and participatory approach, design teams can enhance the usability, effectiveness, and alignment of the system with real-world workflows. This ensures that the final product is not only intuitive and practical but also meets both user expectations and operational demands [27].

4.3 Types of user requirements

User requirements extraction involves identifying and analysing user needs, preferences, and constraints to design user-centred systems effectively [28]. These requirements are defined as statements that provide the basis for designing and evaluating interactive systems to address some or all of the identified user needs [29]. They encapsulate the needs gathered from user input, specifying what is necessary for users to perform tasks efficiently and effectively. When gathering requirements, all stakeholders within the system's network are typically regarded as potential users [30].

These requirements are broadly categorised into **functional requirements** and **non-functional requirements**, each addressing distinct aspects of system design and performance.

4.3.1 Functional requirements

Functional requirements refer to the specific features, capabilities, or behaviours that a system must have to fulfil user needs and achieve its intended goals. These requirements focus on what the system is expected to do and outline the essential operations or tasks it should support [31], [32].

Questions to Ask:

- What tasks should the system help users perform?
- What specific features or functions are necessary for the system?
- What are the core objectives that the system must achieve?

Examples:

- The system must allow users to log in using a username and password.
- The application must enable users to leave comments and annotations on datasets or models to facilitate team collaboration.
- The platform must automatically detect and handle missing values in datasets, providing options for imputation or removal.

Functional requirements serve as the foundation for the system's core functionality and ensure that it directly supports user objectives.

4.3.2 Non-Functional requirements

Non-functional requirements specify the quality attributes, performance criteria, or constraints of a system. These requirements focus on how the system should behave or operate to ensure it performs effectively and meets user expectations [33]. Unlike functional requirements, these focus on attributes such as usability, performance, and compliance, which support the overall experience and reliability of the system.

Questions to Ask:

- How quickly should the system respond to user actions?
- What level of reliability or uptime is required for the system?
- What standards or regulations must the system comply with?

Examples:

- The system must be accessible to users with disabilities and comply with WCAG 2.1 standards [34].
- Text on the platform must meet readability standards to ensure content is comprehensible for a broad audience.
- The platform should include customisable font sizes and themes to improve accessibility and usability for diverse user preferences.

Non-functional requirements enhance the system's usability, reliability, and inclusivity, ensuring it performs well under various conditions and aligns with legal and regulatory standards.

It is important to note that the distinction between functional and non-functional requirements is not always clear-cut. In some cases, requirements may overlap or be interpreted differently depending on the context. For instance, a requirement such as "*the system must provide a screen reader-compatible navigation menu*" could be considered functional, as it describes a feature, but it may also touch on non-functional aspects like accessibility standards and compliance with guidelines such as WCAG. While it is a functional feature to ensure compatibility with screen readers, it inherently ties into non-functional considerations like inclusivity, usability, and legal compliance. In the ALFIE project, where we extract user requirements for both functional and non-functional aspects, this adaptable interpretation has been applied, recognising that some requirements may simultaneously encompass both functional and non-functional dimensions.

5 User requirements elicitation process

The user requirements elicitation process in ALFIE has been systematically designed to systematically identify, gather, and analyze user needs to ensure that the resulting design solutions are user-centric, inclusive, and aligned with ethical AI principles. This section outlines the comprehensive three-phase approach employed to collect requirements from diverse stakeholder groups, building upon the inclusive and user-centred design methodologies described in Section 4.

The elicitation process (Figure 2) was structured around a progressive methodology that began with exploratory activities involving external user groups, evolved through internal workshops with project partners, and culminated in validation activities with broader stakeholder communities. This sequential approach allowed the project team to build comprehensive understanding of user needs across the complete spectrum of potential ALFIE users, from end users who will interact with AI-generated solutions to designer users who will develop those solutions using the AutoML platform.

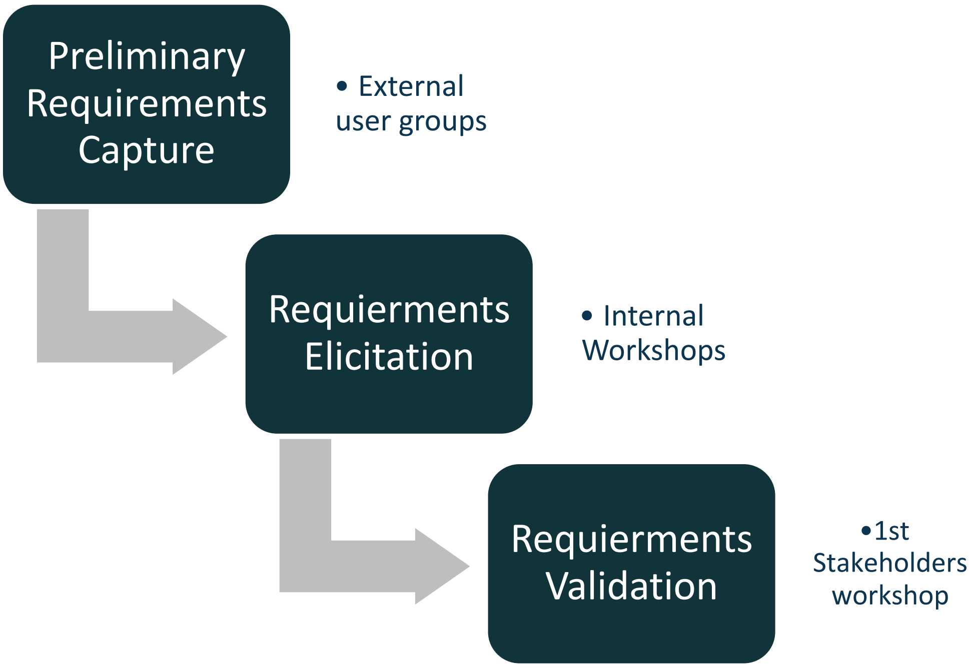

Figure 2. Requirements elicitation process

5.1 Preliminary requirements capture

The preliminary requirements capture phase represented the initial exploration of user needs and expectations for the ALFIE platform. This phase focused on engaging external user groups to understand fundamental requirements and establish the foundation for more detailed requirements gathering in subsequent phases. The approach was designed to be exploratory and inclusive, allowing diverse perspectives to inform the project's direction from the earliest stages.

5.1.1 Initial user group formation

Three distinct user groups were established to provide diverse perspectives on potential ALFIE applications and requirements. Each group was selected to represent different domains and use cases, ensuring broad coverage of potential AI ethics challenges and accessibility needs.

Educators, designers, and scientists addressing climate change – ClearClimate[4]

Context of use:

This group comprises climate communication specialists across different disciplines such as climate science, psychology, accessibility and agriculture. They focus on how AI could assist in addressing climate-related challenges, streamlining workflows, and enhancing cross-disciplinary collaboration.

Accessibility experts - Transmedia Catalonia[5]

Context of use:

This group consists of experts in the field of accessibility related to traditional and digital media, as well as emerging technologies. They focus on understanding the unique barriers faced by individuals with diverse abilities when interacting with AI systems. Their aim is to explore how AI systems can support usability and inclusivity, ensuring that the technology catered to a wide range of users.

AI specialists - ALFIE consortium

Context of use:

This group consists of project experts specialising in engineering, design, ethics, and technology. Their work focuses on promoting trustworthy AI models and frameworks by exploring methods to prevent and address bias. They are also involved in the development of an AutoML platform and an ETD Hub, a forum dedicated to discussing ethical issues related to AI.

5.1.2 Postcard-based requirements collection

To facilitate early engagement and encourage creative input, a postcard-based requirements collection methodology was implemented. This approach allowed participants to express their needs and ideas concisely while fostering engagement and collaboration. The postcard format served as a structured yet flexible tool for gathering both functional and non-functional requirements.

Participants were invited to share their requirements using a structured postcard format that encouraged concise and thoughtful input. This approach enabled participants to express both functional and non-functional requirements in a way that supported the project's commitment

[4] https://clear-climate.com/

[5] https://transmediacatalonia.uab.cat/

to ethical AI. The creative format helped participants think beyond immediate technical needs to consider broader implications of AI systems on their work and communities.

In the postcard-based requirements collection, we invited participants to share their needs using a structured postcard format, as illustrated in Figure 3. This approach encouraged concise and thoughtful input, allowing participants to express both functional and non-functional requirements in a way that supported our commitment to ethical AI. By using this format, we were able to collect input that reflected user concerns about fairness, accountability, and trust in the system.

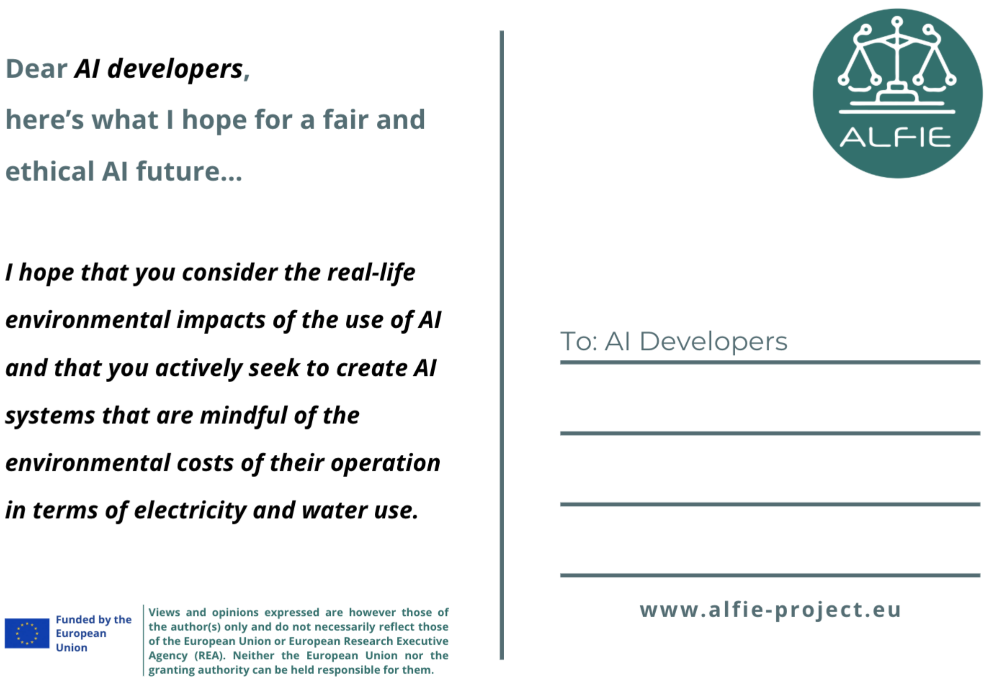

Figure 3. Example of a postcard response from User Group participant

5.1.3 Early results and thematic analysis

A total of 34 end user requirements were gathered from the three initial user groups during this preliminary phase. The distribution of contributions reflected the diverse perspectives and priorities of each group:

- **Transmedia Catalonia** contributed 10 requirements in October 2024, focusing primarily on accessibility, inclusivity, and barrier removal in AI systems, with an example shown in Figure 4.
- **ClearClimate** submitted 12 requirements in December 2024, emphasizing multilingual support, scientific text simplification, and environmental considerations, with an illustrative example provided in Figure 5.

- **ALFIE internal team** identified 12 requirements in November 2024, addressing technical functionality, bias detection, and platform integration needs, one of which is presented in Figure 6.

Theme	Quotations	User Requirements
Environmental sustainability (SDG 12)	"I hope that you consider the real-life environmental impacts of the use of AI and that you actively seek to create AI systems that are mindful of the environmental costs of their operation in terms of electricity and water use.	- AI systems should be energy efficient and minimise environmental impact. - Transparency about resource usage, such as electricity and water, should be prioritised.
Diversity and representation (SGD 10)	"When I request an image depicting a conference on technology and accessibility, I hope to see a diverse representation around the table – including women – not just white men" "I hope that minority and minorized languages are actively supported and represented in new technologies so that AI can speak these languages with the same fluency and accuracy as major ones"	- AI outputs (e.g. images or responses) should reflect diversity in gender and ethnicity, and cultural contexts. - Active support for minority and endangered languages should be incorporated into AI systems.

Figure 4. Examples of end user requirements collected with the Transmedia Catalonia group

Themes	Quotations	User Requirements
Accuracy and reliability (SDG 13)	"As a user, I wish there were less false research, scientific references. Sometimes AI compiles non-existing references." "As a user I would like to know how accurate AccWeather's rainfall forecast is for the next 48 hours in postcode X, in UK/Serbia/Florida/Madrid."	- Ensure AI outputs are accurate and reliable, particularly for critical use cases like weather forecasts. - Prevent AI systems from generating false or non-existent references in research or information dissemination
Support for vulnerable groups (SDG 10)	"As a user, I'd like a system that can reach the most vulnerable people in the city." "As a user, I would like to know how can we ensure timely evacuation for the most vulnerable residents in Santander, Spain."	- Prioritize AI applications that reach and support vulnerable populations (e.g., during emergencies like extreme heat or natural disasters). - Provide timely, actionable recommendations to ensure safety and equitable access.
Accessibility and inclusion (SDG 10)	"As a user, I want to know how we can create an AI that will help everyone, for example, change any website to be more accessible not only for the blind but also for the old" "As a user I would like to help colour blind children play with other kids without any limitations"	- Design AI systems to improve accessibility for people with disabilities (e.g. visually impaired, colour blind) and older adults. - Prevent AI systems from generating false or non-existent references in research or information dissemination

Figure 5. Examples of end user requirements collected with the ClearClimate group

Theme	Quotations	User Requirements
Bias Mitigation (SDG 9)	"As a user, I hope that AI learns from the past, but does not repeat its mistakes, leaving behind biases and prejudices that could harm people."	- Develop AI systems that identify and remove biases, ensuring decisions do not perpetuate past mistakes or prejudice.
User-Friendliness (SDG 10)	"As a user, I hope for an easy-to-use tool that will help me, as a leader of a small startup, to use AI (without AI expertise) to improve my product and better serve my customers."	- Create user-friendly AI tools accessible to non-experts, empowering startups and small businesses to leverage AI effectively.
Support for Speech Recognition (SDG 10)	"As a user, I would like you to recognise multiple accents translating speech to text so that I do not have to edit my generated transcript afterwards."	- Build speech recognition systems that accurately transcribe diverse accents, minimizing the need for manual corrections.

Figure 6. Examples of end user requirements collected with the ALFIE internal group

These requirements reflect the specific needs and priorities of each group and will guide the ongoing development of the project. Appendix A contains the complete list of requirements, along with their prioritisation based on the MoSCoW technique used during the workshop described in Section 5.1

At this stage we run an initial data analysis phase where we applied qualitative analysis tools such as NVivo[6] to identify key themes in the responses. After recognising recurring patterns, we organised individual inputs into clearly defined user requirements, as shown in Figure 7. This step helped ensure that users' values and priorities were meaningfully represented in the design process.

Theme	Quotations	User Requirements
Bias mitigation (SDG 10)	"I hope to see speech-to-text application capable of accurately recognising diverse accents and accommodating speech impairments, removing any bias rooted in language diversity." "I hope that you take into account the biases and prejudices in our society, so they aren't reflected in the AI systems you create. My hope is for AI that genuinely represents and respects people of all gender identities, sexual orientations, religions, and nationalities"	- Ensure unbiased recognition of accents, dialects, and speech impairments. - AI systems should actively prevent the reflection of societal prejudices, promoting equity across identities.
Transparency and trustworthiness (SDG 16)	"When receiving information from the AI system, I would appreciate knowing the original sources to ensure that the information is trustful."	- AI systems must provide transparent sourcing for information. - Users should be able to verify the trustworthiness and credibility of AI outputs.

Figure 7. Example of thematic analysis and user requirement definition

This thematic analysis revealed several cross-cutting concerns:

[6] https://www.nvivo-spain.com

Accessibility and inclusivity: All groups emphasized the need for AI systems that accommodate diverse user abilities, languages, and cultural contexts. Requirements focused on ensuring that AI tools do not disadvantage individuals with disabilities and provide customizable interfaces.

Transparency and trust: Participants consistently requested clear explanations of AI decision-making processes, reliable sourcing of information, and mechanisms to verify the trustworthiness and credibility of AI outputs.

Bias detection and mitigation: All groups identified the need for robust bias detection functionality and tools to develop AI systems that identify and remove biases, ensuring fair and ethical use.

Multilingual and multicultural support: Requirements emphasized the importance of processing information in multiple languages and localizations, with active support for minority and endangered languages.

Environmental responsibility: Several requirements addressed the need for energy-efficient AI systems that minimize environmental impact and provide transparency about resource usage.

5.1.4 Initial prioritization and categorization

The preliminary phase concluded with an initial categorization of requirements into functional and non-functional types. Functional requirements described specific actions the AI system should perform, such as bias detection, multilingual processing, and accessibility feature implementation. Non-functional requirements referred to broader system qualities such as usability, reliability, responsiveness, and environmental efficiency.

This preliminary categorization provided the foundation for more detailed analysis in subsequent phases and helped identify areas where additional investigation would be needed. The early insights gathered during this phase informed the design of the internal workshops (Section 5.2) and helped ensure that subsequent requirements gathering activities would build upon rather than duplicate the initial findings.

5.2 Internal requirements capture

The internal requirements capture phase involved intensive workshops with ALFIE project partners to gather detailed requirements from both end user and designer user perspectives. This phase built upon the preliminary insights gathered from external groups while leveraging the technical expertise and project knowledge of the consortium members. The internal workshops were designed to bridge the gap between high-level user needs identified in the preliminary phase and specific technical requirements needed for platform development.

5.2.1 Workshops methodology and platform

The workshop methodology employed the Miro[7] collaborative platform, which provided a digital workspace that enabled effective remote participation and real-time collaboration among participants. This platform choice proved particularly valuable given the distributed nature of the ALFIE consortium and the need to accommodate participants from different organizations, use case domains, and geographical regions. The Miro environment facilitated dynamic group formation and allowed participants to contribute ideas simultaneously while maintaining visibility of all contributions throughout the workshop process.

During each workshop, participants were organized into smaller working groups that promoted focused discussion and ensured that all voices could be heard effectively, regardless of their technical background or domain expertise. Each group was tasked with identifying and proposing different requirements relevant to their user category, drawing upon their professional experience and understanding of current challenges and future needs within their respective domains, whether industrial compliance, accessibility design, or automotive safety.

A critical component of the elicitation process involved the systematic classification of identified requirements into functional and non-functional categories. Functional requirements encompass the specific features, capabilities, and behaviours that the AutoML platform and resulting AI tools must exhibit to fulfil user needs and achieve intended objectives across all use case domains. These requirements focus on what the system should do and outline the essential operations and tasks it must support, from automated alert prioritization to accessible interface generation to unbiased emotion detection. Non-functional requirements, by contrast, specify quality attributes, performance criteria, and constraints that define how the system should behave or operate to ensure effective performance and user satisfaction across diverse application contexts. Figure 8 shows a screenshot of the requirements elicitation process using the Miro platform, whose results will be presented in Sections 6 and 7.

The classification process required participants to engage in detailed discussions about each proposed requirement, considering not only its immediate practical implications within specific use cases but also its broader impact on system design, implementation, and long-term sustainability across the entire ALFIE platform. This collaborative classification approach ensured that requirements were thoroughly understood and appropriately categorized while building consensus among participants about priorities and relationships between different requirements that span the trial use cases described in Section 3.

Both workshops followed iterative refinement processes that allowed participants to build upon initial ideas, combine related concepts, and eliminate redundancies or conflicts between proposed requirements. This iterative approach proved essential for developing a coherent set of requirements that accurately reflected the complex needs of different user groups across diverse application domains while avoiding contradictions or incompatibilities that might complicate later development phases.

[7] https://miro.com/

Figure 8. Screenshot of the user requirements capture using the Miro platform.

5.2.2 End user requirements workshop

The end user requirements workshop was conducted on January 17, 2025, focusing on gathering requirements from project partners acting in their capacity as potential end users of AI tools generated by the AutoML platform. Participants were organized into smaller working groups that promoted focused discussion and ensured that all voices could be heard effectively, regardless of their technical background or domain expertise.

Workshop structure and process

During the workshop, participants were organized into smaller working groups that promoted focused discussion and ensured comprehensive coverage of user needs across different application domains. Each group was tasked with identifying and proposing different requirements relevant to end user interactions with AI systems, drawing upon their professional experience and understanding of current challenges in their respective fields.

The workshop followed an iterative process that included:

- **Initial brainstorming** sessions where participants generated requirements without constraint
- **Clustering and thematic organization** of related requirements
- **Classification** into functional and non-functional categories
- **Prioritization** using the MoSCoW methodology
- **Validation and refinement** through group discussion

Classification and prioritization process

After reviewing the thematic groupings, participants worked together to categorise the requirements into functional and non-functional types. To support informed design decisions, we applied the MoSCoW prioritisation technique (Figure 9). This method enabled participants to rank each requirement based on necessity: Must-Have for essential features, Should-Have for important but non-critical functions, Could-Have for optional enhancements, and Won't-Have for items not relevant in the current phase but potentially useful in the future.

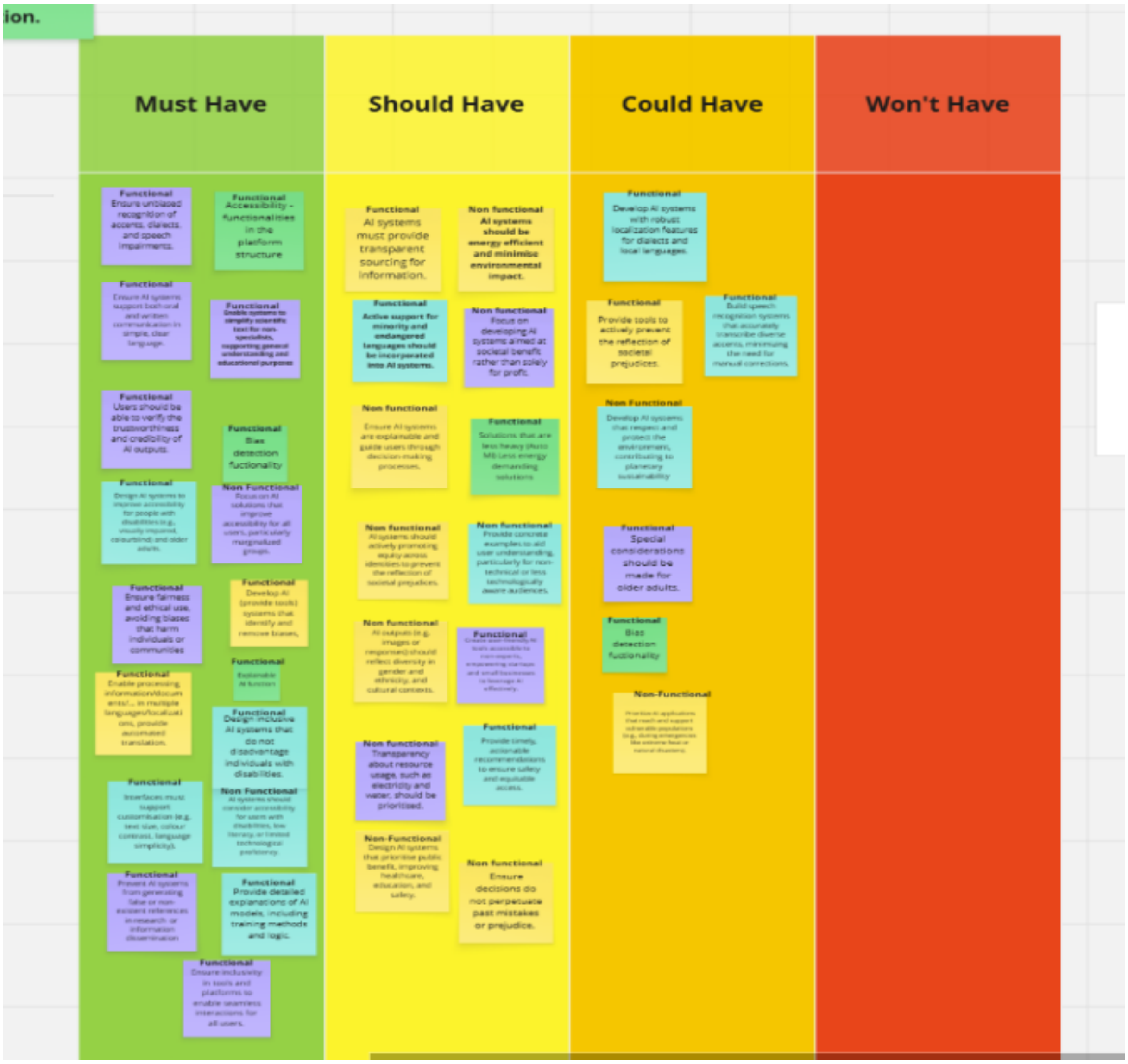

Figure 9. Screenshot of the MoSCoW evaluation process

5.2.3 Checklist for end user requirement extraction

The following Table 4 provides a structured approach to guide the user requirement extraction process in the ALFIE project. It serves as a practical checklist, ensuring that all key steps are systematically followed while maintaining ethical compliance, methodological consistency, and UCD principles.

The table is divided into three columns:

1. **Task**: Specifies the essential actions required at each stage of the user extraction process. Each task corresponds to a crucial step in defining, gathering, prioritising, and implementing user requirements.
2. **Choices/Examples**: Provides concrete examples, methodologies, or best practices that can be applied to complete each task effectively. This column offers guidance on different approaches that users can choose from depending on project needs and available resources.
3. **Status**: Allows users to track progress by indicating the completion status of each task. This column can be updated regularly to monitor implementation and ensure alignment with project timelines and objectives.

To make the most of this table, partners should:

- **Follow the sequence of tasks** to ensure a logical and structured progression from ethical guidelines to requirement evaluation.
- **Select appropriate methodologies** from the "Choices/Examples" column based on the project's scope and target users. For example, different methods such as ethnographic research, surveys, or interviews may be chosen depending on the context and user group.
- **Update the "Status" column** regularly to track progress and document any necessary modifications, ensuring all actions are completed before moving to the next stage.
- **Use the table as a reference** during team discussions, evaluations, and reporting to maintain consistency and transparency in user requirement extraction.

By following this structured approach, stakeholders can conduct user extraction while ensuring ethical integrity, methodological rigour, and alignment with the overarching goals of the ALFIE project.

Table 4. Checklist to conduct user extraction for end users

Task	Choices/examples	Status
Establish ethical guidelines for user requirement extraction	Follow GDPR, ISO/IEC 29100, informed consent practices	Completed
Prepare consent forms and ethical documentation	Create consent forms, provide information sheets, ensure the right to withdraw	Completed

Identify and define user groups	Define designer users and end users	In progress Various new user groups will be identified and taken into consideration.
Decide methodology for understanding the context of use	Ethnographic research, contextual inquiries, questionnaire, observational studies	Completed (Questionnaire has been selected as the methodology used to understand the context of use for the internal ALFIE group, while observations have been conducted with ClearClimate and Transmedia.
Choose methods for gathering user needs	Workshops, focus groups, questionnaires, interviews	In progress Workshops have been established as the initial method, with additional methods to be assessed.
Categorise requirements	Functional and non-functional	First categorisation completed
Select prioritisation methods for user requirements	MoSCoW method, Kano Model, User Story Mapping	Moscow has been selected as the initial prioritisation method.

Ensure secure data storage and GDPR compliance	Encrypted storage, limited access policies, compliance checks	Completed
Plan follow-up strategies for validation and refinement	User review sessions, follow-up surveys, feedback workshops	In progress
Develop initial prototypes based on user requirements	Wireframes, low-fidelity prototypes, usability testing	Not started
Evaluate design solutions against user needs	Usability testing, Scenario-based evaluations	Not started
Document outcomes and lessons learned	Summarise findings, update documentation, plan next iteration	Not started

5.2.4 Designer user requirements workshop

The designer user requirements workshop was conducted on April 2, 2025, focusing specifically on the needs of users who would utilize the AutoML platform to create, configure, and deploy AI models. This workshop built upon insights gained from the end user workshop while addressing the distinct needs of technical users responsible for developing AI solutions.

Participant profile and context

The designer user group included ALFIE's technological partners and other project members with expertise in engineering, design, ethics, and technology. While these participants also represented potential end users, their technical expertise positioned them in a design-focused role within the ALFIE ecosystem. Their context involved practical scenarios for deploying AI systems, emphasis on ethical challenges during implementation, and detailed understanding of bias detection and mitigation mechanisms.

Technical requirements focus

The designer user workshop addressed requirements specific to the AutoML platform's technical functionality, including:

- **Model development and training**: Requirements for automated model selection, hyperparameter tuning, and performance optimization
- **Bias detection and mitigation**: Specifications for proactive and reactive bias checking mechanisms throughout the model development lifecycle
- **Explainability and interpretability**: Needs for comprehensive explanations of model behaviour and decision-making processes
- **Integration and deployment**: Requirements for seamless integration with existing systems and standardized deployment options
- **Multi-language and accessibility support**: Technical specifications for supporting diverse languages and accessibility features
- **Data management and privacy**: Requirements for secure data handling, connection to local storage systems, and privacy-preserving techniques

Collaborative requirements definition

Similar to the end user workshop, the designer user session employed collaborative techniques to ensure comprehensive coverage of technical requirements. Participants worked in smaller groups to identify specific platform capabilities needed to support ethical AI development, with particular attention to features that would enable non-expert users to develop sophisticated AI solutions.

The workshop structure incorporated iterative refinement processes that allowed participants to build upon initial ideas, combine related concepts, and eliminate redundancies or conflicts between proposed requirements. This approach proved essential for developing a coherent set of requirements that accurately reflected the complex needs of AI developers while ensuring compatibility with the broader ALFIE platform vision.

5.2.5 Synthesis and integration

The internal requirements capture phase concluded with comprehensive synthesis activities that integrated insights from both the end user and designer user workshops. A total of 33 designer user requirements were collected and prioritized, complementing the 34 end user requirements gathered during the preliminary phase.

The synthesis process involved:

- **Cross-workshop analysis** to identify common themes and potential conflicts between end user and designer user requirements
- **Gap analysis** to identify areas where additional requirements might be needed
- **Dependency mapping** to understand relationships between different requirements
- **Technical feasibility assessment** to ensure requirements could be realistically implemented within project constraints

This comprehensive requirements base provided the foundation for the validation activities described in Section 5.3 and informed the development of the 1st stakeholder workshop methodology.

5.3 Requirements validation through the 1st stakeholders workshop

The requirements validation phase involved a comprehensive stakeholder workshop designed to validate, refine, and expand upon the requirements gathered in the preliminary and internal capture phases. This workshop represented a critical milestone in the requirements elicitation process, bringing together diverse external stakeholders to review and provide feedback on the proposed ALFIE platform components and associated requirements.

5.3.1 Workshop design and methodology

The 1st stakeholder workshop was held on June 3rd, 2025, at BOSCH Spain Headquarters in Madrid, during the 1st General Assembly of the ALFIE project. The workshop was designed as a collaborative session titled "*Shaping the Future of Ethical AI"* with the primary objective of gathering comprehensive feedback on the three main components of the ALFIE platform: the AutoML platform, the ETD-Hub, and the Use Cases. The agenda of this workshop is included as Appendix H.

The workshop employed a World Café methodology (Figure 10), a structured conversational process designed to facilitate collaboration and knowledge-sharing among diverse stakeholders. This methodology was specifically adapted to gather feedback on the three core components of the ALFIE platform, ensuring comprehensive coverage of all major system elements.

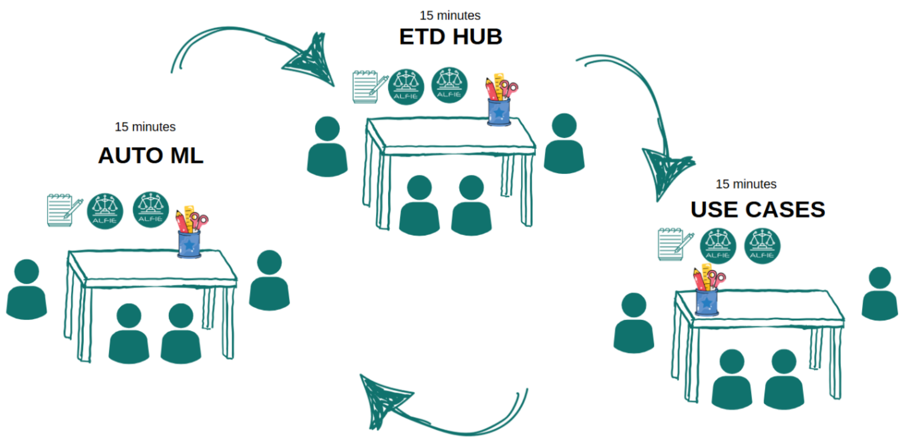

Figure 10. The World Cafè methodology structure

World Café Structure and Implementation

Setup: We set up three dedicated tables, each concentrating on a specific component:

- **AutoML Platform Table**: Focused on technical functionality, user experience, and ethical considerations
- **ETD-Hub Table**: Concentrated on stakeholder engagement, content moderation, and inclusivity
- **Use Cases Table**: Addressed ethical concerns, legal challenges, and practical implementation

Each table was assigned to at least two ALFIE moderators from the leading partners, a note-taker to capture essential points, and a facilitator to steer the discussions.

Rotation system: Participants rotated among tables, dedicating 15 minutes to each station. This format enabled everyone to share feedback on all three key areas. ALFIE moderators stayed at their designated tables during the session to ensure consistent oversight and thorough documentation.

Documentation process: Real-time note-taking was conducted using dedicated ALFIE workshop predefined templates. These templates guided the structured capture of insights, concerns, and recommendations for each component.

Sharing and synthesis: Following the three rotation rounds, moderators from each table presented a 5-minute summary of key insights. This way, cross-component themes and concerns were identified and documented.

5.3.2 Participant profile and engagement

The workshop brought together more than 20 stakeholders representing diverse perspectives essential for comprehensive requirements validation. Participants included:

- **AI professionals** with expertise in machine learning, ethics, and platform development
- **Legal experts** specializing in AI regulation, data protection, and compliance frameworks
- **Accessibility specialists** focused on inclusive design and barrier removal
- **Cybersecurity experts** addressing data protection and system security concerns
- **Researchers** from academic institutions working on ethical AI and bias mitigation
- **Representatives from various disciplines** ensuring broad coverage of potential application domains

The diverse participant's profiles ensured that requirements validation considered technical feasibility, legal compliance, ethical implications, and practical usability across different domains and user communities.

5.3.3 Structured feedback collection

The workshop employed structured templates and guided discussions to ensure systematic feedback collection on each platform component. Specific guiding questions were designed for each table to focus discussions on the most critical aspects of requirements validation.

AutoML platform feedback focus

Feedback on the AutoML platform concentrated on four key areas:

- **Features and functionality**: Participants evaluated proposed platform capabilities against their professional needs and experience
- **Transparency and explainability**: Discussion of requirements for model interpretability and decision explanation
- **Accessibility and user experience**: Assessment of platform usability across different user skill levels and abilities
- **Bias management and ethics**: Evaluation of proposed mechanisms for detecting and mitigating bias in AI systems

ETD-Hub feedback focus

The ETD-Hub discussions emphasized:

- **Platform features**: Requirements for forum functionality, content organization, and user interaction mechanisms
- **Accessibility and inclusivity**: Needs for accommodating diverse stakeholder groups and communication preferences
- **Content quality and moderation**: Requirements for ensuring productive discussions and preventing misuse
- **Integration with AutoML**: Feedback on the bi-directional relationship between ethical discussions and technical development

Use Cases Feedback Focus

Use case discussions addressed:

- **Ethical concerns**: Identification of potential ethical challenges specific to each trial use case
- **Legal and regulatory challenges**: Assessment of compliance requirements and potential barriers
- **Practical implementation**: Feedback on proposed demonstration approaches and evaluation methodologies
- **Broader applicability**: Discussion of how use case insights might apply to other domains

5.3.4 Checklist for designer requirement extraction

The guidelines for completing the checklist are the same as those outlined in Section 5.2.1, but here they are applied specifically to the designer-users (Table 5).

Table 5. Checklist to conduct user extraction for designer users

Task	Choices/examples	Status
Establish ethical guidelines for user requirement extraction	Follow GDPR, ISO/IEC 29100, informed consent practices	Completed
Prepare consent forms and ethical documentation	Create consent forms, provide information sheets, ensure the right to withdraw	Completed

Identify and define user groups	Define designer users and end users	In progress Various new user groups will be identified and taken into consideration.
Decide methodology for understanding the context of use	Ethnographic research, contextual inquiries, questionnaire, observational studies	Completed Questionnaire has been selected as the methodology used to understand the context of use for the internal ALFIE group. The ALFIE internal group is represented by both end users and designer users.
Choose methods for gathering user needs	Workshops, focus groups, questionnaires, interviews	In progress Workshops have been established as the initial method, with additional methods to be assessed.
Categorise requirements	Functional and non-functional	First categorisation completed
Select prioritisation methods for user requirements	MoSCoW method, Kano Model, User Story Mapping	Moscow has been selected as the initial prioritisation method.

Ensure secure data storage and GDPR compliance	Encrypted storage, limited access policies, compliance checks	Completed
Plan follow-up strategies for validation and refinement	User review sessions, follow-up surveys, feedback workshops	In progress
Develop initial prototypes based on user requirements	Wireframes, low-fidelity prototypes, usability testing	Not started
Evaluate design solutions against user needs	Usability testing, Scenario-based evaluations	Not started
Document outcomes and lessons learned	Summarise findings, update documentation, plan next iteration	Not started

5.4 Ensuring ethical consent

Ethical considerations are a cornerstone of the requirements-gathering process in User Centric Design. The ALFIE Project adhered to established frameworks such as the General Data Protection Regulation (GDPR)[8] and ISO/IEC 29100 Privacy Framework [35] to guide its data collection and processing activities. These regulations emphasise key principles, including:

- **Lawfulness, fairness, and transparency:** Data must be collected and processed transparently, with clear communication about its purposes.
- **Purpose limitation:** Data collection is restricted to specific, explicit purposes and cannot be repurposed without additional consent.
- **Data minimisation:** Only the data strictly necessary for project objectives is collected.
- **Accountability:** Organisations must demonstrate compliance with ethical standards through documentation and audits.

[8] Regulation (EU) 2016/679

5.4.1 Consent forms and informed participation

Consent forms were integral to the ALFIE project ethical framework. Before initiating any user requirements extraction, it is essential to obtain informed consent from all participants. This ensures that users fully understand the purpose, scope, and potential implications of their participation in the study or design process. Informed consent protects user rights, promotes ethical research practices, and fosters trust between researchers and participants. It should include clear information about data collection methods, how the information will be used, and any potential risks or benefits. Additionally, participants should have the right to withdraw at any stage without any consequences. By prioritising informed consent, the process upholds ethical standards and ensures that user engagement is voluntary, transparent, and respectful of individual autonomy.

As part of the ALFIE project, UAB have created a set of consent form templates accompanied by information sheets that outline the following:

- Purpose of the research
- What participation in the study involves
- Duration
- Risks and benefits
- Compensation
- Confidentiality
- Voluntary participation
- Right to withdraw
- Subsequent publication/re-use/processing of basic data and conservation period
- Recording and use of contributions made
- Contact person
- Consent

These templates are available in Appendices C, D, and E of this document.

5.4.2 Ethical risk mitigation

The ALFIE Project implemented measures to address potential ethical risks:

Transparency: Participants were provided with detailed documentation about data collection processes and intended uses.

Equity in representation: Recruitment efforts prioritised diverse user groups to ensure that marginalised populations were included in the design process.

Ethical oversight: Regular reviews by an ethics committee ensured ongoing compliance with regulatory and ethical standards.

5.5 Ensure Secure Data Storage

To protect user data in the ALFIE project, stringent security measures will be implemented in line with GDPR standards. All collected data will be stored securely, with role-based access controls (RBAC) restricting access to authorized personnel only. Anonymisation and pseudonymisation techniques will be applied to minimize exposure of personally identifiable information. Regular security audits and compliance checks will be conducted to identify and mitigate vulnerabilities.

Data will be retained for 5 years, ensuring compliance with legal and ethical requirements. Once no longer required, data will be securely deleted. These measures guarantee that user data remains secure, private, and compliant throughout the project lifecycle.

6 End-user requirements results

This section presents the detailed results from the end user requirements capture activities described in Section 5, focusing on the comprehensive analysis of the 65 end user requirements collected from diverse stakeholder groups. The requirements presented here are documented in full detail in Appendix A and represent the consolidated outcomes from both preliminary and internal requirements capture phases.

6.1 Overview of end-user requirements

The end user requirements capture process successfully identified 65 distinct requirements that establish the foundation for user-centric design of the ALFIE platform. These requirements reflect the diverse needs and priorities of end users who will interact with AI systems generated through the AutoML platform, encompassing individuals from various professional backgrounds, technical expertise levels, and accessibility needs.

The 65 end user requirements span the complete spectrum of user interactions with AI systems, from basic accessibility needs to advanced functionality expectations. Each requirement has been systematically categorized as either functional or non-functional and prioritized using the MoSCoW methodology to guide development decisions and resource allocation.

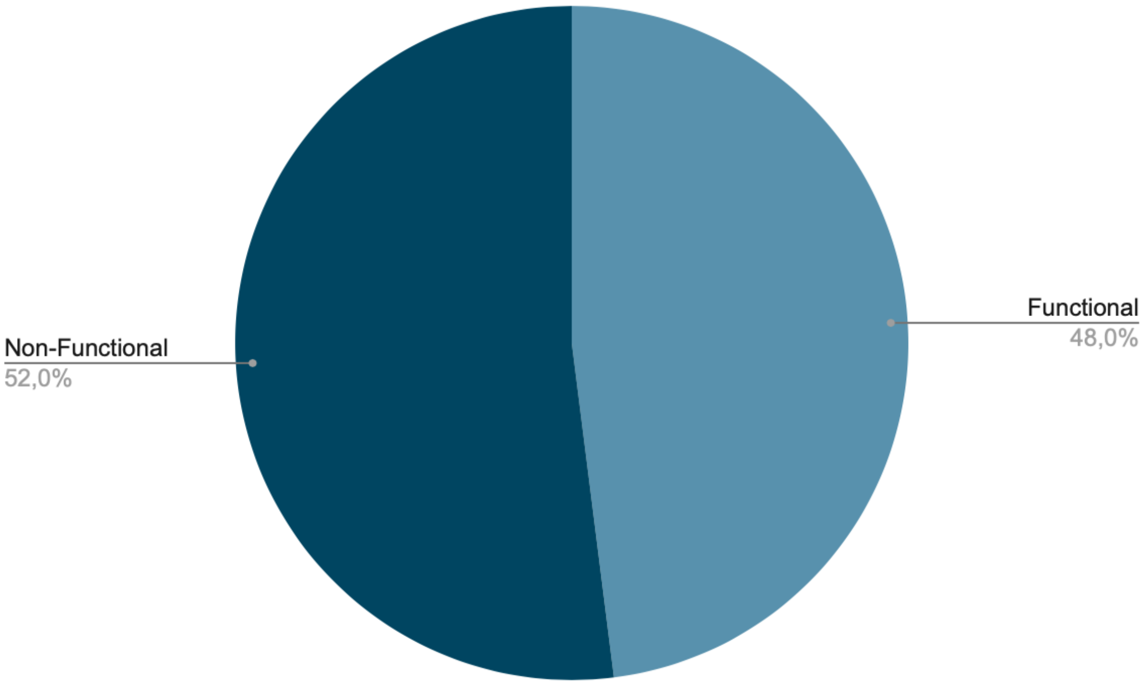

Figure 11. Distribution of end user requirements

6.2 Functional requirements analysis

The analysis reveals that 31 of the 65 end user requirements (48%) are classified as functional requirements, specifying concrete capabilities and behaviours that AI systems must exhibit to meet user needs effectively.

Core accessibility functions

A significant portion of functional requirements address accessibility and inclusive interaction capabilities:

EU-1: Unbiased recognition of accents, dialects, and speech impairments - This requirement ensures that voice-based interactions accommodate diverse linguistic backgrounds and speech patterns, preventing discrimination against users with different accents or speech characteristics.

EU-2: Accessibility functionalities in the platform structure - This requirement mandates that accessibility features are integrated into the fundamental platform architecture rather than added as secondary considerations.

EU-3: Support for oral and written communication in simple, clear language - This requirement ensures that AI systems can interact effectively with users who prefer different communication modes or have varying literacy levels.

EU-4: Scientific text simplification for non-specialists - This functional requirement enables AI systems to translate complex technical information into accessible language for general audiences.

AI system capabilities

Several functional requirements specify essential AI system behaviours and capabilities:

EU-5: User verification of trustworthiness and credibility - This requirement specifies that users must be able to independently verify the reliability of AI outputs through transparent sourcing and evidence provision.

EU-6 & EU-37: Bias detection functionality - These requirements specify that AI systems must include automated mechanisms for identifying potential biases in data, algorithms, and outputs.

EU-10: Systems that identify and remove biases - This requirement goes beyond detection to specify active bias mitigation capabilities.

EU-15: Prevention of false or non-existent references - This requirement ensures AI systems maintain information integrity by preventing the generation of fabricated citations or sources.

Multilingual and communication functions

Language and communication capabilities represent a critical category of functional requirements:

EU-11: Multi-language processing and automated translation - This requirement specifies comprehensive multilingual support for information processing and user interaction.

EU-16: Detailed explanations of AI models and logic - This requirement mandates that AI systems provide comprehensive explanations of their decision-making processes in understandable terms.

EU-20: Active support for minority and endangered languages - This requirement extends language support to include specialized linguistic communities.

Interface and interaction functions

User interface and interaction requirements specify how users should be able to engage with AI systems:

EU-13: Customizable interfaces (text size, contrast, language simplicity) - This requirement specifies that users must be able to adapt interface elements to their individual needs and preferences.

EU-17: Inclusive tools enabling seamless interactions - This requirement ensures that platform tools are designed to accommodate diverse user interaction patterns and capabilities.

EU-27: User-friendly tools accessible to non-experts - This requirement specifies that AI capabilities must be accessible to users without technical expertise.

6.3 Non-functional requirements analysis

The remaining 34 requirements (52%) are classified as non-functional, addressing quality attributes, performance criteria, and operational constraints that define how AI systems should behave to ensure effective performance and user satisfaction.

Performance and efficiency requirements

EU-19 & EU-35: Energy efficiency and environmental impact minimization – These two requirements establish performance criteria for sustainable AI system operation.

EU-23: Less energy-demanding solutions - This requirement specifically addresses the computational efficiency of AutoML processes.

EU-28: Transparency about resource usage - This requirement mandates that users receive clear information about the environmental and computational costs of AI operations.

Quality and reliability requirements

EU-8: Focus on accessibility improvements for marginalized groups - This requirement establishes quality standards for inclusive design that prioritize underserved populations.

EU-14: Accessibility considerations for diverse user capabilities - This requirement defines performance standards for accommodating users with disabilities, low literacy, or limited technological proficiency.

EU-22: Explainable systems guiding decision-making - This requirement establishes quality standards for AI transparency and user guidance.

Ethical and social requirements

EU-21: Focus on societal benefit rather than profit - This requirement establishes ethical constraints on AI system development priorities.

EU-24: Active promotion of equity across identities - This requirement defines standards for preventing discrimination and promoting fairness.

EU-26: Diverse representation in AI outputs - This requirement establishes quality standards for inclusive AI-generated content.

EU-30: Prioritization of public benefit - This requirement defines operational priorities for AI system deployment and use.

6.4 Prioritization results using MoSCoW methodology

The MoSCoW prioritization process revealed clear consensus among end users regarding the relative importance of different requirements, providing crucial guidance for development planning and resource allocation.

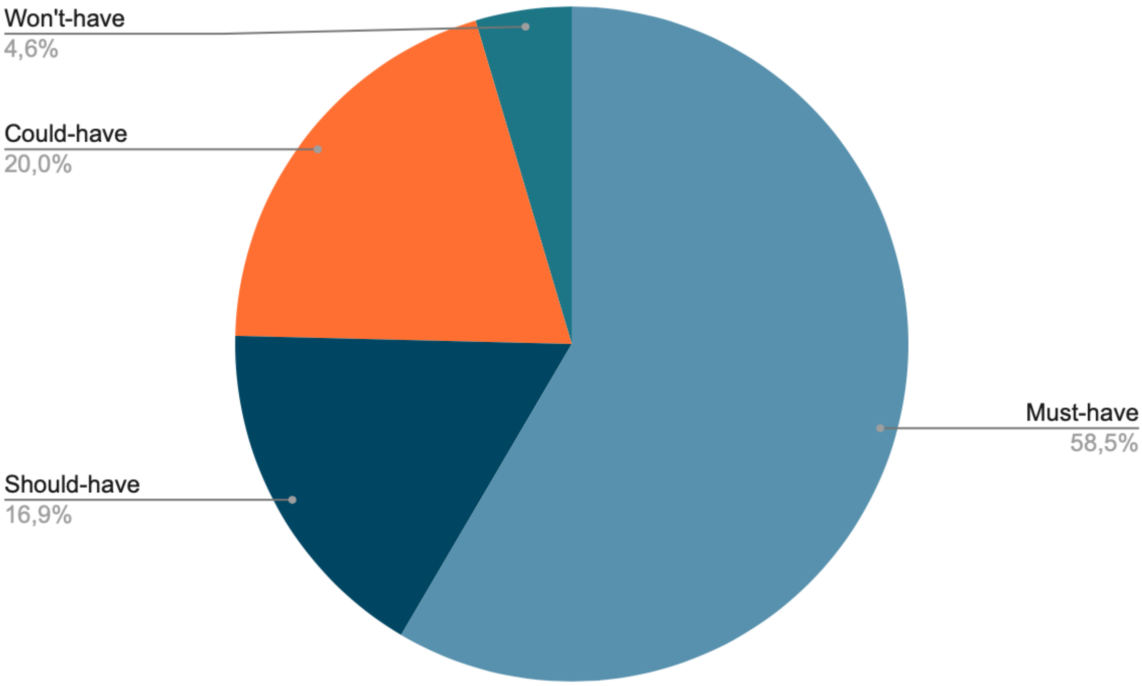

Figure 12. Distribution of end user requirements according to MoSCoW priorization

6.4.1 Must-Have requirements

The Must-Have category includes 38 requirements requirements that end users consider absolutely essential for platform acceptance and basic functionality. These requirements represent non-negotiable capabilities that form the foundation of the ALFIE platform:

Critical accessibility requirements (8 requirements):

- EU-1: Unbiased recognition of accents, dialects, and speech impairments
- EU-2: Accessibility functionalities in platform structure
- EU-3: Support for oral and written communication in clear language
- EU-7: Accessibility improvements for people with disabilities and older adults
- EU-8: Focus on accessibility for marginalized groups
- EU-12: Inclusive systems that don't disadvantage individuals with disabilities
- EU-46: User profile customization
- EU-47: Create a system that works for everyone, making sure it is accessible and understandable to all users.

Essential AI capabilities (11 requirements):

- EU-5: User verification of trustworthiness and credibility
- EU-6, EU-56: Bias detection functionality
- EU-9, EU-42, EU-60: Fairness and ethical use, avoiding harmful biases
- EU-10: Systems that identify and remove biases
- EU-40, EU-41, EU-61: Model interpretability and transparency
- EU-48: Ethical and contextual aware AI

Core functionality (12 requirements):

- EU-4: Scientific text simplification for non-specialists
- EU-11: Multi-language processing and automated translation
- EU-13: Customizable interfaces
- EU-14: Accessibility for users with diverse capabilities
- EU-16, EU-53, EU-54: Detailed explanations of AI models and logic
- EU-17: Inclusive tools enabling seamless interactions
- EU-45: Speech input capability
- EU-55, EU-62, EU-63: Explainability of the system

Regulatory Compliance (7 requirements)

- EU-39, EU-49, EU-52, EU-57, EU-58, EU-64: Compliance with current regulations (including privacy and data protection, GDPR)
- EU-51: Ethical AI, Compliance, and System Integrity

The high proportion of Must-Have requirements (58%) demonstrates strong consensus among end users about fundamental platform capabilities, with particular emphasis on accessibility, ethical AI behaviour and regulatory compliance.

6.4.2 Should-Have requirements

Should-Have requirements enhance platform value and user experience but are not critical for initial implementation. These 13 requirements support advanced functionality and improved user satisfaction:

Enhanced user experience (5 requirements):

- EU-18: Transparent sourcing for information
- EU-22: Explainable systems guiding decision-making
- EU-25: Concrete examples for user understanding
- EU-27: User-friendly tools for non-experts
- EU-29: Timely, actionable recommendations

Ethical and social enhancement (5 requirements):

- EU-20: Support for minority languages
- EU-21: Focus on societal benefit rather than profit
- EU-24: Active promotion of equity across identities
- EU-26: Diverse representation in AI outputs
- EU-31: Prevention of perpetuating past mistakes

Performance optimization (3 requirements):

- EU-19, EU-23: Energy efficiency and environmental impact minimization
- EU-28: Transparency about resource usage

6.4.3 Could-Have requirements

The 11 Could-Have requirements represent desirable enhancements that would add significant value but are not necessary for platform success:

Advanced language support (2 requirements):

- EU-32: Robust localization features for dialects and local languages
- EU-33: Accurate transcription of diverse accents

Enhanced features (9 requirements):

- EU-34: Tools to prevent reflection of societal prejudices
- EU-35: Environmental protection and sustainability focus
- EU-36: Special considerations for older adults
- EU-37: Bias detection
- EU-38: Support for vulnerable populations during emergencies
- EU-15, EU-18, EU-43, EU-59: Reliability of sources.

6.4.4 Won't have requirements

Won't-Have requirements refer to those which are ouside the scope of the AutoML platform at the time of writing. These do not mean that these requirments cannot be implemented at a later stage. These 3 tentative requirements primarily relate to misuse prevention and user control and feedback.

Misuse prevention

- EU-44 and EU-50: Ethical misuse and prevention
- EU-65: User control and feedback.

7 Designer user requirements results

This section presents the comprehensive analysis of the 39 designer user requirements collected during the internal requirements capture phase described in Section 5.2 and the 1st stakeholder's workshop. These requirements, documented in full detail in Appendix B, reflect the specific needs of technical/designer users who will utilize the AutoML platform to create, configure, and deploy AI models with ethical considerations integrated throughout the development process.

7.1 Overview of designer user requirements

The designer user requirements represent a comprehensive specification for AutoML platform capabilities that enable ethical AI development by users with varying levels of machine learning expertise. The 39 requirements span the complete AI development lifecycle, from initial data exploration and model specification through deployment, monitoring, and maintenance.

Unlike end user requirements that focus on interaction with completed AI systems, designer user requirements address the technical infrastructure and capabilities needed to create those systems. These requirements reflect the dual challenge of making advanced AI development accessible to non-experts while providing the sophisticated functionality needed for ethical AI development.

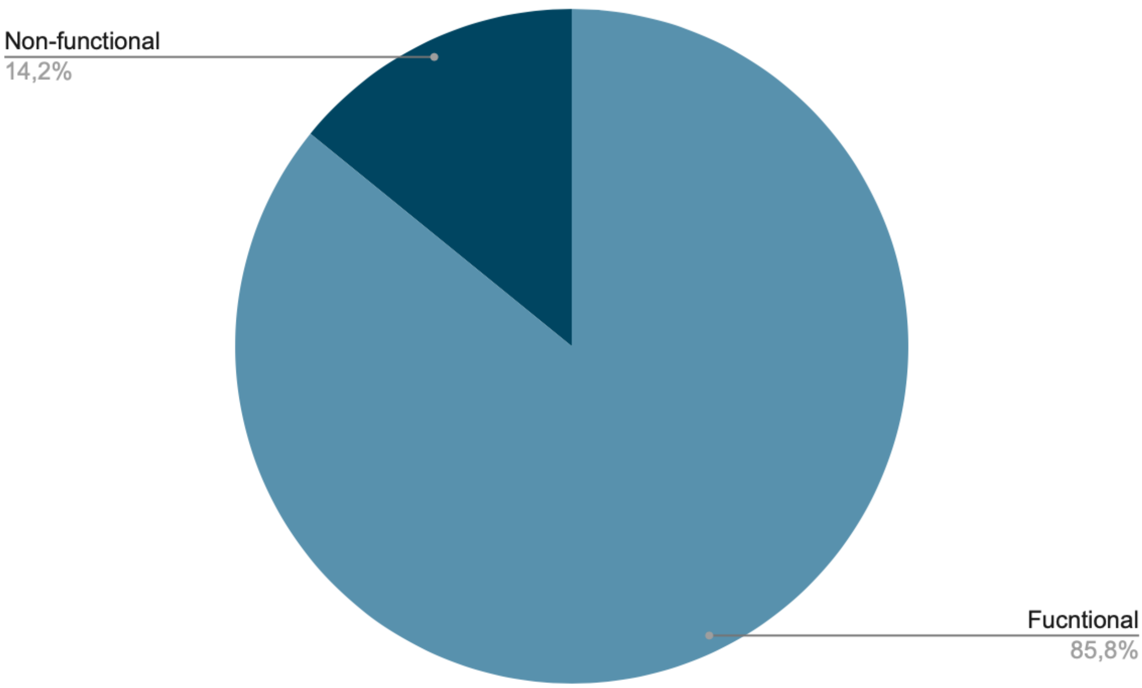

Figure 13. Distribution of designer user requirements

7.2 Functional requirements analysis

The analysis reveals that 34 of the 39 designer user requirements (86%) are classified as functional requirements, reflecting the technical nature of AI development tasks and the need for specific platform capabilities.

Data management and integration functions

A significant category of functional requirements addresses data management capabilities essential for ethical AI development:

DUR-31: Search for datasets in the AutoML data warehouse - This requirement enables users to discover existing datasets that may be suitable for their AI development needs, promoting data reuse and consistency.

DUR-32: Upload new datasets to the AutoML data warehouse - This requirement allows users to contribute new datasets to the shared repository, expanding the available resources for AI development.

DUR-33: Connection to local storage when upload is not possible - This requirement accommodates scenarios where data cannot be shared due to privacy, security, or proprietary concerns, enabling AI development with private datasets.

DUR-30: Comprehensive dataset description including variables and target identification - This requirement ensures that datasets are thoroughly documented with metadata describing their structure, content, and intended use, enabling informed model development decisions.

DUR-25: Support for multiple data format standards - This requirement ensures platform compatibility with diverse data sources and organizational systems.

DUR-35: The system should provide functionality for text summarisation and keyword recognition – This requirement requests additional functionality to the platform.

Model development and configuration functions

Core model development capabilities represent the largest category of functional requirements:

DUR-28: Request model generation with specified parameters - This requirement enables users to initiate AI model development with specific requirements and constraints.

DUR-29: Describe model goals (classification, forecasting, etc.) - This requirement allows users to specify the type of AI task they want to accomplish, enabling appropriate algorithm selection and configuration.

DUR-12: Specify technical requirements (real-time processing, device constraints, performance trade-offs) - This requirement enables users to define deployment constraints that influence model architecture and optimization decisions.

DUR-23: Describe the domain of the target model - This requirement allows users to provide context about the application domain, enabling domain-specific optimization and bias consideration.

DUR-34: Users should be able to select a specific model type for their task – This requirement enables the generation of a custom ML model.

DUR-36: It should raise alerts if harmful or unethical content is detected, ensuring that general users are protected and aware of potential risks – This requirement implies the analysis of any content uploaded to the platform to detect harmful content (in any format).

DUR-38: Prompt the user to ask questions or clarify doubts if something is unclear – This requirements enables the resolution of users' questions.

DUR-39:Fake News Alerts It should flag potential misinformation. This requirement is analogous to DUR-36 but focuses on textual content.

<u>**Bias detection and mitigation functions**</u>

Ethical AI development capabilities represent a critical category of functional requirements:

DUR-26: Proactive bias checking throughout model development - This requirement ensures that bias detection occurs automatically during the AI development process, identifying potential issues before they become embedded in deployed systems.

DUR-27: Reactive bias checking on user request - This requirement allows users to explicitly request bias analysis at any point during development, enabling targeted investigation of specific concerns.

DUR-21: Detailed explanations of identified biases - This requirement ensures that when biases are detected, users receive comprehensive information about the nature, source, and potential impact of those biases.

DUR-24: Automated bias mitigation solution proposals - This requirement goes beyond bias detection to provide actionable recommendations for addressing identified bias issues.

DUR-9: Clear explanation of bias mitigation actions taken - This requirement ensures transparency about what steps have been taken to address bias concerns, enabling user validation and audit trail maintenance.

DUR-37: I want a clear, visual traffic light system (red, yellow, green) that indicates risk levels. This requirement aims to democratize the use of the model by making risk assessments understandable and accessible. It should also explain how decisions are made, allowing to see why the model reached its conclusions, and giving the choice to approve or reject them.

<u>**Model deployment and integration functions**</u>

Deployment capabilities enable the transition from development to production use:

DUR-7: Simple and standard model deployment methods - This requirement ensures that model deployment is accessible to users without specialized DevOps expertise.

DUR-8: Model download in standard formats for integration - This requirement enables integration with external systems and proprietary software by providing models in widely-supported formats.

DUR-14: Deployment as endpoints in various environments - This requirement supports diverse deployment scenarios, from cloud services to on-premises systems.

DUR-11: Automatic front-end generation for model endpoints - This requirement reduces the technical burden of creating user interfaces for deployed AI models.

Code generation and local development functions

Advanced capabilities support users who need greater control over the AI development process:

DUR-3: Provide code for local training when data cannot be uploaded - This requirement enables AI development with sensitive or proprietary data that cannot be shared with the platform.

DUR-5: Provide code for local fine-tuning - This requirement allows users to customize and optimize models using their own computational resources and expertise.

DUR-4: Upload locally generated models for ethical bias checking - This requirement enables users who develop models independently to leverage the platform's bias detection capabilities.

7.3 Non-functional requirements analysis

The remaining 5 requirements (15%) are classified as non-functional, addressing quality attributes and operational characteristics that define how the platform should behave to support effective AI development.

Usability and accessibility requirements

DUR-10: Accessible user interface design - This requirement ensures that the platform itself is accessible to users with disabilities, extending the accessibility focus beyond generated AI systems to the development tools themselves.

DUR-19: Multilingual platform with multiple interaction modes - This requirement addresses the need for platform usability across diverse linguistic and cultural contexts, supporting global AI development efforts.

Performance and monitoring requirements

DUR-6: Comprehensive explainability using standard XAI methods - This requirement establishes quality standards for model interpretability, ensuring that generated AI systems can be understood and validated by users.

DUR-15: Standard performance metrics for generated models - This requirement ensures that users receive standardized measures of model quality and effectiveness, enabling informed deployment decisions.

Intelligence and adaptation requirements

DUR-2: Intelligent warning system for missing information or resources - This requirement establishes standards for platform responsiveness and user guidance, ensuring that users receive appropriate feedback about development process status and requirements.

7.4 Prioritization results using MoSCoW methodology

The MoSCoW prioritization process revealed clear consensus among designer users regarding the relative importance of different technical requirements, providing crucial guidance for development planning and resource allocation.

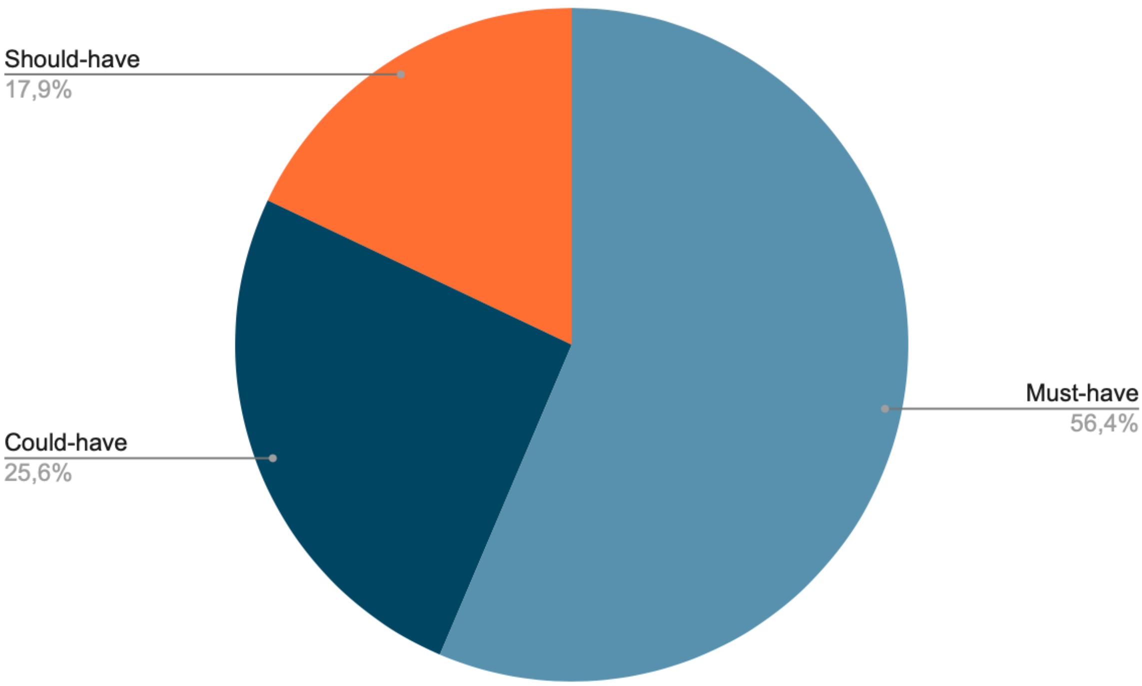

Figure 14. Distribution of designer user requirements according to MoSCoW priorization

7.4.1 Must-Have requirements

The Must-Have category includes 22 requirements that designer users consider absolutely essential for platform functionality and acceptance. These requirements represent non-negotiable technical capabilities that form the core of the AutoML platform:

Critical data management requirements (5 requirements):

- DUR-31: Search for datasets in the AutoML data warehouse
- DUR-32: Upload new datasets to the AutoML data warehouse
- DUR-33: Connection to local storage when upload is not possible
- DUR-30: Comprehensive dataset description including variables and target identification
- DUR-28: Ask to generate the model

Essential AI development capabilities (9 requirements):

- DUR-29: Describe the goal: classification, forecasting, etc.
- DUR-23: Describe the domain of the target model
- DUR-12: Technical requirements specification for models
- DUR-26: Proactive bias checking throughout model development
- DUR-27: Reactive bias checking on user request
- DUR-21: Detailed explanations of identified biases
- DUR-35: Text summarisation and keyword recognition.
- DUR-36: Detection of harmful or unethical content.
- DUR-37: Clear indication of risk levels.

Core platform infrastructure (8 requirements):

- DUR-24: Automated bias mitigation solution proposals
- DUR-9: Clear explanation of bias mitigation actions taken
- DUR-19: Multilingual platform with multiple interaction modes
- DUR-6: Comprehensive explainability using standard XAI methods
- DUR-2: Intelligent warning system for missing information or resources
- DUR-16: Model retraining and evaluation after bias mitigation
- DUR-15: Standard performance metrics for generated models
- DUR-38: Prompt the user to ask questions or clarify doubts if something is unclear.

The high proportion of Must-Have requirements (56) demonstrates strong consensus among designer users about fundamental platform capabilities, with particular emphasis on data management, bias detection, and explainability.

7.4.2 Should-have requirements

Should-Have requirements enhance platform capabilities and support advanced use cases but are not critical for initial implementation. These 10 requirements support sophisticated functionality and improved developer experience:

Advanced development features (5 requirements):

- DUR-3: Provide code for local training when data cannot be uploaded
- DUR-5: Provide code for local fine-tuning
- DUR-18: User-specified performance metrics
- DUR-1: Multi-language support across platform components
- DUR-13: Alternative solution suggestions when ML is inappropriate

Enhanced integration and deployment (5 requirements):

- DUR-7: Simple and standard model deployment methods
- DUR-25: Support for multiple data format standards
- DUR-4: Local model upload for ethical bias checking
- DUR-10: Accessible user interface design
- DUR-17: Training cost and environmental impact calculation

7.4.3 Could-have requirements

Could-Have requirements (5) represent advanced features that would add significant value but are not necessary for platform success:

Advanced automation features (3 requirements):

- DUR-11: Automatic front-end generation for model endpoints
- DUR-8: Model download in standard formats for external integration
- DUR-14: Model deployment as endpoints in various environments
- DUR-34: Users should be able to select a specific model type for their task.

Enhanced monitoring and analytics (2 requirements):

- DUR-20: Language-aware performance monitoring
- DUR-22: Risk ranking for identified issues (critical, warning, etc.)
- DUR-39 Fake News Alerts.

8 First requirements validation results

8.1 Overview

The 1st external stakeholder's workshop had as primary objective to gather comprehensive feedback on the three main components of the ALFIE platform: the AutoML platform, the ETD-Hub, and the Use Cases.

The workshop followed a structured agenda that included:

- Introduction to the ALFIE ecosystem and its bi-directional feedback loop between ETD-Hub and AutoML
- Detailed presentations of each platform component, including the EthiTech Dialogue Hub (ETD-Hub), the AutoML platform architecture, and the Trial Use Cases (TUCs)
- Interactive discussion and feedback sessions focused on legal & ethical considerations, data aspects, technical features, and accessibility & user-centric design
- Summary of key insights and next steps for stakeholder engagement

The workshop brought together more than 20 stakeholders, including AI professionals, legal experts, accessibility specialists, cybersecurity experts, researchers, and representatives from various disciplines to ensure comprehensive coverage of ethical, technical, and practical considerations.

8.2 Results

This section presents an analysis of the feedback collected using the methodology outlined earlier. The guiding questions used during the sessions, along with the note taker template, are provided in Appendix G.

8.2.1 Feedback for the AutoML platform

The feedback on the AutoML platform revealed several critical themes across functional requirements, transparency, and ethical considerations.

Security and compliance requirements: Participants emphasized the paramount importance of regulatory compliance, particularly with GDPR in Europe and equivalent regulations in other jurisdictions. There was strong demand for cybersecurity standards compliance and transparent risk assessment mechanisms. A traffic light system (red, yellow, green) was suggested for visual risk indication to democratise model understanding and decision-making.

Transparency and explainability: A major concern centered on achieving true transparency in AutoML systems. Participants questioned whether transparency is genuinely achievable with black-box models and requested features that allow exploration of model decision-making processes. They emphasised the need for clear explanations tailored to different user expertise levels, using simple language and visual aids. The system should provide explanations of feature importance, model decision paths, and the ability to adjust models when necessary.

Accessibility and user experience: Strong emphasis was placed on making AutoML accessible to diverse user groups. Participants requested voice interaction capabilities, especially for users who may have difficulty with traditional interfaces. The system should accommodate different user experience levels (beginner, intermediate, expert) and provide guidance mechanisms, text summaries, and keyword recognition tools.

Bias management and ethical safeguards: Participants highlighted the critical need for handling diverse contexts including ethnicity and discrimination issues. The system must implement strict privacy and security safeguards for personal and medical data, while also detecting and preventing misuse, bad intentions, and hate speech. User protection mechanisms and automatic alerts for harmful or unethical content were deemed essential.

8.2.2 Feedback for the ETD-Hub

The ETD-Hub component generated discussions focused on source reliability, moderation challenges, and inclusivity concerns.

Source quality and reliability: Participants emphasised that ethical sources should extend beyond human inputs to include established references such as Aristotle's ethics, treaties on Natural Sciences, and health-related documents. There was significant concern about the trustworthiness of sources and the need for robust criteria to evaluate compliance across diverse contexts. The question of "who decides" on source reliability was raised as a critical governance issue.

Moderation and bias prevention: Strong concerns were expressed about preventing bias in participatory systems and managing the dynamic nature of chat-based interactions. Participants recommended implementing automatic moderation for risks, establishing filters to detect bots, and closely monitoring participant profiles to ensure authenticity. The potential for echo chamber effects was identified as a significant risk requiring active mitigation strategies.

Flexibility and multidisciplinary approach: The feedback stressed the need for a broad, multidisciplinary approach that accommodates different cultural contexts, religions, and languages. Participants suggested creating multiple systems or frameworks rather than a one-size-fits-all solution, with chat versions tailored to different groups. The importance of grounding abstract topics in practical use cases to establish shared language and semantics was emphasized.

Inclusivity and accessibility: Recommendations included implementing anonymisation of comments, user profile monitoring, and like systems for comment prioritization. The need for guided debates and AI variability for different topics to enable segmentation was highlighted. Clear definition of ethics (legal, cultural, or domain-specific) was identified as fundamental to the platform's effectiveness.

8.2.3 Feedback for the Use Cases

The feedback on use cases revealed consistent concerns across all three trial scenarios regarding data protection, regulatory compliance, and ethical implementation.

Data protection and privacy: Across all use cases (compliance alerts, IRIS driver monitoring, and accessibility evaluation), participants identified data breaches and unauthorized access to personal and sensitive data as primary concerns. Specific attention was given to biometric data storage in the IRIS case, with emphasis on guaranteeing data elimination after established periods. The need for clear data storage policies and robust control measures was consistently highlighted.

Regulatory and legal challenges: Regulatory differences among countries emerged as a significant challenge across all cases, particularly regarding data protection and compliance frameworks. Participants noted the complexity of ensuring consistent compliance across different jurisdictions while maintaining system effectiveness. The issue of consent mechanisms was raised, particularly for website owners in the accessibility use case and for users providing biometric data in the IRIS system.

Ethical implementation concerns: For the compliance alerts use case, concerns centered on subjectivity in handling alerts and the need for objective criteria for compliance decisions. The IRIS use case raised questions about cultural differences in emotional expression and the risk of misinterpretation across different demographic groups. The accessibility use case highlighted the importance of following established accessibility standards while respecting cultural preferences and differences.

Risk Mitigation Strategies: Participants consistently recommended comprehensive risk analysis, considering multiple parameters to address the identified concerns. Data validation mechanisms to prevent infiltration of fake data were suggested, along with the need for transparent criteria sources and trustworthy regulatory references. The development of certification systems or badges for good practices was proposed as a motivation mechanism for stakeholder participation.

9 AutoML validation framework

9.1 Overview

This section presents a comprehensive validation plan that outlines the evaluation criteria, methodologies, and metrics for assessing the performance of the ALFIE AutoML platform across three distinct trial scenarios. The validation framework is designed to ensure systematic evaluation of the platform's capabilities, effectiveness, and trustworthiness while addressing key considerations such as transparency, fairness, and accountability in model selection and deployment.

The validation plan is built upon the foundation of user requirements extracted from diverse stakeholder groups, including end users with special skills and needs, designer users, and technical implementers. It incorporates findings from stakeholder workshops and follows an iterative approach that allows for continuous improvement and adaptation throughout the project lifecycle.

The trial scenarios evaluated include: (1) **Compliance Alerts Classification** for regulatory compliance monitoring, (2) **IRIS Driver Monitoring** for emotion and stress detection, and (3) **Web Accessibility Evaluation** for inclusive interface design. Each scenario presents unique challenges and requirements that collectively provide a comprehensive assessment of the AutoML platform's versatility and effectiveness across diverse application domains.

9.2 Global methodology

9.2.1 Multi-phase validation approach

The validation methodology follows a structured three-phase approach designed to provide comprehensive evaluation throughout the project lifecycle:

Phase 1: Pre-Trial Validation (M9-M15)

- Initial platform assessment using simulated datasets
- User interface and experience validation with designer users
- Baseline performance measurement establishment

Phase 2: Trial Implementation and Evaluation (M15-M30)

- Real-world trial scenario execution across all three use cases
- Continuous monitoring and data collection
- Stakeholder feedback integration and iterative improvements

Phase 3: Post-Trial Analysis and Optimization (M31-M36)

- Comprehensive performance analysis and reporting
- Long-term impact assessment
- Scalability and sustainability evaluation
- Final validation against all defined KPIs

9.2.2 Evaluation criteria

The validation framework employs six primary evaluation dimensions:

Technical performance criteria:

Model accuracy and precision metrics
Processing speed and computational efficiency
Platform stability and reliability
Integration capabilities with existing systems

User experience criteria:

- Usability and accessibility measurements
- User satisfaction ratings
- Learning curve and adoption rates
- Interface effectiveness across diverse user groups

Ethical and compliance criteria:

- Bias detection and mitigation effectiveness
- Data privacy and security compliance
- Regulatory adherence across jurisdictions
- Transparency and explainability metrics

Scalability and adaptability criteria:

- Cross-domain performance consistency
- Multi-language support effectiveness
- Platform growth and adoption rates
- Integration flexibility with different workflows

Innovation and impact criteria:

- Novelty of solutions generated
- Real-world problem-solving effectiveness
- Contribution to accessibility and inclusion
- Industry adoption and standardization potential

Trustworthiness and acceptability criteria:

- Stakeholder confidence levels
- Expert validation and peer review
- Reliability in decision-making scenarios
- Acceptance across diverse user communities

9.3 Trial scenario-specific validation plans

9.3.1 Compliance alerts classification trial use case

This trial evaluates the AutoML platform's capability to classify compliance alerts from internet services, specifically focusing on relevance determination for Bosch Compliance Regulations.

The scenario involves processing approximately a large dataset of unique alerts with multiple classifications per alert.

Specific KPIs and metrics

Table 6. Specific KPIs and metrics for the compliance alerts classification TUC

Metric Category	Specific Metric	Target/Requirement
Technical Performance Metrics		
Technical Performance	**Classification Accuracy**	Target ≥95% precision for "Relevant" class identification
	Processing Speed	Target 5% improvement classification time
	Data Handling Capacity	Successful processing of large datasets
User Engagement Metrics		
User Engagement	**Active User Participation**	Compliance officers actively using the system
	User Satisfaction	Average score ≥4.5 on 5-point Likert scale
	Training Effectiveness	Users able to complete tasks independently after training
	Workflow Integration	Successful integration into existing compliance processes
Ethical and Compliance Metrics		
Ethical & Compliance	**Data Privacy Compliance**	100% adherence to GDPR and internal data policies
	Bias Detection	Automated detection of potential classification biases
	Transparency	Clear explanation provided for classification decisions
	Audit Trail	Complete traceability of all classification decisions

Validation methodology:

Data quality assessment:

- Comprehensive analysis of the alert dataset for completeness and consistency
- Validation of annotation quality and inter-rater reliability among compliance officers
- Assessment of data representativeness across different compliance domains

Performance evaluation:

- Cross-validation using 80-20 training-testing split
- Temporal validation using historical data to predict recent alerts
- Comparative analysis against current manual classification processes

User acceptance testing:

- Structured user testing sessions with compliance officers
- Task-based evaluation measuring completion time and accuracy
- System Usability Scale (SUS) questionnaire administration
- Focus group discussions for qualitative feedback collection

Ethical validation:

- Independent audit of classification decisions for potential biases
- Privacy impact assessment for data handling procedures
- Compliance verification against relevant regulatory frameworks
- Transparency evaluation through explainability feature testing

9.3.2 IRIS driver monitoring trial

This trial assesses the platform's ability to develop emotion and stress detection models for driver monitoring applications. The scenario emphasizes biometric data processing, cross-cultural emotion recognition, and real-time performance in automotive environments.

Specific KPIs and metrics:

Table 7. Specific KPIs and metrics for the IRIS driver monitoring TUC

Metric Category	Specific Metric	Target/Requirement
Technical Performance Metrics		
Technical Performance	**Emotion Recognition Accuracy**	Target ≥70% across primary emotion categories
	Cultural Adaptation	Performance variation <5% across different demographic groups
	Real-time Processing	Target <100ms latency for emotion detection
	Biometric Data Security	100% compliance with data protection regulations
User Experience Metrics		
User Experience	**Driver Acceptance**	Minimum 10 drivers from diverse backgrounds participating
	Usability Rating	Average Likert score ≥ 4.5 on accessibility and comfort measures

	Non-intrusive Operation	Minimal impact on driving performance and attention
	Cross-cultural Effectiveness	Consistent performance across ethnic and cultural groups
Safety and Ethical Metrics		
Safety & Ethical	**Privacy Protection**	Automatic data deletion after specified retention periods
	Bias Mitigation	Equal performance across demographic groups
	Consent Management	Clear and informed consent from all participants
	Safety Compliance	No adverse effects on driving safety or performance

Validation methodology:

Multi-phase testing protocol:

- Laboratory phase: Controlled environment testing
- Simulator phase: Driving simulator evaluation under various stress conditions
- Real-world phase: Limited field testing with volunteer drivers in controlled scenarios

Cross-cultural validation:

- Participant recruitment across diverse demographic groups
- Cultural bias assessment through comparative performance analysis
- Expert review by cultural psychology and automotive safety specialists
- Validation against established emotion recognition benchmarks

Privacy and ethics assessment:

- Comprehensive privacy impact assessment
- Data minimization and anonymization protocol validation
- Consent process evaluation for clarity and comprehensiveness
- Independent ethics committee review and approval

Technical validation:

- Accuracy measurement against ground truth emotion labels
- Latency testing under various computational load conditions
- Robustness evaluation under different lighting and environmental conditions
- Integration testing with automotive systems and protocols

9.3.3 Web accessibility evaluation trial

This trial evaluates the platform's capability to assess and improve web accessibility, focusing on generating inclusive interfaces for diverse user groups including elders, children, and individuals with disabilities.

Specific KPIs and metrics:

Table 8. Specific KPIs and metrics for the Web accesibility checker TUC

Metric Category	Specific Metric	Target/Requirement
Accessibility Performance Metrics		
Accessibility Performance	**Issue Detection Accuracy**	Target ≥95% for identifying accessibility barriers
	Compliance Coverage	Assessment against WCAG 2.1 AA standards
	Improvement Measurement	Quantifiable accessibility score improvements
	Multi-modal Support	Effective support for screen readers, voice commands, and adaptive layouts
User Inclusion Metrics		
User Inclusion	**Diverse User Participation**	Minimum 10 users across different ability groups
	Satisfaction Across Groups	Average score ≥4.5 for all user categories
	Accessibility Feature Adoption	More than 70% of users are able to use the platform effectively to perform their tasks

Validation methodology:

Inclusive user testing:

- Recruitment of participants representing diverse ability groups
- Task-based evaluation covering content upload, feedback review, and fix application
- Assistive technology compatibility testing (screen readers, voice control, etc.)
- Longitudinal user experience tracking over multiple sessions

Accessibility compliance assessment:

- Automated WCAG 2.1 compliance checking
- Expert accessibility audits by certified professionals
- Comparative analysis against leading accessibility tools
- Real-world website testing across various content types and complexity levels

Impact measurement:

- Before-and-after accessibility score comparisons
- User engagement and retention analysis
- Industry adoption tracking through deployment monitoring
- Scalability assessment across different sectors and applications

Technical validation:

- HTML structure analysis accuracy testing
- Image analysis and alt-text generation validation
- ARIA label detection and recommendation accuracy
- Integration testing with content management systems and development workflows

9.4 Cross-scenario validation elements

Next we define the validation elements for the AutoML platform

9.4.1 Platform-wide performance metrics

Trustworthiness and explainability:

- Explanation quality assessment across all three scenarios
- User comprehension testing of AI-generated explanations
- Expert validation of decision reasoning
- Consistency measurement of explanations across similar cases

Natural language interaction:

- Speech recognition accuracy across multiple languages and accents
- Text input processing effectiveness
- Multilingual support validation
- Conversational interface usability assessment

Data security and privacy:

- End-to-end encryption of the data transmitted through the platform's components
- Access control mechanism testing
- Data anonymization effectiveness measurement
- Compliance verification across jurisdictional requirements

Integration and interoperability:

- API performance and reliability testing
- Third-party system integration validation
- Workflow compatibility assessment
- Scalability testing under varying load conditions

9.4.2 Stakeholder engagement validation

Multi-stakeholder feedback integration:

- Regular feedback collection from companies, policymakers, and end-users
- Stakeholder satisfaction measurement across different engagement levels
- Expert panel evaluation for technical and ethical assessment
- Community feedback integration through iterative development cycles

Training program effectiveness:

- User competency assessment before and after training
- Training material accessibility and comprehensiveness evaluation
- Knowledge retention measurement over time

- Training scalability assessment for different user groups

9.4.3 Validation success criteria

Technical success criteria:

- Achievement of ≥90% of defined technical KPIs across all scenarios
- Successful integration and deployment in real-world environments
- Demonstrated improvement over existing solutions
- Scalability validation across different application domains

User acceptance success criteria:

- User satisfaction scores ≥4.5 across all user groups
- Successful task completion rates ≥70% for all scenarios
- Active user engagement meeting or exceeding participation targets
- Positive accessibility and inclusion impact demonstration

Impact success criteria:

- Measurable improvement in target application domains
- Adoption by external organizations beyond the consortium
- Contribution to relevant standards and best practices
- Evidence of sustainable long-term value creation
- Academic publications in relevant journals
- Academic presentations at relevant conferences

10 Conclusions

The definition and validation framework for the three trial use cases demonstrates ALFIE's comprehensive approach to developing ethical and inclusive AI solutions through the AutoML platform. The systematic application of User-Centred Design methodology throughout the project has successfully established a robust foundation for evaluating the platform's effectiveness across diverse application domains while maintaining strong emphasis on accessibility, bias mitigation, and stakeholder engagement.

10.1 Trial Use Cases (TUCs) definition

The three trial use cases represent carefully selected scenarios that collectively validate the AutoML platform capabilities across critical areas of AI application. The compliance alerts classification scenario for Bosch addresses the industrial need for automated decision support in regulatory environments, targeting measurable improvements in processing efficiency while maintaining high accuracy standards. The UAB website accessibility evaluation use case demonstrates the platform capacity to generate tools that promote digital inclusion, specifically focusing on creating accessible interfaces for users with diverse abilities. The CTL driver monitoring system tackles the challenge of developing unbiased emotion and drowsiness detection models, addressing critical safety applications while ensuring fairness across different demographic groups.

The stakeholder engagement activities, particularly the 1st stakeholders workshop conducted at Bosch headquarters, have provided valuable validation of the requirements gathering process while identifying additional considerations for platform development. The feedback received across the AutoML platform, ETD-Hub, and use case components confirms the relevance of the identified requirements while highlighting the importance of regulatory compliance, transparency mechanisms, and cultural adaptation in AI system development.

The user requirements extraction process has successfully captured both end user and designer user perspectives, resulting in prioritized requirement sets that guide the platform’s development decisions. The high proportion of Must-Have requirements in both categories demonstrates strong consensus among stakeholders about fundamental platform capabilities, particularly in areas of accessibility, bias detection, and explainability. The systematic application of MoSCoW prioritization provides clear direction for development resource allocation while ensuring that critical functionality receives appropriate attention.

10.2 Validation framework

The validation methodology establishes comprehensive evaluation criteria that span technical performance, user experience, ethical compliance, and real-world impact. Each use case incorporates specific Key Performance Indicators that reflect both quantitative measures such as accuracy thresholds and processing speeds, alongside qualitative assessments including user satisfaction scores and accessibility ratings. This dual approach ensures that the AutoML platform demonstrates not only technical competence but also genuine usability and acceptance among diverse user communities.

The evaluation framework multi-phase approach enables continuous assessment throughout the project lifecycle, from initial platform validation through real-world deployment and long-term impact analysis. This structured progression ensures that lessons learned during early phases inform subsequent development activities while maintaining focus on the ultimate goal of demonstrating measurable improvements in AI system trustworthiness and accessibility.

The integration of ethical considerations throughout the validation process reflects ALFIE's commitment to responsible AI development. Each use case addresses specific ethical challenges relevant to its application domain, from ensuring fair treatment in compliance decisions to protecting driver privacy in biometric monitoring systems. The comprehensive consent procedures and data protection measures establish strong foundations for ethical research practices while enabling meaningful stakeholder participation.

The project emphasis on accessibility and inclusion extends beyond compliance with technical standards to encompass genuine usability improvements for marginalized communities. The validation criteria include specific measures for diverse user group participation and satisfaction, ensuring that the AutoML platform delivers tangible benefits to users who have traditionally been underserved by AI technologies.

This comprehensive validation framework positions ALFIE to make meaningful contributions to the field of ethical AI development while demonstrating practical approaches to democratizing access to advanced machine learning capabilities. The successful completion of the planned trials will provide evidence-based insights into the effectiveness of AutoML platforms for addressing real-world challenges while maintaining high standards for fairness, transparency, and inclusivity.

References

[1] A. Halin, J. G. Verly, and M. V. Droogenbroeck, "Survey and synthesis of state of the art in driver monitoring," *Sensors*, vol. 21, no. 16, p. 5558, Aug. 2021, doi: 10.3390/s21165558.

[2] Asgarian, Azin, et al. Limitations and Biases in Facial Landmark Detection -- An Empirical Study on Older Adults with Dementia. arXiv:1905.07446, arXiv, 17 May 2019. arXiv.org, https://doi.org/10.48550/arXiv.1905.07446.

[3] D. Salem and M. Waleed, "Drowsiness detection in real-time via convolutional neural networks and transfer learning," *J. Eng. Appl. Sci.*, vol. 71, no. 1, p. 122, May 2024, doi: 10.1186/s44147-024-00457-z.

[4] G. S. Krishna, K. Supriya, J. Vardhan, and M. R. K, "Vision Transformers and YoloV5 based Driver Drowsiness Detection Framework," Sep. 03, 2022, *arXiv*: arXiv:2209.01401. doi: 10.48550/arXiv.2209.01401.

[5] S. S. Sengar, A. Kumar, and O. Singh, "VigilEye -- Artificial Intelligence-based Real-time Driver Drowsiness Detection," arXiv.org. Accessed: Jun. 13, 2025. [Online]. Available: https://arxiv.org/abs/2406.15646v1

[6] “What Is Mean Average Precision,” Oslo - Vision AI for all. Accessed: July 24, 2025. [Online]. Available: https://oslo.vision/blog/what-is-mean-average-precision/

[7] Joshi, A., Kale, S., Chandel, S., & Pal, D. K. (2015). Likert scale: Explored and explained. *British Journal of Applied Science & Technology*, 7(4), 396–403.

[8] Boone, H. N., & Boone, D. A. (2012). Analyzing Likert data. *Journal of Extension*, 50(2), 1–5.

[9] A. Savidis and C. Stephanidis, "Unified user interface design: designing universally accessible interactions," *Interacting with Computers*, vol. 16, no. 2, pp. 243-270, 2004.

[10] S. Chew, "An approach to design with people who have special needs," in *HCI International 2013 - Posters' Extended Abstracts: International Conference, HCI International 2013, Las Vegas, NV, USA, July 21-26, 2013, Proceedings, Part I*, vol. 15, pp. 221-225, Springer Berlin Heidelberg, 2013.

[11] International Organization for Standardization, ISO 9241-210:2019 -- Ergonomics of Human-System Interaction -- Part 210: Human-Centered Design for Interactive Systems, ISO, 2019. [Online]. Available: https://www.iso.org/standard/77520.html

[12] C. Spinuzzi, "The methodology of participatory design," *Technical Communication*, vol. 52, no. 2, pp. 163-174, 2005.

[13] I. Sommerville, "Systems engineering for software engineers," *Annals of Software Engineering*, vol. 6, no. 1, pp. 111-129, 1998.

[14] K. Holtzblatt, "Contextual design," in *The Human-Computer Interaction Handbook*, pp. 975-990, CRC Press, 2007.

[15] R. A. Krueger and M. A. Casey, *Designing and Conducting Focus Group Interviews*, vol. 18, 2002.

[16] J. G. Byrne and T. Barlow, "Structured brainstorming: A method for collecting user requirements," in *Proceedings of the Human Factors and Ergonomics Society Annual Meeting*, vol. 37, no. 5, pp. 427-431, Sage, Los Angeles, CA, 1993.

[17] A. Azadegan, K. N. Papamichail, and P. Sampaio, "Applying collaborative process design to user requirements elicitation: a case study," *Computers in Industry*, vol. 64, no. 7, pp. 798-812, 2013.

[18] S. Weibelzahl, A. Jedlitschka, and B. Ayari, "Eliciting requirements for an adaptive decision support system through structured user interviews," 2006.

[19] C. Courage and K. Baxter, *Understanding Your Users: A Practical Guide to User Requirements Methods, Tools, and Techniques*. Gulf Professional Publishing, 2005.

[20] G. Lucassen, F. Dalpiaz, J. M. E. van der Werf, and S. Brinkkemper, "Improving agile requirements: the quality user story framework and tool," *Requirements Engineering*, vol. 21, pp. 383-403, 2016.

[21] C. Andrews et al., "A new method in user-centered design: Collaborative prototype design process (CPDP)," *Journal of Technical Writing and Communication*, vol. 42, no. 2, pp. 123-142, 2012.

[22] I. P. Ramayasa and I. G. A. Candrawibawa, "Usability evaluation of lecturer information systems using Sirius framework and Moscow technique," *Scientific Journal of Informatics*, vol. 8, no. 1, pp. 16-23, 2021.

[23] J. Patton and P. Economy, *User Story Mapping: Discover the Whole Story, Build the Right Product*. O'Reilly Media, Inc., 2014.

[24] T. Wang and P. Ji, "Understanding customer needs through quantitative analysis of Kano's model," *International Journal of Quality & Reliability Management*, vol. 27, no. 2, pp. 173-184, 2010.

[25] Z. S. Li et al., "Unveiling the life cycle of user feedback: Best practices from software practitioners," in *Proceedings of the 46th IEEE/ACM International Conference on Software Engineering*, pp. 1-13, 2024.

[26] C. Spinuzzi, *Tracing Genres through Organizations: A Sociocultural Approach to Information Design*, vol. 1, MIT Press, 2003.

[27] J. Good and J. Robertson, "Learning and motivational affordances in narrative-based game authoring," in Proceedings of the 4th International Conference for Narrative and Interactive Learning Environments (NILE), Edinburgh, pp. 37-51, 2006.

[28] M. Maguire and N. Bevan, "User requirements analysis: a review of supporting methods," in *IFIP World Computer Congress, TC 13*, Boston, MA, USA: Springer US, 2002, pp. 133-148.

[29] International Organization for Standardization, *ISO 25065:2019 Systems and software engineering --- Software product Quality Requirements and Evaluation (SQuaRE) --- Common Industry Format (CIF) for Usability: User requirements specification*, Geneva, Switzerland: ISO, 2019.

[30] J. Hanson, *The Inclusive City: Delivering a More Accessible Urban Environment through Inclusive Design*, UCL Discovery, 2004.

[31] S. Robertson and J. Robertson, Mastering the requirements process: Getting requirements right. Addison-Wesley, 2012.

[32] I. Sommerville, Software Engineering, 7th ed. Pearson Education, 2004.

[33] L. Chazette and K. Schneider, "Explainability as a non-functional requirement: challenges and recommendations," Requirements Engineering, vol. 25, no. 4, pp. 493-514, 2020.

[34] World Wide Web Consortium (W3C), Web Content Accessibility Guidelines (WCAG) 2.1, 2018. [Online]. Available: https://www.w3.org/TR/WCAG21/

[35] International Organization for Standardization, ISO/IEC 29100:2011 - Information Technology - Security Techniques - Privacy Framework, Geneva, Switzerland: ISO, 2011.

Appendices

Appendix A: End user requirements

ID	End User Requirement(s)	Functional or Non-functional	Priority (Must Have, Should Have, Could Have, Won't Have)
EU-1	Ensure unbiased recognition of accents, dialects, and speech impairments.	Functional	Must have
EU-2	Accessibility - functionalities in the platform structure	Functional	Must have
EU-3	Ensure AI systems support both oral and written communication in simple, clear language.	Functional	Must have
EU-4	Enable systems to simplify scientific text for non-specialists, supporting general understanding and educational purposes	Functional	Must have
EU-5	Users should be able to verify the trustworthiness and credibility of AI outputs.	Functional	Must have
EU-6	Bias detection functionality	Functional	Must have
EU-7	Design AI systems to improve accessibility for people with disabilities (e.g., visually impaired, colourblind) and older adults.	Functional	Must have
EU-8	Focus on AI solutions that improve accessibility for all users, particularly marginalized groups.	Non-functional	Must have

EU-9	Ensure fairness and ethical use, avoiding biases that harm individuals or communities	Functional	Must have
EU-10	Develop AI (provide tools) systems that identify and remove biases	Functional	Must have
EU-11	Enable processing information/documents/... in multiple languages/localizations, provide automated translation.	Functional	Must have
EU-12	Design inclusive AI systems that do not disadvantage individuals with disabilities.	Functional	Must have
EU-13	Interfaces must support customisation (e.g. text size, colour contrast, language simplicity).	Functional	Must have
EU-14	AI systems should consider accessibility for users with disabilities, low literacy, or limited technological proficiency.	Non-functional	Must have
EU-15	Prevent AI systems from generating false or non-existent references in research or information dissemination	Functional	Must have
EU-16	Provide detailed explanations of AI models, including training methods and logic.	Functional	Must have
EU-17	Ensure inclusivity in tools and platforms to enable seamless interactions for all users.	Functional	Must have
EU-18	AI systems must provide transparent sourcing for information.	Functional	Should have
EU-19	AI systems should be energy efficient and minimise environmental impact.	Non-functional	Should have

EU-20	Active support for minority and endangered languages should be incorporated into AI systems.	Functional	Should have
EU-21	Focus on developing AI systems aimed at societal benefit rather than solely for profit.	Non-functional	Should have
EU-22	Ensure AI systems are explainable and guide users through decision-making processes.	Non-functional	Should have
EU-23	Solutions that are less heavy (Auto Ml) Less energy demanding solutions	Functional	Should have
EU-24	AI systems should actively promoting equity across identities to prevent the reflection of societal prejudices.	Non-functional	Should have
EU-25	Provide concrete examples to aid user understanding, particularly for non-technical or less technologically aware audiences.	Non-functional	Should have
EU-26	AI outputs (e.g. images or responses) should reflect diversity in gender and ethnicity, and cultural contexts.	Non-functional	Should have
EU-27	Create user-friendly AI tools accessible to non-experts, empowering startups and small businesses to leverage AI effectively.	Functional	Should have
EU-28	Transparency about resource usage, such as electricity and water, should be prioritised.	Non-functional	Should have
EU-29	Provide timely, actionable recommendations to ensure safety and equitable access.	Functional	Should have

EU-30	Design AI systems that prioritise public benefit, improving healthcare, education, and safety.	Non-functional	Should have
EU-31	Ensure decisions do not perpetuate past mistakes or prejudice.	Non-functional	Should have
EU-32	Develop AI systems with robust localization features for dialects and local languages.	Functional	Could have
EU-33	Build speech recognition systems that accurately transcribe diverse accents, minimizing the need for manual corrections.	Functional	Could have
EU-34	Provide tools to actively prevent the reflection of societal prejudices.	Functional	Could have
EU-35	Develop AI systems that respect and protect the environment, contributing to planetary sustainability	Non-functional	Could have
EU-36	Special considerations should be made for older adults.	Functional	Could have
EU-37	Bias detection fuctionality	Functional	Could have
EU-38	Prioritize AI applications that reach and support vulnerable populations (e.g., during emergencies like extreme heat or natural disasters).	Non-functional	Could have
EU-39	I want AutoML platforms to comply with current regulations, including the GDPR in Europe, regulations in the Americas, and cybersecurity standards. This would give me confidence that my data is protected and handled responsibly.	Non-functional	Must have (GDPR)

EU-40	I want to understand how models are making decisions. Is transparency truly achievable? Many AutoML tools use black-box models, but there are also more interpretable models like decision trees. I would like to see comparisons between these models to understand the trade-offs and make informed choices.	Non-functional	Must have
EU-41	I would appreciate features that allow me to explore how different models work and why they make specific predictions. This could include explanations of features' importance, model decision paths, and the ability to adjust models if needed.	Non-functional	Must have
EU-42	I want a built-in mechanism that guides the model's use of information in an ethical way.	Non-functional	Must have
EU-43	The system should ensure that it does not use unreliable references.	Non-functional	Could have
EU-44	The system should not allow anyone to manipulate or misuse it in ways that would pervert its purpose or violate ethical guidelines. (The bad user)	Non-functional	Won't have
EU-45	The system should provide options for users to input commands via speech.	Functional	Must have
EU-46	Users should be able to specify their experience level when first interacting with AutoML.	Functional	Must have

EU-47	Create a system that works for everyone, making sure it is accessible and understandable to all users.	Non-functional	Must have
EU-48	The system should be able to handle diverse contexts, including ethnicity and discrimination, in a sensitive and respectful way.	Non-functional	Must have
EU-49	The system must treat personal and medical data with strict privacy and security safeguards.	Non-functional	Must have
EU-50	AutoML should be designed to detect and prevent any misuse, including bad intentions, attempts to pervert the system, or hate speech.	Non-functional	Won’t have
EU-51	Some concepts should be considered irremovable, such as an imperative minimum, ethical standards, and respect for human rights.	Non-functional	Must have
EU-52	The system should clearly flag any issues that might conflict with ethical guidelines or legal frameworks. It should be able to say: “Alert: This is against our ethical system.	Functional	Must have
EU-53	Understandability and Bibliography If I don’t understand how the model works or cannot see explanations, I will not use it. The system should include references or a bibliography and avoid generating hallucinations.	Non-functional	Must have
EU-54	Provide a visual timeline to keep the user informed and engaged. The system should actively show what’s happening at each stage. Raise awareness of the process.	Functional	Must have

EU-55	Offer clear and straightforward explanations so the user understands each step.	Non-functional	Must have
EU-56	Decisions should be based on objective criteria. There should be no room for subjective interpretation in critical areas, especially concerning human rights compliance.	Non-functional	Must have
EU-57	AutoML should comply with GDPR	Non-functional	Must have
EU-58	The system should clearly show risks related to legal compliance, including privacy and data protection.	Functional	Must have
EU-59	The system must avoid making things up or generating fake information. It should always tell me when it cannot provide a reliable answer.	Non-functional	Could have
EU-60	It should alert me if the system generates information that is inconsistent with verified sources or potentially violates legal or ethical guidelines.	Non-functional	Must have
EU-61	I would like visual aids that help me understand the model's decisions, such as graphs, charts, or summaries. A reference list should always be available, and the dataset used for training should be documented and accessible.	Functional	Must have
EU-62	I want to see how decisions are made. The system should provide explanations tailored to users, using clear and simple language. It should show how different	Non-functional	Must have

	factors contributed to the outcome and why a specific decision was reached.		
EU-63	Explainability should be easy to understand, not just for experts. The system should show me how decisions are made using clear and simple language. It should also explain which factors influenced the decision and why the system reached that particular result.	Non-functional	Must have
EU-64	AutoML should comply with all relevant laws, including GDPR.	Non-functional	Must have
EU-65	AutoML should comply with all relevant laws, including GDPR.	Non-functional	Won’t have

Appendix B: Designer user requirements

ID	Designer User Requirement(s)	Functional / Non-functional	Priority (Must Have, Should Have, Could Have, Won't Have)
DUR-1	Multi-language Support	Functional	Should have
DUR-2	If the system requires additional info or resources to generate the model, it should warn and ask the designer to provide them.	Non-Functional	Must have
DUR-3	Provide code for local train if the data cannot be uploaded	Functional	Should have
DUR-4	If the model is generated locally, the designer can upload it and the system may check for ethical biases/issue. As the original dataset cannot be uploaded (proprietary), the system could use an alternative (similar) dataset.	Functional	Could have
DUR-5	Provide code for local fine tuning	Functional	Should have
DUR-6	Explain the outcomes of the model (standard XAI local or global methods)	Non-Functional	Must have
DUR-7	The system must provide simple and standard ways to deploy the generated model	Functional	Should have
DUR-8	The model should be downloaded using standard formats to be integrated in proprietary software	Functional	Could have
DUR-9	Explain any bias/issue mitigation action	Functional	Must have

DUR-10	The UI of the system must be accessible.	Non-Functional	Should have
DUR-11	The system should generate a front-end for the generated end-point, using existing templates.	Functional	Could have
DUR-12	When asking for the model, the designer must state additional technical requirements. For example, real time detection. Or generated for specific devices. Or, we can sacrifice performance in order to increase processing speed.	Functional	Must have
DUR-13	If there is a goal that cannot be achieved through a ML model, the system must suggest alternative solutions/applications/tools	Functional	Should have
DUR-14	The model should be deployed as an end-point, either in the system's servers or to the cloud, or manually by the designers in their own servers.	Functional	Could have
DUR-15	After generating a model the system must provide standard performance metrics	Non-Functional	Should have
DUR-16	Once a bias or issue has been addressed the system must retrain and evaluate the new model	Functional	Must have
DUR-17	Calculate and provide the approximate cost of training the model (and possibly its environmental impact?)	Functional	Could have
DUR-18	The designer must ask for performance metrics.	Functional	Should have
DUR-19	The system must be multilingual and enable interaction in different ways.	Non-Functional	Must have

DUR-20	The system may be aware of its performance when interacting in different languages	Functional	Could have
DUR-21	Each potential bias/issue must be explained	Functional	Must have
DUR-22	The system must rank the potential issues identified in the dataset/model (critical, warning, etc)	Functional	Could have
DUR-23	Describe the domain of the target model	Functional	Must have
DUR-24	Once a bias is identified by the system, it must propose a solution (component, data cleaning, ...)	Functional	Must have
DUR-25	Support as many standards as possible to upload the dataset	Functional	Could have
DUR-26	[PROACTIVE BIAS CHECK] The system should check at any time (during the interaction) if there is any kind of bias. It follows the designer's instructions and proactively checks for potential biases.	Functional	Must have
DUR-27	[REACTIVE BIAS CHECK]The designer can ask the system for bias	Functional	Must have
DUR-28	Ask to generate the model	Functional	Must have
DUR-29	Describe the goal: classification, forecasting, etc.	Functional	Must have
DUR-30	Once a tabular dataset is uploaded, it must be described: columns, independent variables, identify target variables, values ... (i.e. variable "var12" represents the gender, 0: male, 1: female)	Functional	Must have

DUR-31	Search for a dataset in the automl data warehouse	Functional	Must have
DUR-32	Upload a new dataset to the automl data warehouse	Functional	Must have
DUR-33	If upload is not possible (proprietary data), a connection to the local storage must be provided/enabled (probably through API calls)	Functional	Must have
DUR-34	Users should be able to select a specific model type for their task.	Functional	Could have
DUR-35	The system should provide functionality for text summarisation and keyword recognition.	Functional	Expected
DUR-36	It should raise alerts if harmful or unethical content is detected, ensuring that general users are protected and aware of potential risks.	Functional	Expected
DUR-37	I want a clear, visual traffic light system (red, yellow, green) that indicates risk levels. This helps democratize the use of the model by making risk assessments understandable and accessible. It should also explain how decisions are made, allow me to see why the model reached its conclusions, and give me the choice to approve or reject them.	Functional	Expected
DUR-38	Prompt the user to ask questions or clarify doubts if something is unclear.	Functional	Expected
DUR-39	Fake News Alerts It should flag potential misinformation. A reference	Functional	Could have

	list should always be available to back up the information provided.		

Appendix C: Information Sheet

INFORMATION SHEET

Project: ALFIE: Assessment of Learning technologies and Frameworks for Intelligent and Ethical AI.

The objective of this interview/focus group/survey is to get feedback on specific needs and requirements of the end users (i.e. customers that want to deploy some AI solution), on which the applications and software solutions proposed by the project will be based.

This will allow us to identify the needs of diverse audiences and research how the quality of experience and the quality of the service can be improved. It will also help towards drafting requirements to design AI technologies from a user-centered approach.

This idea is new and experimental, so we need to understand what works – and what could be better. For that reason, we will gather information about:

- What happens in the interview/focus group/survey in relation to the project
- The people who take part
- What they want to get from it
- Their experience in the interview/focus group/survey in relation to the project

During this session you will be asked to provide some demographic data about yourself namely your gender, age group, educational background or country of residence. Then, you will be asked more questions and give your opinion on various aspects. No recording of the session will take place. During the session, notes will be taken to gather your feedback and comments. At the end of the session, these comments and feedback will be read aloud for you to approve. These will be considered the result of the interview/survey/focus group. Please ask as many questions as needed to clarify the procedure.

If you encounter any type of discomfort, you can stop at any time without prior justification.

Now, please read the consent form.

Appendix D: Informed consent form (over 18s)

ALFIE: Assessment of Learning technologies and Frameworks for Intelligent and Ethical AI

Alfie is a European project under the Horizon program (grant Agreement number 101177912) led by Catalink Limited.

Please read this consent form carefully before deciding whether to take part in this study.

Purpose of the research

In this focus group/questionnaire/interview, we want to talk about your habits in the use of technology. This will allow us to identify the needs of diverse audiences and research how the quality of experience and the quality of the service can be improved. It will also help towards drafting requirements to design AI technologies from a user-centred approach.

What participation in the study involves

During this focus group/questionnaire/interview you will be asked to provide some demographic data namely your gender, age group, educational background or country of residence. Then, you will be asked more questions and give your opinion on various aspects. These will be considered the result of the focus group/questionnaire/interview. Please ask as many questions as needed to clarify the procedure.

Duration

The survey takes around ___ minutes.

Risks and benefits

Your participation involves no risks of any kind.

Compensation

In this case, no compensation is envisaged for taking part.

Confidentiality

If you decide to take part, your identity will remain confidential and only members of the research team will have access to the project data. Pseudonyms will always be used if case studies need to be presented.

This informed consent form will be kept in a safe place by the principal investigators and will be destroyed five years after the end of the project. When the project is over and all data have been analysed, the whole database will be anonymised and made available to other interested researchers.

Voluntary participation

Participation in this study is completely voluntary. There is no penalty for opting not to take part.

Right to withdraw from the study

You can withdraw from the study at any time without giving explanations and with no negative consequences: just by letting us know through any communication channel. As well as this, you can exercise your rights under the European General Data Protection Regulation by making a request to Estella Oncins, estella.oncins@uab.cat, enclosing a photocopy of your ID document. Request forms for this purpose are available on the website of the UAB Data Protection Office (https://www.uab.cat/web/coneix-la-uab/itineraris/proteccio-de-dades/drets-de-les-persones-interessades-1345764799916.html).

You may also file a claim before the Catalan Data Protection Authority (**https://apdcat.gencat.cat/ca/contacte**), or contact the UAB data protection officer (proteccio.dades@uab.cat).

In all cases you will receive a written response within the legal time limit, stating what action has been taken.

Subsequent publication/re-use/other processing of the basic data and conservation period

Research data will be made available in anonymised form to other researchers after 5 years from the end of the project. Personal identifiers will be securely kept confidential and destroyed after 5 years of the finalization of the research.

Recordings and use of contributions made

__I agree to the interview being recorded (audio/video) for research purposes.

__I consent to my contributions being quoted literally with no mention of my name and to audio or video recordings of my statements being played back with no mention of my name.

__I consent to the use of my contributions in audio or video recordings for purposes of scientific dissemination, provided steps are taken to safeguard my privacy.

Contact person

If you have any queries, you can contact Estella Oncins at estella.oncins@uab.cat.

Consent

- I have read the information about the research project and I have had the opportunity to ask questions, which have been answered to my satisfaction.

- I understand that the anonymised information (with no personal identifiers) on this project will be placed at the disposal of other researchers some time after the project has ended.

- I agree to take part and I have received a copy of this consent form.

Full name of the participant ______________________________________

Signature ______________________________________ **Date:** _______________

Researcher:

Signature ______________________________________ **Date:** _______________

Appendix E: General video release consent form

By signing this,

- I allow the partners of the Alfie project to make photo, video and sound recordings of me in the light of my image right and in accordance with the privacy notice included on the next page. This is completely voluntary and up to me. I am aware that my photo, video and sound recordings may be used as part of Alfie materials and dissemination.

- I hereby grant Alfie a worldwide, royalty-free, non-exclusive, non-transferable and non-sublicensable license to make the recording(s) and to copy and communicate the recording(s) and the intellectual property rights included therein (performance rights and possibly copyrighted material as included in the recording, e.g. slides), in any known format, for strictly dissemination, training, educational or scientific purposes, for the duration of the intellectual property rights.

Signature:

Printed name:

Date:

Appendix F: Questionnaire

- How do you understand trustworthy AI in your practice?
- Which mechanisms have you established to deal with ethical issues when developing or studying AI solutions?
- Describe the specific process you follow when you or your colleagues approach some ethical issues in your Research & Development (R&D) practice.
- How do you understand bias in AI in your practice?
- What bias do you think will be most significant in your R&D work in ALFIE?
- What specific features or design elements would you prioritise to create an intuitive and accessible ALFIE platform (ETD-Hub and AutoML), and why?
- What ethical considerations should be prioritised in the development and deployment of ALFIE platforms (ETD-Hub and AutoML), and why?
- We are looking to expand our network of external stakeholders in the AI space using the snowball technique. Could you recommend three initiatives working on trustworthy AI that you believe would be valuable?

Appendix G: Template to collect feedback during the World Cafè methodology

Auto ML table

Guiding questions	Group 1 (15 minutes)	Group 2 (15 minutes)	Group 3 (15 minutes)
Participants key roles and background.			
As a potential user of AutoML, what features and functionalities would you like to see to better meet your needs?			
What types of explanations or visual displays would help people trust and use AutoML models responsibly?			
In your opinion, what steps can we take to make AutoML more accessible and intuitive?			

ETD HUB table

Guiding questions	Group 1 (15 minutes)	Group 2 (15 minutes)	Group 3 (15 minutes)
Participants key roles and background.			
As a potential user of ETD Hub, what features or functionalities would you like to see to better meet your needs?			

Use cases table

Note: Make sure you cover all the use cases (ideally, 5 minutes per use case)

Guiding questions	Group 1 (15 minutes)	Group 2 (15 minutes)	Group 3 (15 minutes)
Participants key roles and background.			
What ethical concerns could arise in the three technical use cases?			
What legal or regulatory challenges could these use cases face?			

Appendix H: Agenda of the First Stakeholders Workshop

ASSESSMENT OF LEARNING TECHNOLOGIES AND FRAMEWORKS FOR INTELLIGENT AND ETHICAL AI

Invitation to ALFIE workshop

Shaping the Future of Ethical AI

Date: 3 June 2025

Time: 10:30 – 13:00

Location: BOSCH Spain Headquarters, Madrid, Spain

You are warmly invited to the upcoming workshop for the Horizon EU project ALFIE (Assessment of Learning technologies and Frameworks for Intelligent and Ethical AI) , which is scheduled to take place on the 3nd of June 2025 at BOSCH Spain Headquarters (Calle de los Hermanos Garcia Noblejas 19 | 28037 Madrid | SPAIN).

ALFIE is an EU project aimed at ensuring AI technologies are developed and assessed in ways that consider potential bias, promote ethical standards, and ensure accessibility.

As part of the ALFIE project this workshop promises a thoughtfully curated agenda designed to facilitate meaningful **collaboration** and enhance **knowledge-sharing** among diverse stakeholders, like you, on key ethical, legal, and societal aspects of AI.

We greatly value your **insights**, **feedback**, and **recommendations**, which are essential for shaping the development and broader impact of the ALFIE project. Your contributions will be crucial in strengthening our commitment to **ethical AI** and ensuring that the project upholds the highest standards of social responsibility.

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.

ALFIE Workshop Agenda

Workshop Agenda

Time	Activity
10:30 – 10:40	Welcome and Introduction
10:40 – 11:10	The ALFIE Ecosystem: ETD-Hub and AutoML
11:10 – 11:20	Coffee break
11:20 – 11:30	Introduction to Trial Use Cases (TUCs)
11:30 – 11:40	Q&A
11:40 – 11:50	Coffee break
11:50 – 12:50	Discussion and Feedback Sessions
12:50 – 13:00	Summary and Next Steps
13:00 – 14:00	Lunch Break
14:00 – 15:00	Bosch Tour

Additional Information

- **Refreshments** and small bites will be provided during the break.
- **Recording/Documentation:** The session may be recorded for internal purposes.
- **Certification:** A certificate of participation is available upon request.

We kindly request you to confirm your attendance by completing the RSVP form available here by May 29th.

We look forward to your positive response and the opportunity to collaborate with you. Should you have any questions or require additional information, please do not hesitate to contact us at events@alfie-project.eu.

Thank you for considering our invitation, and we hope to see you on June 3nd.

The ALFIE team

2