CERTAIN

Certification for Ethical and Regulatory Transparency in Artificial Intelligence

D3.2: ETHICAL CONSIDERATIONS FOR AI COMPLIANCE

Co-funded by the European Union

Project funded by

Schweizerische Eidgenossenschaft
Confédération suisse
Confederazione Svizzera
Confederaziun svizra

Swiss Confederation

Federal Department of Economic Affairs, Education and Research EAER
State Secretariat for Education, Research and Innovation SERI

Work package	WP3
Task	T3.2
Due date	31/12/2025
Submission date	23/12/2025
Deliverable lead	UT
Version	1.0
Authors	Yasaman Yousefi (DEX), Fenia Giannakopoulou (ANAD), Anna Rizzo (DEX), Camilla Ravot Licheri (DEX)
Reviewers	Fabian Kovac (SPU), Eleni Mangina (UCD)
Abstract	This deliverable advances CERTAIN's commitment to building trustworthy, inclusive, and ethically robust AI systems by operationalising ethical and gender governance structures aligned with EU regulations. It translates high-level principles such as fairness, transparency, accountability, and privacy into practical tools, including self-assessment checklists, gender audits, risk registers, and monitoring processes. A dedicated ethical self-assessment checklist is the result of this work, which ensures that inclusivity is embedded throughout the AI lifecycle, preventing bias and digital divides. By combining legal, ethical, technical, and gender perspectives, the framework equips CERTAIN's consortium with concrete mechanisms for compliance and certification readiness, while reinforcing Europe's ambition to make trustworthy, auditable, and human-centric AI the standard.
Keywords	High-level principles, practical tools, compliance, certification readiness, human-centric AI

Document Revision History

Version	Date	Description of change	List of contributor(s)
V1.0	23/12/2025	Final editing and deliverable submission	Yasaman Yousefi (DEX), Fenia Giannakopoulou (ANAD), Anna Rizzo (DEX), Camilla Ravot Licheri (DEX)

Reviewer's comments

Name of reviewer: Irina Xezal Urumova	**Affiliation:** INCOM	**Date of review:** 31/10/2025

- Recommended insights during the internal WP3 discussions that helped form the deliverable.
- Recommended separating overlapping material by placing it in a dedicated contextual analysis section, while keeping the sections describing the project's methodology clear and concise.
- Suggested reordering certain sections to improve clarity and overall flow.
- Recommended revising this section to clearly explain how the Gender Audit Tool will be applied for the purposes of this exercise and how it adds value; a comparative analysis of existing approaches would be better placed elsewhere.
- Suggested making the text more succinct and less theoretical, with a stronger focus on purpose, scope, and added value.

Name of reviewer: Victoria Sovatzoglou	**Affiliation:** INCOM	**Date of review:** 31/10/2025

- Recommended insights during the internal WP3 discussions that helped form the deliverable.
- Commented: The document is solid and well-grounded in the literature. To enhance clarity, flow and reader-friendliness, they suggested making the presentation more concise and less theoretical, clearly separating the contextual analysis from the explanation of the methodological approach, highlighting its distinct features and added value.

Name of reviewer: Fabian Kovac	**Affiliation:** USTP	**Date of review:** 10/12/2025

- Recommended grammatical and language revisions to improve clarity and readability.
- Commented: The executive summary is well-structured and effectively communicates the deliverable's purpose and contributions. The four practical instruments are clearly articulated. Consider adding a brief mention of how this deliverable relates to the project timeline.
  The scoring system and color-coded risk matrix are valuable contributions. The four-point scale is appropriately calibrated. Consider adding example scenarios to illustrate how scores translate into risk levels.
  The QETAM model is innovative. The BIU formula is clearly presented (after the proposed changes). However, the section could benefit from a worked example showing how the formula would be applied in practice. In summary, the deliverable is of high quality and achieves its stated objectives. The identified issues are primarily editorial (grammatical corrections) and some minor structural suggestions that would enhance clarity. No fundamental methodological or content revisions are required.

Name of reviewer: Eleni Mangina	**Affiliation:** UCD	**Date of review:** 09/12/2025

- Recommended to justify why CERTAIN uses QETAM
- Suggested to clarify operationalisation per pilot & WP
- Suggested to clarify the use of tool were demonstrated briefly with an example scenario from one CERTAIN pilot or a hypothetical case.
- Suggested to describe the existing governance mechanisms of the project.

Grant Agreement No: 101189650 | **Topic:** HORIZON-CL4-2024-DATA-01-01
Call: HORIZON-CL4-2024-DATA-01. | **Type of action:** HORIZON-IA

DISCLAIMER

Co-funded by the European Union

Project funded by
Schweizerische Eidgenossenschaft
Confédération suisse
Confederazione Svizzera
Confederaziun svizra
Swiss Confederation
Federal Department of Economic Affairs, Education and Research EAER
State Secretariat for Education, Research and Innovation SERI

The CERTAIN project received funding from the European Union's Horizon Europe Research and Innovation Programme under Grant Agreement No 101189650. Views and opinions expressed are, however, those of the author(s) only and do not necessarily reflect those of the European Union. Neither the European Union

nor the granting authority can be held responsible for them. This work has received funding from the Swiss State Secretariat for Education, Research and Innovation (SERI).

COPYRIGHT NOTICE.

Project funded by the European Commission in the Horizon Europe Programme		
Nature of the deliverable:	R	
Dissemination Level		
PU	Public, fully open, e.g. web (Deliverables flagged as public will be automatically published in CORDIS project's page)	x
SEN	Sensitive, limited under the conditions of the Grant Agreement	
Classified R-UE/ EU-R	*EU RESTRICTED under the Commission Decision No2015/ 444*	
Classified C-UE/ EU-C	*EU CONFIDENTIAL under the Commission Decision No2015/ 444*	
Classified S-UE/ EU-S	*EU SECRET under the Commission Decision No2015/ 444*	

* *R: Document, report (excluding the periodic and final reports)*
DEM: Demonstrator, pilot, prototype, plan designs
DEC: Websites, patents filing, press & media actions, videos, etc.
DATA: Data sets, microdata, etc.
DMP: Data management plan
ETHICS: Deliverables related to ethics issues.
SECURITY: Deliverables related to security issues
OTHER: Software, technical diagram, algorithms, models, etc.

TABLE OF CONTENTS

EXECUTIVE SUMMARY

Deliverable 3.2 - Ethical Considerations for AI Compliance (DEX, M12, R, PU) advances CERTAIN's commitment to developing trustworthy, inclusive, and ethically aligned AI systems. It translates ethical and legal principles into practical governance mechanisms that support compliance with European regulations and standards, most notably the European Union's *Artificial Intelligence Act* (AI Act) and its underpinning standards, and prepares the project for future certification.

The purpose of this deliverable is to move from abstract values such as fairness, transparency, accountability, privacy, and inclusiveness, towards concrete, auditable practices. It does so by establishing a governance framework that integrates ethics and gender perspectives across the AI lifecycle: design, development, deployment, monitoring, and post-market oversight.

It combines ethical theory, regulatory requirements, and operational tools. It draws on international and European frameworks and aligns itself with CERTAIN's work under D3.1 with the AI Act's requirements on human oversight, technical robustness, transparency, data governance, fairness, societal well-being, and accountability.

The added value of this deliverable lies in the development of four practical instruments that make AI ethics actionable:

- ***Ethical Self-Assessment Checklist***, used by partners to evaluate ethical maturity through a scoring system and risk matrix.
- ***Gender Audit Tool***, which assesses how gender equality is embedded in project practices.
- ***Ethical and Gender Risk Register*** to record, assess, and monitor ethical and gender-related risks.
- ***Sex And Gender Impact Assessment*** **(SGIA)** framework to systematically integrate gender into design, data, user research, and technology acceptance.

Alongside these tools, the deliverable defines processes for risk classification, continuous monitoring, and compliance documentation. This ensures that ethical and gender-related risks are identified early, managed transparently, and recorded in a way that supports audits, accountability, and conformity assessment under the AI Act.

By operationalising ethics and gender governance, this deliverable provides CERTAIN with a practical structure for ethical oversight, regulatory alignment, and long-term certification readiness. It enables the consortium to make its value commitments measurable, demonstrable, and auditable.

LIST OF FIGURES

LIST OF TABLES

ABBREVIATIONS

AI	Artificial Intelligence
AI Act	European Union Artificial Intelligence Act
ALTAI	Assessment List for Trustworthy Artificial Intelligence
BIU	Batch Index Unit
EC	European Commission
EAD	Ethically Aligned Design (IEEE framework)
EIGE	European Institute for Gender Equality
EU	European Union
FC	Facilitating Conditions
FRIA	Fundamental Rights Impact Assessment
GANs	Generative Adversarial Networks
GDPR	General Data Protection Regulation
GI	Gender Influence (parameter in QETAM model)
GIA	Gender Impact Assessment
HLEG	High-Level Expert Group on Artificial Intelligence
HRIA	Human Rights Impact Assessment
HUDERIA	Human Rights, Democracy and Rule of Law Impact Assessment for AI Systems
IEEE	Institute of Electrical and Electronics Engineers
PEOU	Perceived Ease of Use
PT	Perceived Trust
PU	Perceived Usefulness
QETAM	Conceptual model combining TAM/UTAUT with gender parameters (full name not provided in text)
SGIA	Sex and Gender Impact Assessment
SI	Social Influence
STEM	Science, Technology, Engineering, and Mathematics
TAM	Technology Acceptance Model
TCP	Transmission Control Protocol
UTAUT	Unified Theory of Acceptance and Use of Technology
WP	Work Package

1. INTRODUCTION

1.1. Introduction and purpose of the deliverable

CERTAIN is committed to building trustworthy, lawful, and human-centric AI systems. As part of this mission, *Work Package 3* (WP3) plays a foundational role in operationalising ethical, legal, and societal values across the project lifecycle. Within WP3, ***Task 3.2 - Ethical Requirements*** focuses specifically on defining the ethical and gender implications of the project, and on **establishing a practical framework to ensure compliance with ethical standards, fairness requirements, and gender equality** objectives.

This deliverable builds directly on ***Deliverable 3.1 – Legal requirements (UT, R, PU, M11)***, which established **the comprehensive legal and regulatory mapping for CERTAIN**. D3.1 defines the **compliance boundaries, obligations, and risk categories under EU law** that serve as the **normative foundation** for the ethical and gender-governance methods developed in D3.2. As such, the ethical tools, checklists, and governance mechanisms presented in this report are designed to operationalise and extend the legal requirements identified in D3.1 into actionable ethical practice.

This deliverable presents the first version of the ethical and gender governance method underpinning CERTAIN. It contributes to the project's broader goals by:

- **Establishing foundational ethical and gender principles** tailored to AI and data-driven environments;
- **Developing self-assessment tools and checklists** for project partners as well as other interested individuals or organisations to evaluate their activities against these principles;
- **Designing a gender audit methodology**, including readiness metrics and application scenarios;
- **Proposing monitoring and reporting structures** for continuous ethical oversight;
- **Aligning with key *European Union* (EU) policy and regulatory instruments**, including *the* ***Ethics Guidelines for Trustworthy AI***, the ***Artificial Intelligence Act*** **(AI Act)**, and emerging **CEN-CENELEC standards** on trustworthy and auditable AI, in which DEXAI is actively involved.

By grounding the CERTAIN approach in concrete, actionable tools and governance procedures, this deliverable supports the project's contributions to **AI certification readiness**. It provides ethical foresight to ensure that compliance addresses technical parameters on top of structural fairness, inclusiveness, and gender-sensitive design, in alignment with EU values and fundamental rights.

1.2. Supporting certification, compliance, and trust

Fostering **trustworthy AI** involves both **technical robustness and legal conformity**, as well as **embedding ethical principles into the systemic design of AI solutions**. As regulatory landscapes in the EU evolve, most notably through the AI Act and related standardisation efforts, compliance and certification emerge as instruments of institutional trust, reputation, and legitimacy.

In this context, this deliverable supports that aim by laying the groundwork for ethical and gender-informed governance structures that can be meaningfully integrated into the processes of risk management, conformity assessment, and third-party audit. The development of ethical self-

assessment checklists, gender audit tools, and monitoring procedures presented herein is thus intended as a prototype for future certification regimes capable of recognising the multidimensionality of trust in AI systems.

Certification, in the frame of CERTAIN, is approached as both a **technical and ethical act**, a means of demonstrating adherence to verifiable requirements and values such as fairness, transparency, non-discrimination, and gender equality. Ethical readiness for certification entails the anticipation of downstream risks, the demonstration of procedural integrity, and the establishment of documentation practices that make value alignment legible and auditable. The deliverable, therefore, aims to bridge the gap between normative aspirations and operational procedures, proposing a framework where compliance is seen as a mode of ethical governance.

In line with this vision, the tools and structures developed here will support CERTAIN's efforts to contribute to the **translation of high-level ethical and legal principles into certifiable practice**, reinforcing public trust in AI while advancing the project's scientific, regulatory, and societal impact.

The ethical aspect of the certification-readiness approach outlined here follows the legal compliance framework established in D3.1. While D3.1 clarifies the statutory obligations for providers, deployers, and data holders under the AI Act and related EU legislation, D3.2 translates these obligations into concrete ethical and gender-responsive governance procedures. Together, the two deliverables form a coordinated pathway from legal compliance (D3.1) to ethical assurance (D3.2).

1.3. Methodological approach

The development of trustworthy AI systems in the European context requires synthesising three critical areas: **(1) foundational ethical principles**, **(2) mandatory legal compliance mechanisms**, and **(3) rigorous, operationally focused design methodologies**. The overall methodology for this deliverable is built upon an integrated, multi-layered framework designed to translate high-level ethical and legal principles into **concrete, auditable governance practices** throughout the AI system lifecycle. The approach is fundamentally socio-technical, recognising that algorithmic fairness is an inherently social construct and that AI development pathways risk replicating or amplifying existing structural biases, particularly those related to gender. Consequently, the methodology synthesises legal mandates, such as those established under the **EU AI Act**, with academic critical fairness frameworks to ensure that compliance functions not merely as a regulatory requirement, but as an **active mode of ethical governance**.

A landmark comparative study by Fjeld and others maps the convergence of thirty-six influential AI ethics principles authored by governments, private companies, advocacy groups, and multilateral institutions[i].

The report identifies **eight thematic clusters**: *Privacy, Accountability, Safety and Security, Transparency and Explainability, Fairness and Non-Discrimination, Human Control of Technology, Professional Responsibility*, and the *Promotion of Human Values*, forming the **emerging normative core for responsible AI governance**. These themes, reinforced by human-rights-based frameworks such as the **Toronto Declaration**[1], underscore the shared global recognition that AI systems must be developed in ways that safeguard individual rights and promote social justice. However, Fjeld et al. also highlight significant implementation gaps between principle formulation

[1] See https://www.torontodeclaration.org

and operationalisation: an insight that directly informs CERTAIN's focus on **translating ethical foresight into actionable practice**.

Building on this foundation, CERTAIN integrates and operationalises insights from multiple **international and European ethical governance frameworks**, including the **OECD AI Principles (2019)** [2], the **UNESCO Recommendation on the Ethics of AI (2021)** [3], and the **EU Ethics Guidelines for Trustworthy AI** [4]. These frameworks collectively emphasise human agency, transparency, fairness, accountability, and societal well-being as the cornerstones of trustworthy AI. The AI HLEG's guidelines are complemented by the ***Assessment List for Trustworthy AI*** **(ALTAI)** [5], which provides a practical self-assessment mechanism. CERTAIN adopts and extends this model through a tailored **Ethical Self-Assessment Checklist**, designed both for project partners and accessible to any external stakeholder interested in evaluating their activities against ethical and gender-responsive AI principles.

Complementary methodologies such as the **IEEE *ethically aligned design* (EAD)** framework [6] and the **Z-Inspection® process** [7] further inform CERTAIN's operational design. The EAD framework provides detailed engineering guidance for embedding ethical values (transparency, agency, and accountability) throughout the AI lifecycle, while Z-Inspection® offers an iterative, multidisciplinary methodology for evaluating ethical tensions in real-world deployments, particularly in high-impact contexts such as healthcare. Together, these tools bridge the gap between abstract ethical aspirations and implementable technical standards, providing developers and decision-makers with structured pathways for **ethically aligned innovation**.

In addition, CERTAIN aligns with **rights-based assessment frameworks** that extend beyond traditional data protection toward broader human-rights considerations. The ***Human Rights Impact Assessment*** **(HRIA)**[8], and the ***Fundamental Rights Impact Assessment*** **(FRIA)**, as required under **Article 27 of the EU AI Act** for high-risk systems, provide a model for identifying and mitigating potential rights violations. Notable examples such as Council of Europe's **HUDERIA** (***Human Rights, Democracy and Rule of Law Impact Assessment for AI Systems***, developed by the Alan Turing Institute for the Council of Europe) [9] and **FRAIA (*Fundamental Rights and Algorithms Impact Assessment,*** developed by the Dutch government)[10] illustrate how risk-based and participatory approaches can ensure that human rights and democratic values are systematically considered in AI design and deployment.

To ensure these ethical and legal principles are not confined to the theoretical level, the CERTAIN methodology embeds them directly into the project's design and governance workflow through structured participatory tools. Drawing inspiration from operational frameworks discussed above, like **ALTAI**, **Z-Inspection®**, and the **IEEE EAD**[2], as well as rights-based approaches such as **HRIA** and **FRIA**, the methodology emphasises inclusivity, reflexivity, and stakeholder participation. This integration ensures that ethical reflection is not a one-off event but a continuous, collaborative process accessible to all interested parties. Through this self-assessment tool and transparent reporting mechanism, CERTAIN promotes a culture of shared responsibility and ethical literacy, extending engagement beyond project partners to include external researchers, civil society, and policymakers.

The methodology is proactive and rooted in an **ex-ante ethical design philosophy** [11], employing specialised tools to anticipate and mitigate risks before deployment. A core component is the ***Sex and Gender Impact Assessment*** **(SGIA)**, which provides a structured, gender-sensitive lens over the entire research and development process. Drawing on the ***European Institute for Gender Equality*** **(EIGE)** ***Gender Impact Assessment*** **(GIA)**, this process involves defining the system's purpose, checking its gender relevance, and conducting gender-sensitive analyses to identify

[2] See https://standards.ieee.org/wp-content/uploads/import/documents/other/ead_general_principles.pdf

potential structural inequalities in data, design assumptions, and access. These qualitative assessments are complemented by the ***Ethical Self-Assessment Checklist***, which employs a quantifiable risk matrix and a four-point scoring system to evaluate preparedness and prioritise mitigation actions based on the likelihood and severity of identified ethical and compliance risks.

For accountability and formalised risk management, the core of the governance structure is the ***Ethical And Gender Risk Register***. This structured instrument enables the systematic identification, classification, and monitoring of risks such as dataset bias, user exclusion, or discriminatory design patterns, across social, legal, technical, and gender-specific dimensions. The register formalises the recording of risk ratings, mitigating measures, and clearly assigned responsibilities for risk treatment, ensuring that commitments to non-discrimination, fairness, and transparency are **traceable, managed, and auditable**.

The methodological design of D3.2 is intentionally aligned with D3.1, which provides the authoritative interpretation of the EU regulatory landscape, including the GDPR, the Data Act, Digital Services Act, Digital Markets Act, Cybersecurity Act, Cyber Resilience Act, and the AI Act's risk-based obligations.[3]

Accordingly, the ethical tools developed here build on the legal foundations and risk categories defined in D3.1, ensuring that ethical assessment, gender auditing, and monitoring processes align coherently with the legal obligations of AI actors in CERTAIN.

The methodology culminates in a cycle of **continuous monitoring and compliance**, ensuring that ethical integrity is maintained throughout the system's operational phase. This includes periodic risk reviews, iterative impact assessments, and reflexive evaluation of emergent biases or unintended consequences as systems interact with real-world contexts. All monitoring activities and corrective actions are comprehensively documented, forming the **evidence base for external audit, conformity assessment procedures, and certification readiness**, thereby directly linking CERTAIN's internal governance mechanisms with the external regulatory expectations of the EU AI Act.

1.4. Structure of the deliverable

The deliverable aims to provide a comprehensive and actionable framework for supporting ethical compliance in CERTAIN, while contributing to the long-term goal of aligning AI development with European values, legal standards, and social expectations. To do so, it follows a layered progression from conceptual foundations to practical tools and implementation strategies, responding to the need for an integrated framework that is ethically robust and procedurally applicable.

Section 2 surveys relevant **ethical governance frameworks**, offering a comparative analysis of existing European and international guidelines, with particular attention to their alignment with the regulatory landscape in the EU. This section establishes the normative baseline upon which the CERTAIN ethical requirements are built.

Section 3 sets out the core **ethics and gender assessment guidelines**, articulating the project's foundational ethical values, operationalised through an Ethics-by-Design toolkit. It introduces the gender audit tools and self-assessment instruments developed to assist partners in embedding

[3] As noted in D3.1, the regulatory proposal known as the Digital Omnibus may add uncertainty to the EU regulatory landscape. While none of its proposed amendments currently affect the issues addressed in this deliverable, its scope and potential impact, particularly regarding data governance, AI training practices, and enforcement structures, could influence compliance pathways.

ethical foresight and gender responsiveness throughout their activities. The section concludes by mapping ethical and gender dimensions to future certification processes.

Section 4 introduces the **risk management and monitoring** infrastructure, including a typology of ethical and gender risks, a preliminary version of the CERTAIN risk register, compliance indicators, and the design of internal monitoring dashboards. This section also anticipates future integration with project-wide quality assurance and audit mechanisms.

Section 5 is dedicated to the ***Sex And Gender Impact Assessment*** **(SGIA)** framework. It outlines the methodological foundations, metrics, and indicators for assessing gender-related implications across different project domains and provides illustrative application scenarios to support contextual interpretation.

Section 6 presents the **CERTAIN ethical guidelines**, which are a synthesis of the deliverable's preceding components, adapted to the technical and organisational structure of the project. It outlines how ethical oversight, monitoring responsibilities, and escalation procedures will be coordinated across partners and phases of development.

Section 7 defines the **implementation and integration plan**, specifying roles, responsibilities, and timelines for the operational deployment of ethical and gender tools within the broader CERTAIN governance ecosystem.

Section 8 concludes the deliverable by reflecting on the systemic relevance of the developed tools and processes and outlining future steps for continuous improvement and responsive adaptation as the project progresses.

2. Ethical Governance Frameworks

2.1. Why govern ethically?

In this section, we explore **how ethical guidelines, especially those developed within the European Union, function as an intermediary governance tool in this shifting landscape**. Understanding this requires engagement with AI as a technical domain, as well as with the societal and normative contexts in which it operates. For instance, machine learning systems learn from data that reflect historical and structural patterns, which may reproduce social biases or cause new and unaccounted-for issues. This interaction underscores the importance of **approaching AI governance as a socio-technical issue** and positions ethics guidelines not as peripheral commentary but as central instruments in shaping AI development pathways. By examining the role and limits of ethical guidelines, this section contributes to an ongoing discussion on how best to regulate AI in a way that is principled, context-sensitive, and future-facing.

As AI advances, the imperative to govern its societal impacts becomes increasingly urgent. Whether manifesting as personalised recommendations, automated credit scoring, predictive policing, or diagnostic tools, **AI systems are no longer confined to technical domains; they actively mediate human experiences, influence institutional decisions, and shape social outcomes**. With this expanding reach, AI has come to exemplify a class of socio-technical systems whose effects are deeply normative.

AI systems promise **efficiency and innovation while raising ethical concerns around opacity, bias, discrimination, and loss of human agency**. These tensions have led to increasing demands for frameworks that can guide the responsible development and deployment of AI, especially within the European context[12].

In this light, the European Commission launched a coordinated AI strategy in 2018, which included the establishment of a ***High-Level Expert Group on AI*** **(AI HLEG)**. This group was tasked with advancing a normative foundation for AI grounded in EU values and rights. In 2019, the AI HLEG issued its ***Ethics Guidelines for Trustworthy AI***, articulating a set of principles, such as *transparency, accountability,* and *human-centricity*, meant to operationalise ethical concerns beyond legal compliance[13]. Although the guidelines expressly refrain from offering legal advice, they play an important role in setting ethical expectations within the AI ecosystem.

At the centre of the guidelines lies the concept of *trustworthy AI*, which rests on three core elements:

- **Lawful** – AI systems must comply with existing laws and regulations. As technology evolves, legal compliance becomes critical to safeguarding rights, managing risks, and maintaining public trust.
- **Ethical** – Beyond legal obligations, AI should respect human dignity, uphold fundamental rights, and contribute positively to society.
- **Robust** – AI must perform reliably and securely in real-world conditions, remaining resilient to errors, misuse, and adversarial attacks. This includes both technical robustness and broader social reliability.

To translate these principles into practice, the guidelines outline **seven key requirements** for trustworthy AI:

1. **Human agency and oversight** – AI should support human decision-making and preserve autonomy, with mechanisms ensuring meaningful human control.

2. **Technical robustness and safety** – Systems must be secure, resilient and capable of handling errors or attacks, especially in high-risk sectors.
3. **Privacy and data governance** – Personal data must be protected through responsible collection, storage, and processing in line with GDPR and data protection standards.
4. **Transparency** – AI systems should be understandable. This includes clarity around their purpose, functionality, and decision-making processes.
5. **Diversity, non-discrimination and fairness** – AI must be inclusive, accessible, and free from unjust bias to ensure equal treatment for all individuals and groups.
6. **Environmental and societal well-being** – AI should contribute positively to society and minimise environmental impact, supporting sustainability and social good.
7. **Accountability** – Clear responsibility and oversight mechanisms must be in place to address potential harms, ensure traceability, and enable redress.

These guidelines reflect the EU's commitment to shaping a future where AI strengthens societal well-being and human rights. By combining legal compliance, ethical responsibility and technical reliability, they offer a practical roadmap for policymakers, developers and organisations striving to build trustworthy, human-centric AI.

The emergence of these and similar ethical frameworks, from governmental bodies, research institutions, and private actors, shows a broader shift in AI governance toward principles-based regulation. Yet, despite their normative clarity, such guidelines often lack mechanisms for enforcement or concrete implementation. This gap raises critical questions regarding their effectiveness and their relationship to binding legal norms.

At a political level, the centrality of ethics to European AI governance is further evidenced by the **2020 *White Paper on Artificial Intelligence***, which reaffirmed the Commission's commitment to fostering "an ecosystem of trust"[14]. The challenge, however, lies in translating aspirational values into actionable requirements and verifiable procedures. The dynamic nature of AI technologies exacerbates this problem: regulation tends to evolve more slowly than innovation, resulting in a temporal disconnect between technological capability and normative oversight [15].

AI governance is a challenge. At the heart of this challenge lies the recognition that these systems learn from the world and replicate what is fed to them through data. For instance, ML systems adapt based on patterns in training data, which are themselves embedded in histories of social inequality, cultural bias, and structural exclusion. As a result, outputs may reflect, reproduce, or even amplify existing forms of discrimination, whether along the lines of gender, race, ethnicity, or socioeconomic status. This phenomenon renders AI a potentially recursive agent of social stratification, even unintentionally.

Empirical studies have illustrated the extent of these risks. Image recognition systems trained on culturally homogenous datasets exhibit markedly lower accuracy for underrepresented groups[16]; automated decision systems used in judicial contexts, such as COMPAS, have demonstrated racial bias in predicting recidivism, and facial recognition algorithms have shown significantly higher error rates for women and black people[17]. In the employment sector, algorithmic job recommendation engines have exhibited gendered salary suggestions, disadvantaging female users[18]. These examples underscore a broader pattern: AI systems, when trained on imbalanced data or deployed without ethical safeguards, risk perpetuating societal harms.

Beyond discriminatory outcomes, the opacity of AI processes is an additional governance concern. Many models, particularly those employing deep neural networks, function as “black boxes” whose internal logic resists interpretability [19,20]. This opacity complicates efforts to assign accountability, verify fairness, or explain decisions to affected individuals, a fundamental requirement under European data protection and human rights frameworks.

In parallel, the capacity of generative models, such as *Generative Adversarial Networks* (GANs), to produce synthetic content introduces novel risks related to authenticity, misinformation, and manipulation[21]. While these technologies hold potential for socially beneficial uses (e.g., synthetic medical data for privacy-preserving research), they also pose clear risks for disinformation, identity fraud, and cultural distortion. These tensions underscore the need to govern AI in its broader *sociotechnical assemblages*, the data sources, deployment contexts, and value systems within which it operates.

Consequently, the governance of AI must extend beyond technical assurance and risk minimisation. It must interrogate what kind of social order AI systems instantiate or disrupt. For CERTAIN, this entails developing tools and guidelines that are both reactive and anticipatory, capable of detecting and addressing ethical risks before harm occurs and designed to align AI development with EU fundamental rights and democratic values.

2.2. Overview of existing frameworks

A landmark report by Fjeld et al. [22] maps the emergence of consensus across thirty-six influential AI principles documents authored by a range of stakeholders, governments, private companies, advocacy groups, and multilateral bodies. Through a comparative analysis of these normative statements, the report reveals a convergence around a core set of ethical themes, while also exposing important contextual divergences and limitations in scope, enforceability, and alignment with existing rights-based frameworks.

The analysis results in eight thematic clusters of principles: **Privacy, Accountability, Safety and Security, Transparency and Explainability, Fairness and Non-discrimination, Human Control of Technology, Professional Responsibility, and the Promotion of Human Values**, plus an auxiliary category tracking references to international human rights law. These themes together outline an emerging normative core for responsible AI governance.

There is a substantial overlap in the values and safeguards endorsed, despite differences in origin, audience, and institutional format. The themes of **Privacy, Fairness, and Accountability** appear in nearly every document studied, indicating a shared recognition that data-driven AI systems require governance mechanisms to protect individual rights and promote justice. The privacy theme, for instance, encompasses both traditional data protection concerns, such as consent and rectification, and more proactive design choices, such as “privacy by design.” Documents influenced by the GDPR and broader European data protection regimes tend to express this theme with particular granularity and legal sophistication.

Similarly, **Fairness and Non-discrimination** principles seek to ensure that AI systems do not exacerbate existing social inequalities. Here, the emphasis is on inclusive datasets, equity in design and deployment, and the mitigation of algorithmic bias. The *Accountability* theme underscores the importance of identifying responsible actors across the AI lifecycle and of developing clear avenues for redress when harms occur. This includes calls for impact assessments, auditability, and legal liability regimes. Together, these themes reflect widespread concern about the opacity, unpredictability, and asymmetrical power embedded in AI applications.

In addition to identifying thematic convergence, the report distinguishes among three structural levels at which ethical principles operate: design, monitoring, and redress. These phases correspond to different stages of the AI system lifecycle, and ethical principles can apply variously across them. For example, "impact assessments" (design), "evaluation requirements" (monitoring), and "ability to appeal" or "remedy for automated decisions" (redress) collectively shape a more comprehensive accountability architecture. Notably, some frameworks propose institutional innovations such as independent ethics boards, algorithmic oversight bodies, or data trusts to formalise and sustain these functions.

Moreover, several documents emphasise professional responsibility as a crucial enabler of ethical practice. These principles focus on the obligations of researchers, developers, and deployers to anticipate long-term consequences, uphold scientific integrity, and engage with diverse stakeholders. This theme mirrors broader trends in technology ethics that stress individual agency within institutional and socio-technical systems.

One of the most compelling contributions of the report is its engagement with the relationship between AI ethics and international human rights law. While 64% of the documents reference human rights explicitly, only a few adopt them as a comprehensive framework for governance. These include the Toronto Declaration and several civil society-led initiatives that frame AI risks in terms of the right to non-discrimination, freedom of expression, and access to remedies. The authors argue that rights-based approaches offer a legally grounded and globally legitimate foundation for AI ethics, especially in adjudicating conflicts among competing ethical imperatives.

The report also notes that many principal sets, particularly those issued by corporations, stop short of embracing binding legal obligations. Instead, they tend to frame ethics in aspirational or voluntary terms. This soft-law orientation limits the enforceability of the principles and raises concerns about ethics washing or strategic compliance. Thus, the authors emphasise the importance of embedding principles within broader governance ecosystems, including national policies, sector-specific regulation, and professional norms.

Although the authors identify a growing convergence around the eight thematic clusters, they acknowledge the cultural, political, and linguistic diversity underlying the documents. Principles drafted in Asia, for instance, often emphasise social harmony, collective well-being, and innovation leadership, while European and North American documents tend to stress individual rights, accountability, and risk mitigation. The ***Chinese White Paper on AI Standardization***[4] and the ***Japanese Principles of Human-Centric AI***[5] are notable for advocating state-led coordination and for embedding AI within broader national development strategies. In contrast, documents from civil society actors highlight distributive justice, transparency, and public participation.

These differences suggest that while ethical consensus is emerging, its articulation remains context-dependent. The normative content of "fairness" or "accountability," for example, may vary depending on the institutional setting, legal system, or political tradition in which it is invoked. Thus, the authors caution against one-size-fits-all interpretations and instead propose a pluralistic and dialogic approach to global AI governance.

A recurring theme throughout the report is the implementation gap, that is, the disjunction between principal articulation and operationalisation. Most ethical principles are not accompanied by detailed guidelines for enforcement, nor are they linked to formal accountability mechanisms. To be effective, ethical principles must be translated into specific practices, tools, and standards within the

[4] See https://cset.georgetown.edu/publication/artificial-intelligence-standardization-white-paper-2021-edition/

[5] See https://www8.cao.go.jp/cstp/english/humancentricai.pdf

institutions that design, deploy, and regulate AI. This includes integrating ethics into technical documentation, procurement procedures, certification schemes, and impact assessment protocols.

Some initiatives have already begun to address the challenge of translating high-level ethical principles into concrete operational practices. Among the most prominent is the **IEEE's *Ethically Aligned Design* (EAD) framework**, which goes beyond aspirational declarations to offer detailed guidance for practitioners across the AI development lifecycle. The EAD initiative encompasses practical tools for integrating values such as transparency, agency, and accountability into the technical architecture of AI systems. It proposes ethically grounded system requirements, offers design principles for human-machine interaction, and suggests institutional reforms to support professional responsibility, including the creation of interdisciplinary ethics boards and the embedding of ethical checkpoints in software development pipelines. In this respect, the IEEE's work reflects a sustained effort to bridge the gap between ethical foresight and engineering implementation, thereby equipping developers with a vocabulary and toolkit for ethically aligned innovation.

Similarly, the European Commission's ***Assessment List for Trustworthy AI* (ALTAI)** represents a pioneer attempt to operationalise the seven key requirements outlined in the *Ethics Guidelines for Trustworthy AI*. The ALTAI tool enables AI developers and deployers to conduct structured self-assessments of their systems, focusing on issues such as **human oversight, robustness, privacy**, and **fairness**. It functions as a compliance instrument, and an internal governance aid, encouraging organisations to engage in reflexive processes of ethical evaluation. Importantly, the tool is adaptable across sectors and allows users to contextualise ethical principles based on specific deployment scenarios, technical capabilities, and societal impacts.

While ALTAI provides a robust framework for self-assessment, more novel methodologies are necessary to offer a holistic evaluation of AI trustworthiness, particularly in complex, real-world contexts. The **Z-Inspection®** process, developed by an international and interdisciplinary team of experts, is one such methodology. It is specifically designed to assess the ethical implications of AI systems in their operational environment, making it uniquely suited for healthcare applications where a nuanced approach is required.

The Z-Inspection® process is defined by its collaborative nature and its structured, iterative approach to ethical evaluation. It begins with the identification of **ethical tensions** that arise in the deployment of AI systems, such as the trade-off between patient privacy and the need for data sharing in medical research. Once these tensions are identified, the process maps them to specific **technical and legal issues**, providing a detailed analysis of how these issues impact the overall trustworthiness of the AI system. A key feature is the involvement of a multidisciplinary team of experts, including ethicists, legal scholars, AI developers, and healthcare professionals. This team-based approach, which also includes input from stakeholders like patients and providers, ensures that the evaluation is comprehensive and considers a wide range of perspectives. The process concludes by providing **actionable recommendations** that are tailored to the specific context of deployment, ensuring that they are both practical and effective. Z-Inspection® has been successfully applied in various healthcare contexts, demonstrating its effectiveness in identifying potential biases and ensuring that AI systems support, rather than replace, human judgment.

In the context of data-intensive AI technologies, it is essential to move beyond assessments focused solely on data protection. A ***Human Rights Impact Assessment* (HRIA)** offers a more comprehensive evaluation, addressing the broader spectrum of human rights implications that may arise from the use of AI. The HRIA process begins with a preliminary screening to identify potential risks, followed by a more detailed analysis of the specific rights and stakeholders involved. This

thorough examination helps to anticipate and address potential human rights violations before they occur.

Furthermore, under Article 27 of the AI Act, high-risk AI systems are required to undergo a ***Fundamental Rights Impact Assessment*** **(FRIA)** to assess their potential impact on fundamental human rights, such as equality and health. While the CERTAIN project aims to develop limited-risk systems, the principles and structural framework of a FRIA provide an excellent model for our ethical monitoring plan. There are several available FRIAs that can inspire this approach, including:

- ***Human Rights, Democracy, and the Rule of Law Impact Assessment for AI Systems*** **(HUDERIA)**: Developed by the Alan Turing Institute for the Council of Europe, this framework provides a risk-based approach to assess an AI system's impacts on human rights, democracy, and the rule of law. It includes risk analysis, stakeholder engagement, and ongoing impact assessments [17].

- ***Impact Assessment Fundamental Rights and Algorithms*** **(FRAIA)**: A tool developed by the Dutch government, FRAIA helps to assess and mitigate risks to fundamental rights when using algorithms. It fosters dialogue between professionals and ensures that the impact on human rights is considered systematically, thereby preventing unintended consequences.
These frameworks, in combination with the overarching ethical principles, provide a robust and actionable strategy for CERTAIN to ensure that its AI solutions are both technically innovative and ethically sound, legally compliant, and aligned with human values.

However, despite these commendable efforts, structural and cultural barriers continue to impede the full integration of ethical principles into everyday AI development and deployment. In many cases, ethical assessments are still treated as add-on procedures, conducted post hoc or for reputational purposes, rather than embedded within the core design methodology. This is especially problematic in organisational contexts where commercial incentives, regulatory ambiguity, or time pressures discourage meaningful engagement with ethical reflection. Without institutionalised accountability mechanisms, such as mandatory audits, enforceable standards, or external oversight bodies, there is a risk that even the most sophisticated ethical tools will be underutilised or ignored.

Within the European Union, the ***Ethics Guidelines for Trustworthy AI*** s a foundational reference. These guidelines define trustworthy AI as resting on three interdependent components: legality, ethical soundness, and technical robustness. They outline seven key requirements, including human agency and oversight, transparency, fairness, and societal well-being. Crucially, the guidelines are accompanied by an Assessment List designed to support voluntary self-evaluation, although critics have noted the absence of binding implementation mechanisms. The AI HLEG's work has deeply influenced the CERTAIN project's orientation, particularly in shaping the design of ethics checklists, oversight tools, and partner-facing self-assessment instruments.

Other European frameworks have complemented and extended these efforts. **The AI4People initiative**, for example, proposed a set of five ethical principles (beneficence, non-maleficence, autonomy, justice, and explicability) that informed the HLEG's final recommendations. AI4People emphasised the need to translate high-level ethical values into actionable governance structures, offering concrete guidance on how institutions might embed ethics into the design and deployment of AI systems.

The **GE-RRI Toolkit (*Gender Equality in Responsible Research and Innovation*)** and the Gendered Innovations 2 policy report have been instrumental in mainstreaming gender as a critical dimension of responsible AI. These resources provide methodological tools to assess how sex and gender shape data quality, system functionality, and societal impact concerns directly

addressed in CERTAIN's gender audit tools and SGIA (*Sex and Gender Impact Assessment*) framework.

AI ethical governance, however, is far from limited to Europe. At the international level, the Organisation for Economic Co-operation and Development (***OECD) Principles on AI* (2019)** adopted by all G20 countries, offer a set of intergovernmental commitments anchored in human-centred values, robustness, transparency, and accountability. These principles are intended to guide both national policy development and transnational cooperation.

The ***UNESCO Recommendation on the Ethics of Artificial Intelligence* (2021)** takes a broader global view, addressing digital divides, environmental sustainability, and cultural diversity. It insists on the centrality of human rights, particularly in cross-border or public-sector applications.

Parallel efforts in industry and academia have further contributed to this evolving ethical landscape. The **IEEE *Ethically Aligned Design* (EAD) initiative**, developed by an international consortium of engineers and ethicists, provides standards and design principles for embedding ethical considerations throughout the AI lifecycle. Its emphasis on transparency, algorithmic agency, and human data dignity resonates with CERTAIN's technical work packages and *value-sensitive design* (VSD) approach. Likewise, the ***Montreal Declaration for Responsible AI***, shaped through broad public consultation, foregrounds participatory governance, equity, and democratic accountability.

Other efforts, such as the Asilomar AI Principles, represent early attempts by interdisciplinary experts to articulate a roadmap for safe and beneficial AI development. Their influence persists, particularly through their focus on long-term risks, cooperation, and value alignment, which are topics that remain salient today.

Beyond ethics-specific initiatives, several regulatory and integrity-focused frameworks are important to mention. The ***Horizon Europe Ethics Appraisal Procedure*** establishes a baseline for identifying and mitigating ethics and gender risks in EU-funded research. CERTAIN draws on this model through its use of structured self-assessment forms and internal ethics review triggers. The All European Academics (***ALLEA) Code of Conduct for Research Integrity*** complements this by reinforcing principles of honesty, accountability, respect, and reliability[23].

The question now is, why do these frameworks matter for us?

The international and European ethical AI frameworks reviewed above serve as operational anchors for CERTAIN's governance approach. Their relevance to the project can be summarised along three core dimensions:

These frameworks offer concrete procedural elements, such as principles of trustworthy AI (EU HLEG), value-sensitive design guidelines (IEEE EAD), and participatory assessment methods (Z-Inspection®), that can be translated into the checklists, audit tools, and monitoring structures developed in D3.2. They provide the methodological backbone for operationalising ethics into reproducible and auditable processes.

Further, many of the high-level requirements embedded in the EU AI Act (e.g., human oversight, technical robustness, transparency, fairness, and accountability) mirror the ethical foundations articulated in these frameworks. Aligning CERTAIN's ethical governance with them ensures that the project's practices are coherent with the legal obligations set out for deployers and providers under the Act.

Lastly, we should note that the development of the *Ethical self-assessment checklist*, the Gender Audit Tool, and the SGIA methodology is grounded in the insights of these frameworks, which emphasise structured evaluation, risk anticipation, and human-centric design. By drawing from globally recognised standards, the tools introduced in D3.2 gain normative legitimacy and practical

relevance, supporting CERTAIN's long-term objective of certification readiness and trustworthy AI development.

2.3. Alignment with EU regulations

This section builds on the regulatory mapping performed in D3.1, which analysed the AI Act's requirements on human oversight, transparency, data governance, and risk management. D3.2 complements this by situating these legal obligations within an ethical governance framework, ensuring that ethical assurance mechanisms are consistently anchored in the legal constraints identified in D3.1.
CERTAIN will develop its AI solutions in compliance with the seven key requirements for trustworthy AI, which serve as guiding principles for our work:

- **Human Agency and Oversight**: We ensure that human oversight is integrated into our systems to prevent the automation of tasks that require human judgment and accountability, thereby preserving human autonomy.
- **Technical Robustness and Safety**: Our systems will be technically sound, reliable, and resilient against errors, vulnerabilities, and attacks, ensuring a high degree of safety.
- **Privacy and Data Governance**: We adhere to stringent data protection regulations, including the GDPR, to safeguard user data and ensure ethical and responsible data handling throughout the AI lifecycle.
- **Transparency**: CERTAIN's AI systems will be designed to be transparent, allowing stakeholders to understand their purpose, capabilities, and limitations, as well as the data and logic that inform their outputs.
- **Diversity, Non-discrimination, and Fairness**: The project is committed to developing AI that is fair and free from bias, with an emphasis on inclusive datasets and equitable design to ensure just and impartial outcomes for all users.
- **Environmental and Societal Well-being**: We consider the broader societal and environmental impact of our AI systems, aiming to develop technologies that contribute to sustainable and beneficial outcomes for society.
- **Accountability**: CERTAIN establishes clear accountability mechanisms, ensuring that responsibility for the outcomes and impacts of its AI tools can be traced and managed, with clear avenues for redress.

CERTAIN engages proactively with existing and emerging legal frameworks. In particular, the project aligns its outputs with the EU AI Act, especially Title III (requirements for high-risk AI systems) and Annex IV (technical documentation). Through this alignment, CERTAIN's deliverables contribute to early-stage conformity assessment readiness, mapping key ethical and technical practices, such as bias testing, human oversight procedures, and transparency documentation, onto the regulatory expectations defined by the AI Act.

Taken together, these frameworks, spanning ethical theory, legal compliance, technical guidance, and gender equity, provide CERTAIN with a multi-layered governance foundation. Rather than treating them as isolated instruments, the project integrates their insights into a coherent architecture of ethical oversight. In doing so, CERTAIN aims not merely to comply with ethical norms but to operationalise them in a way that is reflexive, participatory, and aligned with both regulatory developments and societal expectations.

3. ETHICS AND GENDER ASSESSMENT GUIDELINES

3.1. Foundational values and ethical principles

The foundational values operationalised here directly reflect the duties and actor-specific responsibilities outlined in D3.1. In particular, fairness, transparency, human oversight, and accountability correspond to the legal obligations defined for providers, deployers, and other actors under the AI Act. D3.2, therefore, serves as the ethical operational layer to the compliance baseline mapped in D3.1.

The exponential development of AI has brought a heightened focus on the ethical implications of these technologies. In response, a consensus has emerged around several foundational values and ethical principles that must guide the design, development, and deployment of AI systems within the CERTAIN project. These principles are rooted in a **human-centric approach** that aims to ensure that AI serves as a "force for good" and upholds fundamental human rights and societal values. Prominent frameworks, such as the EU's White Paper on AI and the HLEG's Ethics Guidelines for Trustworthy AI, consistently prioritise a core set of values including **fairness, transparency, non-discrimination, privacy, and accountability.** Given the possibility of different interpretations surrounding these principles and the peril of "ethics shopping [24]," or "ethics bluewashing"[25], i.e., cherry-picking and adapting ethics principles to one's needs, it is crucial to rely solely on consolidated principles and interpret them consistently with the literature.

Fairness and Non-Discrimination are essential for creating AI systems that operate equitably. These principles require that AI not produce prejudiced or unjust outcomes. A significant challenge lies in the fact that algorithms can learn and even amplify existing societal biases present in their training data. Historically, this was first proven by the investigative journalism of ProPublica (2016) [26] revealed that a widely used criminal risk assessment tool was more likely to falsely flag Black defendants as future criminals than their white counterparts This issue was then highlighted by Buolamwini and Gebru (2018)[27]; in their Gender Shades study, which demonstrated significant accuracy disparities in commercial facial recognition systems for darker-skinned women. Adopting the principle of non-discrimination necessitates a proactive approach to identifying and mitigating bias throughout the entire AI lifecycle, from data collection to model evaluation.

Transparency is the principle that AI systems and their decision-making processes should be understandable and open to scrutiny at several levels, including by competent authorities, users, and persons affected by the system. This is particularly challenging for complex "black-box" machine learning models[28], like deep neural networks, which can operate in ways that are opaque to human beings. The ability to comprehend how an AI reached a particular conclusion is essential for building trust and enabling meaningful control by authorities and users. As discussed by Larsson and Heintz (2020), transparency is a multifaceted concept that is central to the governance of AI. De Laat (2018) argues that transparency is a prerequisite for restoring accountability, as without it, it is virtually impossible to understand the causal chain of an AI's decisions.

Privacy and Personal Data Protection are fundamental rights (Art. 7 and Art. 8 of the European Charter of Fundamental Rights) that must be safeguarded in the age of AI. Many AI applications, particularly those involving machine learning, rely on vast amounts of data, much of which is personal in nature. The ethical principle of privacy requires that personal data be protected from unauthorised access, use, and disclosure. This includes adhering to data protection regulations like GDPR, and it goes further to encompass the ethical obligation to collect only the necessary data, ensure it is secure, and provide users with control over their information, even beyond what is merely requested by data protection laws.

Accountability ensures that a natural or legal person (i.e., not the AI system) can be held responsible for the consequences of an AI system's actions or decisions. This is a critical principle for both ethical and legal reasons. When an AI system causes harm, it is essential to be able to identify who is responsible, be it the developer, the deployer, the user, or other entities provided by the AI Act. In the literature, Lepri et al. (2018)[29] and de Laat (2018)[30], as well as many others, have emphasised the necessity for transparency to achieve accountability. Without a clear chain of responsibility, it is difficult to implement oversight, provide recourse for those who are harmed, and learn from mistakes. Therefore, accountability also encompasses "neighbour" areas of governance, such as the definition of effective redress mechanisms for individuals negatively impacted by the AI system.

3.2. GENDER Audit Tool

Within the CERTAIN project, the Gender Audit Tool[6] serves **as a structured mechanism to evaluate and enhance gender inclusivity across all stages of AI and technology development**. The tool is situated within three high-level phases; the preparation phase provides leadership buy-in and articulates the rationale for the audit, as well as the aims that match the project's aims for gender equality; the implementation phase uses self-assessment checklists that have been adapted to the project's context with self-assessment, interviews, and document reviews to provide an assessment of the practices of partners; and finally, follow-up entails the development of the findings into an action plan, with the incorporation of gender performance indicators into ongoing monitoring.

[6] A gender audit is a tool that institutions use to examine the extent to which gender equality is mainstreamed within their structures, policies, and procedures. While a financial audit examines an organization's monetary aspects, a gender audit focuses on organizational and societal dimensions, assessing whether an institution's culture, operations, and services promote or hinder equality between women and men. Conducting a gender audit allows organizations to unveil gaps, set standards, and draw concrete measures towards the improvement of gender mainstreaming. See https://eige.europa.eu/gender-mainstreaming/tools-methods/gender-audit#toc-what-is-gender-audit

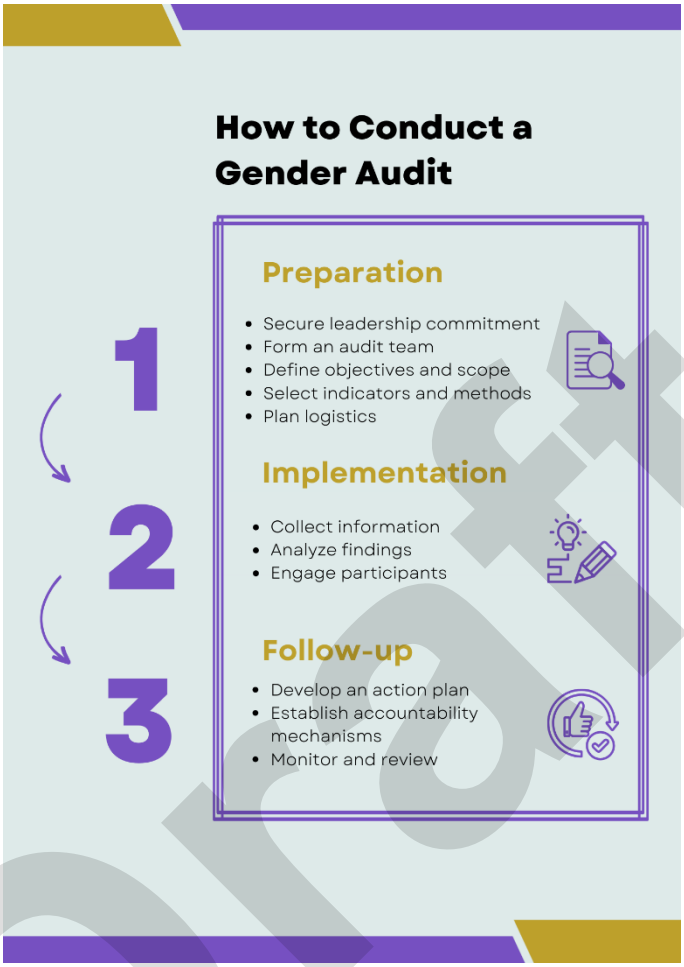

Figure 1 Gender Audit Tool

Complementing this, self-assessment tools specifically designed for our partners embed key gender audit questions, enabling a structured and comprehensive evaluation of gender-related practices. The self-assessment checklist guides the partners in explicitly addressing a number of key concerns, such as the extent to which gender audits have occurred with datasets, algorithms, and procedures; the extent to which gender stereotypes are identified and mitigated in outputs, interfaces, or communications; the extent to which underrepresented genders are included in meaningful ways in design and testing; whether performance metrics are disaggregated by gender and other intersecting identities; to what extent gender-sensitive methodologies are integrated into research and design; and the extent to which institutional mechanisms exist that ensure there is accountability and responsibility for gender inclusion and gender equity.

Through cooperative application of the tool across partners, the project embeds gender inclusion as a core part of the ethical and responsible development of AI, rather than as a stand-alone compliance issue. It results in evidence-based insights into gender representation and bias within tech systems, supports targeted improvement actions through feedback and monitoring, and enables organisations to learn to partner in utilising gender sensitive approaches in their wider operating context. The Gender Audit Tool is designed to operationalise gender responsiveness, as it applies at the technical, organisational, and cultural levels, and takes equality goals from the conceptual and abstract into measurable and tangible outcomes.

3.3. Ethical self-assessment checklist

The ethical self-assessment checklist is intended to be a practical, structured tool to help partners systematically evaluate, document, and improve ethical and gender-responsive practices throughout the lifecycle of the AI system. It brings to life the project's ethical principles and regulatory alignment goals into actionable self-assessment questions, allowing for both accountability and continuous improvement. The checklist is divided into three implementation stages: **Design & Development**, **Deployment & Operationalization**, and **Monitoring & Continuous Post-Market Improvement**, each of which contains targeted thematic areas and guiding questions.

In Phase I: Design & Development, partners examine foundational ethical and social dimensions incorporated during the creation of the AI system. This phase includes four thematic areas:

1. **Benefit–Harm Anticipation**, which evaluates how potential benefits and harms are identified, compared, and documented, including environmental and societal impacts;
2. **Fairness and Non-Discrimination**, focusing on methods to detect and mitigate bias, ensure accessibility, and promote inclusion;
3. **Human Oversight**, assessing how ethical principles, governance structures, and accountability mechanisms are embedded into design and decision-making; and
4. **Gender Inclusivity**, ensuring that gender perspectives are integrated into data, design, team composition, and project culture.

In **Phase II: Deployment & Operationalization**, the checklist supports ethical governance during system rollout. It includes:

- **Transparency**, evaluating documentation, communication, and explainability practices for different audiences;
- **Accountability**, ensuring that responsibility, reporting, and grievance mechanisms are in place and functioning; and
- **Legal Compliance**, assessing adherence to data protection, AI, and digital regulations, including GDPR, the AI Act, and the Data Act.

In **Phase III: Monitoring & Continuous Post-Market Improvement**, the checklist focuses on maintaining ethical and gender standards during real-world use. The last two sections - Deployment, Detection of Harm & Monitoring and Sustainability & Continuous Improvement - are meant to support partners to set up feedback loops, review and analyse risks based on their own environments, assess with and in collaboration monitoring tools, and take it upon themselves to be active in the continuous review and attention to ethical, social and gender consequences in the long-term.

To complement these qualitative insights, the checklist incorporates a **risk matrix**, providing a structured method to assess ethical and operational risks across the system's lifecycle. By combining two central dimensions, likelihood and severity, the risk matrix illustrates the overall risk, which is associated with estimating responses to each issue, within the ethical self-assessment. This matrix helps make decisions about where to prioritise responses and ensures that responses to potential ethical or compliance risks are proportionate to the potential severity of those risks.

Each question was first assigned **a score of 0 to 3 i**n the checklist using **a four-point scale**, indicating the extent to which there was sufficient preparedness and documentation in place.

3 – Fully embedded, reviewed, and continuously improved: The element is an integral part of organisational culture and operations. It is regularly reviewed, evaluated, and refined to ensure ongoing improvement and alignment with best practices.

2 – Documented and implemented: The element is clearly defined, documented, and put into practice. Responsibilities and procedures are established and generally followed.

1 – Partially addressed / informal: The element is acknowledged or applied inconsistently, often through ad hoc or informal practices, but lacks formal documentation or systematic implementation.

0 – Not evident/absent: There is no indication that the element is considered, documented, or implemented in practice.

The table below shows how the descriptive answers were scored on a 0–3 scale.

Table 1 Scoring System

Score	Meaning	What You Look For in the Description	Evidence or Indicators Needed
3 = Fully embedded, reviewed, and continuously improved	The element is fully integrated into operations and culture, regularly reviewed, and continuously enhanced.	Description shows a mature, consistent, and proactive practice — includes regular reviews, updates, and demonstrable improvements.	Comprehensive evidence such as policy updates, review reports, audit logs, meeting minutes, performance indicators, or continuous improvement records.
2 = Documented and implemented	The element is clearly defined, documented, and implemented in practice.	Description details specific processes, responsibilities, and documentation, showing that procedures are consistently followed.	Evidence such as official policies, procedural documents, implementation records, reports, or assigned responsibilities.
1 = Partially addressed / informal	The element is acknowledged and may be applied inconsistently or informally.	Description reflects awareness or ad hoc activities without formal documentation or full implementation.	Limited or inconsistent evidence — drafts, informal notes, partial data, or examples of unsystematic practice.
0 = Not evident / absent	The element is not evident in documentation or practice.	No mention of concrete actions, responsibilities, or established processes.	No supporting evidence; absence of documentation, data, or designated ownership.

Based on the scores to each question, the assessor can also attempt to estimate the **likelihood** of the risk or issue occurring or manifesting, and the **severity** of the consequences if it does. The intersection of these two axes produces a risk level using a colour-coded matrix, as shown below.

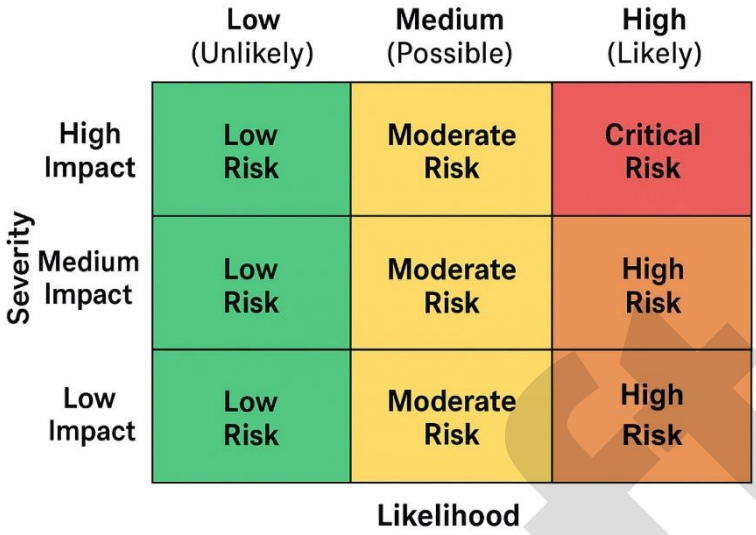

Figure 2 Colour-coded Matrix

To support interpretation, partners can use the following illustrative scenarios demonstrating how the scoring of likelihood and severity translates into practical risk levels:

- **Low Risk (e.g., Green):**

  A partner scores **2 or 3** on data-protection documentation and implementation, indicating clear policies and regular reviews. A minor documentation inconsistency is identified during the checklist (e.g., outdated wording in a non-critical internal form). The likelihood of harm is **low**, and the severity is **minimal**, resulting in a low risk level that requires only routine correction.

- **Moderate Risk (e.g., Yellow):**

  An element related to human oversight receives a **1**, showing partial but inconsistent implementation. For example, model-update decisions are sometimes reviewed by a designated person but lack a formally documented approval process. The likelihood of oversight gaps is **medium**, and the potential consequences, such as delayed identification of model drift, are **moderate**, producing a risk level requiring targeted improvements.

- **High Risk (e.g., Orange/Red):**

  A partner scores **0 or 1** on fairness and bias mitigation practices because no systematic bias review has been conducted for a model deployed in a sensitive domain (e.g., health triage or social-service eligibility). The likelihood of discriminatory outcomes is **high**, and the severity (impact on individuals' health, rights, or access to services) is **significant**, resulting in a high-risk rating that demands urgent mitigation.

These examples help clarify how descriptive scores (0–3) inform the two key risk dimensions and allow partners to more confidently calibrate their assessments.

The usability of this matrix lies in its simplicity and visual clarity. It allows decision-makers to:

- **Quantify ethical performance** and identify weak areas across multiple dimensions (benefit–harm, bias, inclusivity, etc.).
- **Prioritise actions** by distinguishing between acceptable risks and those requiring urgent mitigation.
- **Track progress over time**, since repeating the checklist periodically provides a measurable indicator of ethical maturity and compliance improvement.
- **Communicate results effectively** both internally (to development and governance teams) and externally (to auditors, regulators, or stakeholders).

Overall, the risk matrix transforms qualitative ethical reflections into a **quantitative, evidence-based assessment tool**, promoting accountability, transparency, and continuous improvement in responsible AI governance.

4. Risk Management and Monitoring

4.1. Risk Identification and categorisation

To build ethical and gender-aware AI systems, an essential first step is the structured identification and categorisation of potential risks. This provides a better understanding of where and how the system may cause harm, either directly or indirectly, at every juncture in its lifecycle. This spans consideration from the design and development phase through to when it is deployed, and by the time it is ultimately in use.

The risk identification approach adopted in this section extends the legal risk categories established in D3.1, particularly those concerning high-risk AI systems, transparency obligations, and rights-based constraints. While D3.1 analysed legal exposure and statutory duties, D3.2 translates these into ethical, social, and gender-sensitive risk categories suitable for ongoing monitoring and pre-deployment evaluation.

In identifying risks, the means of risk assessment can include a variety of techniques, including informed opinions of expertise, consultation with stakeholder perspectives, and systematic reviews of sources of data, processes of design and development, and context of use. Subsequently, the identified risks will be categorised according to their characteristics, source, and impact.

Key risk categories include:

- **Dataset bias:** training data might represent different social or gender groups, the training data can reinforce biased outcomes or stereotypes, as well as reinforce systemic inequality.
- **User exclusion:** The design of the system and the user interface could omit, or disadvantage, certain users, for example as a result of a language barrier, accessibility gaps, or gendered assumptions regarding functionality or design.
- **Discriminatory patterns:** Algorithms can create or exacerbate discriminatory patterns during automated decision-making, resulting in unequal treatment, particularly along gender lines, and affecting opportunities, resources, and outcomes.
- **Lack of Redress Mechanisms:** A lack of obvious accountability structures or grievance pathways may limit opportunities for users to challenge unfair or damaging outcomes, specifically for those groups already vulnerable to marginalisation.

Each identified risk is allocated a risk typology – ethical-based, gender-based, or both and is defined in terms of its potential social, legal, technical, reputational, and/or gender specific implications. These risk definitions form the basis for systematic assessment in the Ethical and Gender Risk Register outline in Section 4.2.

4.2. Ethical and Gender Risk Register

The Ethical and Gender Risk Register is a structured tool to specify, assess, and monitor risks that could occur ethically and/or with respect to gender during the lifecycle of an AI system. The purpose of the register is to address risks proactively for compliance with the EU AI Act, the Ethics Guidelines for Trustworthy AI, and the Gender Impact Assessment from the *European Institute for Gender Equality* (EIGE).

For each risk registered the register records a unique identification, the type of risk (ethical, gender, or both), the identification of the risk itself along with the description of it (including details such as possible social impact, legal impact, technical impact, reputational impact, or gender-based impact), projected and calculated based on likelihood of occurrence, harm associated with likely outcomes, and overall risk rating based on these overall features. There are tools in the register to specify key mitigating measures, who is responsible for carrying out risk treatment, and the overall status of the risk (e.g., open, in mitigation, closed).

Ethical risks may include, for example, lack of transparency in algorithmic decisions or "black box" models, negligible explainability of the outputs of the system to users, the risk of personal data breaches, no clear accountability mechanisms, or risks of incorrect or unintended use in morally or legally sensitive contexts.

A risk register is a flexible tool that needs to be revised regularly-such as quarterly or whenever there are significant changes to the design, deployment, or use context of the system. An interdisciplinary team will take part in the recordkeeping and updating process, including legal team members, technical experts, social scientists, and specialists in Gender. The purpose of the data from this process is to ensure that risks are addressed from multiple perspectives. Records of lessons learned and actions taken will also need to be retained to facilitate continuous improvement.

Ultimately, the Ethical and Gender Risk Register has a further requirement as input into any conformity processes, procedures, and assessments under the AI Act, to evidence the risks identified, rated, and mitigated in a systematic way, pursuant with EU regulatory obligations under the EU AI Act. The record will also act as a whole, historical document for audits and conformity certification bodies, to provide transparency and accountability with ethical and gender-aware AI systems.

The Ethical and Gender Risk Register serves as a foundational component for the technical documentation required under Annex IV of the AI Act. By formalizing the recording of risk ratings, mitigating measures, and assigned responsibilities, the register ensures that the identification of ethical and gender-related risks is traceable and auditable. This documentation provides the necessary evidence of a structured risk management system as mandated for conformity assessments.

Risks that may reflect gender include using training data that is biased and replicates gender based stereotypes, gender disparities in access to the relevant system, considerable disparities in terms of levels of trust or acceptance of the system between gender groups, a lack of women and non-binary people involved in its system development and design, or indirect discrimination in automated decision-making options leading to limiting job offers and social opportunities.

4.3. Monitoring

Monitoring refers to the ongoing and systematic process of ensuring that managed ethical risks and gender-related risks do not compromise the AI system throughout its lifecycle. Risk identification and recording begin the process, but monitoring keeps risks visible, measurable, and manageable throughout the system's life as it is actually implemented and transformed in real-world contexts.

Monitoring involves the regular collection, analysis, and review of operations or interactions; user feedback; and performance indicators with a focus on signs of AI bias, discrimination, exclusion, or ethical violations. It also serves the purpose of confirming that any mitigation strategies implemented are purposeful.

In practice, monitoring comprises a number of related activities. These activities include performing regular reviews of risks to reconsider all previously identified risks on a regular (e.g., quarterly) basis to determine whether the likelihood or severity has changed or to assess the efficacy of mitigation efforts. Monitoring also entails conducting ongoing impact assessments to review the outputs of the system and any decision-making patterns to identify any unintended ethical or gender impacts (e.g., emergent biases or disparities in users). Monitoring involves the implementation of clear and transparent means for feedback and reporting of incidents that users, stakeholders, and staff can easily access to report concerns, errors, and harms that relate to the operation of the system. Monitoring will also include a record of all monitoring activities, findings, and any corrective action taken. This record will contribute to ensuring and demonstrating accountability for compliance with the EU AI Act during the process of conformity assessment. Moreover, documentation will help inform decisions demonstrated as necessary when revisiting the register of ethics and gender risks (as in Section 4.2) when new risks arise or when existing ones change in nature or severity.

Continuous monitoring processes, including periodic risk reviews and iterative impact assessments, are designed to generate the evidence base required for external audits. All monitoring records and subsequent corrective actions are systematically documented to fulfil the transparency and accountability obligations of the AI Act. This ensures that the system's operational integrity is maintained and can be verified through the technical documentation throughout its lifecycle.

Our approach is about developing comprehensive guidelines, monitoring frameworks, and actionable checklists that will enable partners to effectively manage their own oversight processes. Each partner will therefore have identifiable responsibilities regarding ethical and gender monitoring, including working through self-assessment checklists, collecting data as needed, and reporting on any identified risks. As a result, the decentralized nature of oversight will allow us to systematically embed and sustain the principles of ethical and gender-sensitive AI across all partners.

While having a qualified interdisciplinary oversight group, consisting of experts from technical, legal, social science, and gender backgrounds, would provide valuable diversity and advance ethical governance for the long haul, such a structure will not be established as part of this project. This approach allows for expanded flexibility and scalability in ensuring that monitoring and ethical considerations are applied directly within each partner's context or the operational context and not concentrated in a single body.

4.4. Compliance

Compliance ensures that the design, development, implementation, and ongoing operation of the AI system comply with all relevant legal, ethical, and gender-specific obligations. Within the CERTAIN project, compliance serves as a more formal bridge between the internal risk management framework and the external regulatory environment, showing that the AI systems do not just intend to adhere to ethical and gender-aware standards, but actually do so across the entire system lifecycle.

For AI systems, compliance means verifying that every aspect of the process, including the selection of datasets, training and developing the model, design and elements of the user interface, and the output of each decision, is verified as compliant with legal obligations, established regulations, and ethical AI guidelines. This means, in particular, that the European Union AI Act, the Ethics Guidelines for Trustworthy AI, and the Gender Impact Assessment frameworks developed by the European Institute for Gender Equality must be adhered to. Compliance must also verify that internal governance rules, as well as industry best practices, for transparency, data protection, fairness,

accessibility, and inclusivity, are all followed, especially in regard to how the system treats its users of different genders or demographic backgrounds.

Compliance-related activities are carefully constructed around the Ethical and Gender Risk Register (Section 4.2), which informs the risk ratings, mitigation measures, and monitoring data elements in Section 4.3 that serve as the evidence base that informs compliance assessments. This requires comprehensive, fully considered, and traceable documentation relating to the design rationale of the project, ethical risk analysis, gender bias audit and mitigations, and monitoring of ongoing performance as it relates to these components. Collectively, this documentation should be well-structured and easily accessible as it relates to external audit, certification, and regulatory inspection processes.

Additionally, the Ethical Self-Assessment Tool[7], developed within the CERTAIN project, is available online for access and participation, supporting the compliance process by identifying high-risk systems and areas requiring particular attention. This tool acts as an early warning mechanism that complements formal compliance checks by flagging potential ethical or gender-related risks before they escalate.

Responsibility for oversight of compliance activities will be assigned to a lead or team responsible for compliance, in close consultation with legal expertise, technical leads, and gender expertise. This interdisciplinary approach ensures compliance is regarded as a continual operational responsibility rather than a box to be checked or a milestone to be achieved. Continuous training and awareness activities for all staff involved in the lifecycle of the AI system further support compliance and promote the culture of ethical and gender-aware practice.

In conclusion, solid compliance protects the project from legal, reputational, and ethical risks, and guarantees that the AI system is being consistently developed and operated with transparency, accountability, and respect for fundamental rights and principles of gender equality.

[7] Available at https://icsa-hua.github.io/checklist/

5. SEX AND GENDER IMPACT ASSESSMENT

5.1. SGIA Methodological Foundations - Contextual analysis

It is important to understand **how individuals view technology to know the potential uses and potential misuses of specific technologies**. However, much of the existing literature has overlooked the complicated relationship that exists between gender and technology exposure and, by extension, technology-knowledge. For example,[31] reported that only 28% of automated vehicle owners are women. The marketing of products towards female consumers often indicates safety features, rather than performance features (e.g. speed) - which in itself reinforces stereotypes of gender roles[32].

Research trying to incorporate a gender lens into perceived and actual technology acceptance models often misses how processes of socialisation differ across gender, thus maintaining binaries and reinforcing broader patterns of the division of labour. Women comprise only 28% of the Science, Technology, Engineering, and Mathematics (STEM) field[33] and the digital gender divide still exists for women and girls[34].

These statistics show that access and engagement with technological innovation are certainly not equitable. Understanding who can access technology and how they respond and leverage technology is essential to understanding their interaction with emerging technology.

5.2. CERTAIN's Methodology and Analytical Framework

Building on the above insights, the CERTAIN project employs a **blended analytical framework** that combines the ***Technology Acceptance Model* (TAM)** and the ***Unified Theory of Acceptance and Use of Technology* (UTAUT)**, enhanced by a gender-sensitive dimension. Key elements from the *Technology Acceptance Model* (TAM), including *Perceived Usefulness* (PU) and *Perceived Ease of Use* (PEOU), are combined with elements from the *Unified Theory of Acceptance and Use of Technology* (UTAUT), such as *Social Influence* (SI) and *Facilitating Conditions* (FC). In addition, our model extends these theories by examining *Gender Influence* (GI) as a moderating factor that influences the process of technology adoption. This integrated perspective allows for a more holistic understanding of how individuals engage with new technologies, particularly concerning gender dynamics. The following image provides a schematic representation of how Qetam operates.

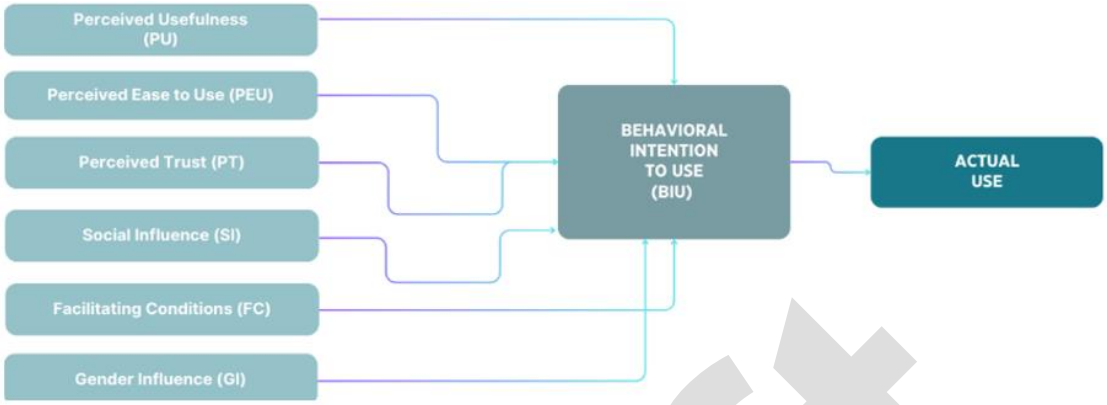

Figure 3 QETAM Conceptual Modelling

The research is based on the creation and distribution of a specially designed questionnaire, containing questions related to both demographic information and knowledge and perceptions associated with the six parameters mentioned above. Responses are provided on a scale from 1 to 10. After collecting the responses, a **correlation matrix** is generated to produce the final BIU number regarding the acceptance of the respective technology. The ***Batch Index Unit*** **(BIU)** number will be calculated using

$$\forall i \in \{1,2,\dots,N\} : BIU_i = PU_i \cdot PEU_i \cdot PT_i \cdot SI_i \cdot FC_i \cdot GI_i \ ,$$

The equation $BIU_i = PU_i \times PEU_i \times PT_i \times SI_i \times FC_i \times GI_i$ means that a user's behavioural intention to use a technology increase when it is perceived as useful, easy to use, trustworthy, socially supported, and well facilitated, while also being influenced by gender-related factors such as differences in technology perceptions, confidence, and usage patterns. A low score in any one factor reduces the overall behavioural intention.

At the same time, using the same model, we measure (potential) gender differences among the participants concerning PU, PEOU, PT, SI, and FC, and proceed with a corresponding analysis of the research findings.

In practice, this model is applied through structured surveys that gather participants' perceptions of AI technologies across the six core dimensions (PU, PEOU, PT, SI, FC, and GI). By analysing these responses, partners can identify which factors most strongly influence acceptance—for example, whether perceived usefulness or trust plays a greater role in shaping willingness to use a specific AI system. The same survey data also enables the exploration of gendered patterns by comparing how different gender groups respond to dimensions such as ease of use, social influence, or facilitating conditions. These insights provide a practical basis for designing targeted interventions, such as improving user training, enhancing transparency, or adjusting communication strategies, with the dual aim of increasing overall acceptance and narrowing potential gender gaps.

The proposed model, previously introduced, is designed to operate in full alignment with the *Gender Impact Assessment* (GIA) framework established by the European Institute for Gender Equality (2016). In this context, the QETAM framework was deliberately selected because it explicitly positions *gender influence* (GI) as a central mediating factor, rather than a peripheral variable. This conceptualisation allows the model to systematically capture how gender-related dynamics shape

relationships among key variables, ensuring methodological coherence with GIA principles and strengthening the explanatory power of the proposed approach.

The GIA serves as a structured methodology for assessing how proposed policies, programmes, or initiatives may influence gender equality, either positively or negatively, and for ensuring that equality considerations are integrated into decision-making processes from the outset. In line with EIGE's methodology, the successful implementation of a GIA consists of the following five interdependent steps:

1. **Definition of the programme's purpose and its connection to gender equality:**

   This step involves clarifying the objectives of the proposed initiative, articulating its intended societal impact, and explicitly mapping how these aims intersect with gender-equality priorities at local, national, and EU levels. The rationale for this link must be transparent, evidence-based, and situated within broader equality frameworks and legal obligations.

2. **Assessment of gender relevance:**

   Here, the programme's potential to either affect or be affected by gender-related factors is assessed. This process includes identifying both direct and indirect effects, considering both the target user group and the wider community and potential spillover effects on gender dynamics.

3. **Conducting a gender-sensitive analysis:**

   This stage requires the collection and examination of sex-disaggregated and intersectional data, assessing existing inequalities, barriers, and enabling factors. It emphasizes understanding how different groups—women, men, and non-binary individuals—experience technology and resources differently due to structural, cultural, and socio-economic factors.

4. **Evaluating the gender impact:**

   Once the analysis is complete, the likely consequences of the programme on gender relations are evaluated. This includes both anticipated benefits and potential risks, using measurable indicators to assess inclusivity and equity outcomes.

5. **Formulating findings and proposals for improvement:**

   The final step transforms analytical insights into actionable recommendations to refine the programme to maximise gender equality benefits and minimise unintended discriminatory effects.

As elaborated in the *Ethical Governance Frameworks* section of this deliverable, Artificial Intelligence (AI) technologies are inherently intertwined with gender-equality considerations. One of the principal reasons is that AI models are trained on large-scale datasets, which often reflect—and sometimes magnify—pre-existing societal patterns of inequality. These patterns can take the form of entrenched gender and cultural biases, which, when embedded into AI algorithms, risk perpetuating discriminatory outcomes at scale (Step 1).

Moreover, user acceptance and the adoption of AI technologies are not gender-neutral processes. The interaction between individuals and emerging technologies can influence gender identities,

shaping and reinforcing socially constructed notions of femininity and masculinity in different ways. Disparities in digital literacy, access to technology, and confidence in usage are well-documented across genders. Evidence from EIGE (2020) and UNICEF (2023) shows that women are statistically more likely to lack foundational digital skills, whereas men are often more experienced, confident, and willing to engage with advanced technological systems, including AI.

These inequalities are further compounded by unequal exposure opportunities. Factors such as occupational segregation, differing educational pathways, socio-economic constraints, and cultural expectations contribute to the unequal distribution of familiarity with AI. If these issues are not actively addressed during the design, testing, and deployment stages of AI solutions, they can exacerbate the digital skills gap and reinforce systemic barriers to participation (Step 2).

Within the CERTAIN project, all research activities will embed gender-sensitive analysis as a fundamental requirement (Step 3). This entails continuously monitoring differences in technology acceptance rates, motivations for adoption, and perceived barriers among women, men, and non-binary individuals. The aim is to ensure that these differences are recognised not as incidental but as central factors shaping technology uptake and trust.

In parallel, CERTAIN partners will undertake a systematic evaluation of projected gender impacts during the design and development of AI systems (Step 4). This evaluation will rely on specific, measurable indicators, such as:

a) the equitable participation of women, men, and non-binary individuals in both user and stakeholder groups, and
b) access to and control over resources that enable effective use of AI technologies.

For example, transport-related studies show that in households with a single car, men are statistically more likely to be the primary users, while women often depend on alternative modes of transport. This dynamic also has implications for the adoption of automated vehicle technologies. As already mentioned, current data indicate that approximately 28% of automated vehicle owners are women, suggesting that women, on average, may be less exposed to these technologies and therefore less likely to demonstrate high levels of acceptance.

Finally, after each technology acceptance study, the CERTAIN project will develop tailored, evidence-based recommendations to enhance inclusivity and address identified gaps (Step 5). These recommendations will be context-specific and will be designed to ensure that the introduction and scaling of AI technologies actively contribute to narrowing, rather than widening, gender-based disparities in digital engagement and participation.

6. CERTAIN'S ETHICAL GUIDELINES

CERTAIN is committed to **building AI systems that are trustworthy, lawful, and human-centric**. These guidelines provide a framework to ensure ethical, fair, and gender-sensitive AI throughout the project lifecycle. They are grounded in EU policies, including the Ethics Guidelines for Trustworthy AI, the EU AI Act, and emerging CEN-CENELEC standards.

Human-centred design serves as a guiding principle that ensures that AI systems are designed with consideration for human rights, dignity, and autonomy. Novelty in technology is subservient to user well-being and societal benefit, and methods are put in place to enable meaningful human oversight of AI output. The AI life cycle integrates fairness and non-discrimination, undertaking proactive measures to find and mitigate bias in data, algorithms, and outcomes. Equal access for all is a priority, and the potential to discriminate based on gender, age, ethnicity, or other social categories is avoided, incorporating fairness measures into the design process, different testing environments, deployment, and operational processes.

AI practices that are gender-sensitive are highlighted through gender impact assessments, which take on defined measures and scenario-based assessments. Gender balance is achieved with regard to data sets, team composition, and evaluation. Gender audit processes are established as part of project governance. Transparency and explainability are paramount when providing information about AI systems, their purpose, capabilities, and limitations. Accordingly, explanatory processes for decision-making are made explainable to users and auditors, inclusive of establishing and maintaining records as part of the expectations for an AI model deployment, inclusive of the data sets, models, and design choices.

Privacy and data protection are diligently respected in accordance with GDPR and other relevant regulators, as data collection is kept to a minimum, adhering to the objectives of the project. Personal or sensitive data is securely stored, processed and shared. The project reinforces accountability and monitoring with specific roles and responsibilities clearly outlined to monitor ethical oversight and monitoring and reporting frameworks to identify and monitor conformance to ethical guidelines. Self-assessment tools and checklists are provided to facilitate ongoing compliance assessments.

Compliance with legal and regulatory standards is one integral part to ensure AI development, deployment, and usage is aligned with the EU AI Act and relevant sectoral regulations in conjunction with following the upcoming CEN-CENELEC standards for auditing and trustworthy AI systems. Continuous updates will be made as policies, standards, and societal expectations change. Finally, sustainability and social impact are assessed through methodologies to consider environmental and societal impacts (e.g., adopting sustainable practices in the development and deployment of AI) and long-term impact assessments to prevent harm to society and unintended social disruption.

All of these guidelines will be **systematically integrated into the project and examined through the creation of self-assessment tools**, ensuring that ethical, fair, and gender-sensitive practices are continuously monitored and evaluated throughout the AI system's lifecycle. To achieve this, it is essential to collect the answers of the CERTAIN partners' answers on the checklist, as their input will provide critical insights into how ethical principles are understood and applied across different

contexts. This collaborative process is crucial for ensuring that the integration of ethical principles is comprehensive, contextually relevant, and effectively embedded in all stages of the project.

7. IMPLEMENTATION AND INTEGRATION PLAN

The ethical self-assessment checklist constitutes a **practical mechanism for embedding principles into practice**, ensuring that ethical and gender considerations are not relegated to abstract guidelines but are operationalised through structured, auditable processes. Its role is to move beyond compliance formalities, functioning instead as a **reflexive instrument** that guides partners in navigating normative commitments, regulatory obligations, and context-specific risks. By organising requirements into concrete questions on consent, transparency, gender sensitivity, inclusiveness, accountability, and data protection, the checklist supports both foresight and continuous reassessment, thus consolidating a culture of responsibility across the project.

To support early validation and real-world applicability, the checklist has already been disseminated to all project partners, including colleagues involved in the seven operational pilots across six different business areas that are currently underway. Together with the checklist, an accompanying evaluation form was distributed to enable structured feedback, assess usability and relevance, and collect empirical insights from practical implementation. This dual deployment is intended to test the effectiveness of the self-assessment tool in operational contexts and to extract evidence-based results that will inform its refinement and long-term integration.

7.1. Integration within the governance architecture

The checklist operates in close interrelation with existing governance mechanisms:

- While the Specific Gender and Inclusion Assessment identifies structural risks and opportunities, the checklist translates them into operational queries and verifiable practices, thereby ensuring that gender and inclusiveness concerns are systematically addressed in day-to-day activities.
- Outputs from the self-assessment feed directly into the project's risk register, ensuring that ethical and gender-related risks are tied to documented mitigation strategies and accountability pathways.
- The design of the tool draws on the EU AI Act (Articles 9–15 on risk management, data governance, transparency, human oversight, and documentation), and HRIA. This grounding secures both regulatory preparedness and methodological consistency with broader human-rights-based approaches.

The integration of ethical oversight into the technical architecture directly supports the requirements for Technical Documentation (Annex IV). By mapping outputs from the self-assessment checklists and the risk register into the project's broader governance ecosystem, CERTAIN ensures that design choices, particularly those involving fairness, transparency, and human oversight, are documented as part of the system's explanatory processes. This creates a coordinated pathway where internal ethical assurance serves as the primary source for regulatory compliance and certification readiness

The governance architecture implemented here aligns also with D3.1's legal mapping to ensure that ethical oversight and gender monitoring operate in continuity with the compliance obligations defined at the regulatory level. In practice, this means that the procedures, checklists, and monitoring tools introduced in D3.2 serve as the ethical-governance complement to the legal-compliance framework articulated in D3.1.

7.2. Pilot implementation and Feedback loop

The ethical self-assessment checklist is being applied across **seven pilots**, covering diverse socio-technical domains and risk profiles. Each pilot provides a concrete context for testing the checklist's

relevance, usability, and capacity to surface ethical, legal, and gender-related considerations in real-world deployments:

- **Pilot 1 – Biometrics**
  Focuses on the use of biometric technologies, examining issues of consent, data protection, and fairness in identity-related AI applications.
- **Pilot 2 – Health**
  Explores AI-supported health applications, with particular attention to sensitive data handling, bias, transparency, and trust in clinical and patient-facing contexts.
- **Pilot 3 – Energy**
  Investigates AI-driven optimisation within renewable energy communities, addressing data privacy, fairness in participation, and trust in automated energy management systems.
- **Pilot 4 – Human Resources**
  Applies AI tools to labour market analysis and recruitment, testing secure data sharing, regulatory compliance, transparency, and accessibility in HR decision-making.
- **Pilot 5 – Data Holders**
  Targets SMEs acting as data holders, assessing how the CERTAIN framework supports ethical, legal, and efficient participation in AI data markets.
- **Pilot 6 – Finance**
  Demonstrates explainable and fair AI for personalised investment recommendations, focusing on transparency, bias mitigation, and user trust.
- **Pilot 7 – Automated MLOps for Regulatory Compliance**
  Embeds automated legal and ethical checks into MLOps pipelines to ensure continuous compliance with GDPR and the EU AI Act throughout the AI lifecycle.

Across all pilots, partners apply the checklist alongside the evaluation form to generate structured feedback. The results collected from these seven implementations feed into an iterative refinement loop, ensuring that the self-assessment tool remains robust, context-sensitive, and aligned with regulatory, ethical, and gender-equality objectives.

8. CONCLUSIONS

This deliverable has outlined a governance framework that translates abstract ethical principles into concrete tools for action, thereby reinforcing CERTAIN's commitment to trustworthy, lawful, and inclusive AI. By integrating ethical oversight with gender-sensitive methodologies, it positions the project in alignment with the European regulatory landscape and at the forefront of anticipating societal risks and embedding values of fairness, transparency, accountability, and non-discrimination into the AI lifecycle.

A distinctive contribution lies in the incorporation of a Sex and Gender Impact Assessment (SGIA) framework, which underscores that ethical robustness cannot be disentangled from gender equality. By addressing inequalities at the level of data, design, and access, the framework helps ensure that CERTAIN's outputs foster inclusivity rather than deepen existing divides. Together with self-assessment tools, gender audits, and an Ethical and Gender Risk Register, these measures create a reflexive system of accountability that makes value commitments demonstrable and auditable.

A practical **ethical self-assessment tool** was designed under the works of this deliverable, as a structured aid for partners to translate principles into practice. Rather than offering abstract guidelines, the self-assessment tool organises key requirements into operational questions on consent, transparency, gender sensitivity, inclusiveness, accountability, and data protection. Its cyclic use supports planning, monitoring, and reassessment throughout the project, ensuring that ethical and gender commitments are revisited at each stage of the research and technical development. In this way, the checklist complements the SGIA and risk register by embedding reflexive and auditable practices into everyday decision-making.

At the same time, the work highlights enduring challenges. The translation of principles into practice is necessarily iterative and context sensitive. While this deliverable provides a first architecture of ethical and gender governance, its success will depend on sustained engagement, cross-partner coordination, and responsiveness to emerging risks and regulatory developments. Ethical foresight must be treated as an ongoing process, not a static checklist, with mechanisms for feedback, monitoring, and continuous improvement.

Together with D3.1, which establishes the regulatory and legal foundations for CERTAIN's approach to trustworthy AI, this deliverable forms the ethical and gender-responsive operational extension of that framework. Where D3.1 has defined what compliance requires, D3.2 defines how these requirements can be implemented ethically, inclusively, and in a manner compatible with future certification pathways.

Looking ahead, the framework developed here should serve as both a guide and a living structure. Its integration into CERTAIN's technical work packages will allow for iterative refinement, while its alignment with evolving European standards and the AI Act will ensure long-term certification readiness. More broadly, by demonstrating that compliance can be a mode of ethical governance rather than a procedural burden, this deliverable contributes to the European ambition of making trustworthy, inclusive, and auditable AI the norm.

APPENDIX A

Ethical Self-Assessment Checklist

Phase	Domain / Section	Questions
Phase I: Design & Development	**1. Intended Use and Proportionality**	1 To what extent are the objectives and intended uses of the AI system clearly defined and documented within your project or institution? 2. To what degree were potential harms to fundamental rights, society, or the environment systematically identified? 3. How effectively are risks of misuse, dual-use, or adversarial applications anticipated and mitigated? 4. To what extent does the proportionality assessment compare expected benefits to potential harms? 5. To what degree are environmental and sustainability impacts integrated into benefit–harm assessments? 6. To what extent did stakeholders (users, domain experts, civil society) contribute to identifying benefits and harms? 7. To what degree are downstream or long-term impacts (economic, psychological, environmental, or cultural) evaluated and addressed after the project concludes? 8. To what extent do institutional mechanisms ensure accountability for identified risks and benefits? 9. To what degree are tensions between ethical principles (e.g., transparency vs. confidentiality) identified, balanced, and justified? 10. To what extent are risk assessments documented, reviewed, and updated over time?
	2. Fairness & Non-Discrimination	1. To what extent are methods used to detect bias in datasets, algorithms, or system outputs rigorous and validated? 2. How effectively are mitigation strategies for identified biases designed, implemented, and evaluated? 3. To what degree is fairness operationalized in your project or institutional framework (criteria, definitions, metrics)? 4. To what extent is fairness tested across demographic groups, including vulnerable populations? 5. To what degree is accessibility for persons with disabilities embedded in system design and evaluation? 6. To what extent are intersectional categories (e.g., race × gender × age) considered in fairness assessments? 7. How effectively are fairness concerns raised by affected communities incorporated into system improvements? 8. To what degree do institutional mechanisms ensure accountability for fairness outcomes and governance of bias issues? 9. To what extent are conflicts between fairness objectives and other ethical principles (e.g., accuracy or privacy) identified and addressed? 10. To what degree is fairness continuously monitored and documented throughout the AI lifecycle?

	3. Human Oversight	1. To what extent are the ethical principles guiding the project clearly defined and justified? 2. To what degree are these principles embedded in the design process and documentation? 3. How robust is the governance framework (e.g., ethics committee, oversight board) ensuring accountability? 4. To what extent are stakeholders consulted to identify ethical concerns during design? 5. How clearly and reliably is human oversight integrated (approval, override, review mechanisms)? 6. To what degree are roles and responsibilities for ethical compliance clearly assigned and tracked within your institution? 7. To what extent are ethical dilemmas documented, discussed, and resolved during design? 8. To what degree are tensions between ethical principles (e.g., autonomy vs. protection) balanced and justified? 9. To what extent are staff trained and supported to identify and manage ethical issues in development? 10. To what degree are ethical guidelines and oversight practices reviewed and updated as risks or norms evolve?
	4. Gender Inclusivity	1. To what extent has a gender audit been conducted on datasets, algorithms, and workflows? 2. How effectively are gender stereotypes identified and mitigated in outputs, interfaces, or communications? 3. To what degree are underrepresented genders meaningfully included in design and testing processes? 4. To what extent are performance metrics disaggregated by gender and intersecting identities? 5. To what degree is gender-sensitive methodology integrated into research and design? 6. To what extent do institutional mechanisms ensure structural accountability for gender inclusion and equity in governance? 7. To what degree is inclusive and gender-sensitive language ensured in documentation and interfaces? 8. To what extent are gender-differentiated impacts identified and addressed, including post-project consequences? 9. To what degree are intersections or conflicts between gender inclusivity and other fairness or privacy goals managed and justified? 10. To what extent is an inclusive and equitable team culture promoted, tracked, and documented?
Phase II: Deployment & Operationalization	**5. Transparency**	1. To what extent is the AI system's functioning (data sources, models, intended use, limitations) documented and communicated within your institution or programme? 2. To what degree are end-users clearly informed that they are interacting with an AI system? 3. How effectively are explanations tailored for different audiences (technical experts, lay users, regulators)? 4. To what extent are explainability techniques validated and reliable? 5. To what degree are uncertainties, limitations, and error rates transparently communicated? 6. To what extent are transparency obligations under applicable laws identified and monitored?

		7. To what degree are decision logs and audit trails maintained to ensure accountability? 8. To what extent are trade-offs between transparency and IP/confidentiality justified and documented? 9. To what degree is transparency ensured when the system is updated or retrained? 10. To what extent are institutional communication plans reviewed and improved to maintain transparency?
	6. Accountability	1. To what extent are accountability roles and responsibilities clearly defined and assigned within your institution? 2. To what degree do internal processes review accountability and ethics before deployment? 3. How effective and accessible are grievance mechanisms allowing individuals to contest or appeal AI decisions? 4. To what extent are accountability responsibilities communicated and audited across organizational levels? 5. To what degree are impact assessments conducted to ensure responsible deployment? 6. To what extent are incidents, failures, or breaches documented and communicated transparently? 7. To what degree are tensions between accountability and confidentiality (e.g., whistleblowing or data disclosure) addressed and justified? 8. To what extent is third-party compliance with accountability requirements verified and enforced? 9. To what degree is staff training on accountability and ethical responsibility maintained and updated? 10. To what extent are accountability frameworks reviewed and improved after project completion?
	7. Legal Compliance	1. To what extent is data minimization achieved in training and deployment? 2. To what degree are anonymization, pseudonymization, or encryption measures effectively implemented for sensitive data? 3. To what extent is user consent obtained, communicated, and managed consistently? 4. To what degree are users' rights to access, rectification, or deletion of data operationalized and respected? 5. To what extent are compliance requirements from the GDPR, AI Act, and Data Act identified, mapped, and integrated into institutional governance strategies 6. To what degree are risks from third-party data sources and integrations assessed and mitigated? 7. To what extent are conflicts between legal compliance requirements and other ethical principles (e.g., transparency or fairness) resolved and justified? 8. To what degree are safeguards against misuse or unauthorized access documented, tested, and verified? 9. To what extent are institutional compliance systems monitored or externally audited for robustness? 10. To what degree are compliance practices adapted and sustained beyond the project lifecycle?
Phase III: Monitoring & Continuous	**8. Deployment, Harm Detection & Monitoring**	1. To what extent are early-warning mechanisms for harm detection established at the institutional or project level? 2. To what degree are early-warning mechanisms tested, validated, and refined in practice?

Post-Market Improvement	3. To what extent was real-world pilot testing conducted and evaluated before full deployment? 4. To what degree are user feedback and harm reports systematically collected, recorded, and acted upon? 5. To what extent are periodic risk reassessments performed and documented? 6. To what degree are correction and mitigation strategies predefined and effectively implemented when harms are detected? 7. To what extent are monitoring results used to improve conformity with ethical and legal standards? 8. To what degree are trade-offs between harm mitigation and operational efficiency evaluated and justified? 9. To what extent are institutional accountability and reporting mechanisms maintained for harm detection and response? 10. To what degree are long-term accountability and ethical exit or phase-out strategies planned after system decommissioning?
9. Sustainability & Continuous Improvement	1. To what extent is a structured feedback loop maintained for continuous evaluation of ethical and social impacts? 2. To what degree are monitoring results systematically reviewed and translated into system improvements? 3. To what extent are marginalized or underrepresented groups meaningfully included in governance and decision-making? 4. To what degree is stakeholder input integrated into institutional governance beyond advisory roles? 5. To what extent are long-term ethical, social, and gender-related consequences evaluated beyond the project's duration? 6. To what degree are trade-offs between sustainability goals and economic or operational priorities evaluated and justified? 7. To what extent are ethics and inclusivity training updated and delivered to staff? 8. To what degree are governance mechanisms adapted in response to evolving regulations and social expectations? 9. To what extent is accountability for continuous ethical oversight clearly defined within your institution? 10. To what degree are sustainability and improvement activities documented for institutional learning and transparency?

The Ethical Self-Assessment Checklist can be accessed at the following link: https://icsa-hua.github.io/checklist/

REFERENCES

[i]Fjeld, Jessica, Nele Achten, Hannah Hilligoss, Adam Nagy, and Madhulika Srikumar. "Principled artificial intelligence: Mapping consensus in ethical and rights-based approaches to principles for AI." Berkman Klein Center Research Publication 2020-1 (2020).
[2] Organisation for Economic Co-operation and Development (OECD), "Health at a Glance 2019: OECD Indicators," OECD Publishing, 2019. https://www.oecd.org/content/dam/oecd/en/publications/reports/2019/11/health-at-a-glance-2019_f58fa178/4dd50c09-en.pdf
[3] United Nations Educational, Scientific and Cultural Organization (UNESCO), "UNESCO Science Report 2021: The Race Against Time for Smarter Development," UNESCO Publishing, 2021.
[4] High-Level Expert Group on Artificial Intelligence (HLEG), "Ethics Guidelines for Trustworthy AI," European Commission, 2019. https://www.aepd.es/sites/default/files/2019-09/ai-definition.pdf
[5] Ala-Pietilä, Pekka, et al. The assessment list for trustworthy artificial intelligence (ALTAI). European Commission, 2020.
[6] Ethically Aligned Design - A Vision for Prioritizing Human Well-being with Autonomous and Intelligent Systems," in Ethically Aligned Design - A Vision for Prioritizing Human Well-being with Autonomous and Intelligent Systems , vol., no., pp.1-294, 31 March 2019.
[7] Zicari, Roberto V., et al. "Z-Inspection®: a process to assess trustworthy AI." IEEE Transactions on Technology and Society 2.2 (2021): 83-97.
[8] Mantelero, Alessandro. "Human rights impact assessment and AI." Beyond data: Human rights, ethical and social impact assessment in AI. The Hague: TMC Asser Press, 2022. 45-91.
[9] Council of Europe. HUDERIA – Risk and Impact Assessment of AI Systems. Council of Europe, https://www.coe.int/en/web/artificial-intelligence/huderia-risk-and-impact-assessment-of-ai-systems. Accessed 18 Dec. 2025.
[10] Gerards, Janneke, et al. "Fundamental rights and algorithms impact assessment (fraia)." (2022).
[11] Malgieri, Gianclaudio, and Frank Pasquale. "From transparency to justification: toward ex ante accountability for AI." Brooklyn Law School, Legal Studies Paper 712 (2022).
[12] Larsson, Stefan. "On the governance of artificial intelligence through ethics guidelines." Asian Journal of Law and Society 7, no. 3 (2020): 437-451.
[13] High-Level Expert Group on Artificial Intelligence (HLEG), 2019
[14] European Commission, "On Artificial Intelligence—A European approach to excellence and trust (White Paper)," COM (2020) 65 final, 2020.
[15] Larsson, “Governance of Artificial Intelligence,” 437–451.
[16] Grigorina, B. O. C. E. "Bias in artificial intelligence." In Smart Cities International Conference (SCIC) Proceedings, vol. 10, pp. 337-344. 2022.
[17] Buolamwini, Joy, and Timnit Gebru. "Gender shades: Intersectional accuracy disparities in commercial gender classification." In Conference on fairness, accountability and transparency, pp. 77-91. PMLR, 2018.
[18] Dubal, Veena. "On algorithmic wage discrimination." Columbia Law Review 123, no. 7 (2023): 1929-1992.
[19] B. Lepri, B. Koenigs, and C. Lepri, "Algorithmic Decision-Making Based on Machine Learning from Big Data: Can Transparency Restore Accountability?," Journal of Technology, vol. 22, pp. 45–67, 2018.
[20] Goodfellow, Ian J., Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. "Generative adversarial nets." Advances in neural information processing systems 27 (2014).
[21] Shoaib, Mohamed R., Zefan Wang, Milad Taleby Ahvanooey, and Jun Zhao. "Deepfakes, misinformation, and disinformation in the era of frontier AI, generative AI, and large AI models." In 2023 international conference on computer and applications (ICCA), pp. 1-7. IEEE, 2023.
[22] Fjeld et al., “Principled Artificial Intelligence.”
[23] ALLEA (All European Academies). The European Code of Conduct for Research Integrity. Revised Edition, 2023. ALLEA. https://allea.org/code-of-conduct/

[24] Nkwo, Makuochi, and Muhammad S. Adamu. "AI “Ethics Shopping” and “Governance Shrinking” in Africa: a critical opinion." In Second edition of the Global Forum on the Ethics of Artificial Intelligence (GFEAI 2024), Brdo Slovenia. 2024.
[25] Floridi, Luciano. "Translating principles into practices of digital ethics: Five risks of being unethical." Philosophy & Technology 32, no. 2 (2019): 185-193.
[26] Angwin, Julia, Jeff Larson, Surya Mattu, and Lauren Kirchner. "Machine bias." In Ethics of data and analytics, pp. 254-264. Auerbach Publications, 2022.
[27] Buolamwini and Gebru, “Gender Shades.”
[28] Pasquale, Frank. "The black box society: The secret algorithms that control money and information." In The black box society. Harvard university press, 2015.
[29] De Laat, Paul B. "Algorithmic decision-making based on machine learning from big data: can transparency restore accountability?." Philosophy & technology 31, no. 4 (2018): 525-541.
[30] De Laat, Paul B. "Algorithmic decision-making based on machine learning from big data: can transparency restore accountability?." Philosophy & technology 31, no. 4 (2018): 525-541.
[31] Kimberly Hurley, “Do Electric Cars Have a Gender Divide?,” CBT News, May 31, 2023, https://www.cbtnews.com/do-electric-cars-have-a-gender-divide/
[32] Zahrah Ahmad, “Why Brands Should Rethink Their Approach to Female Consumers,” Spark Growth, January 30, 2025, https://www.sparkgrowth.com/why-brands-should-rethink-their-approach-to-female-consumers/
[33] Ebru Özdemir, “Why It’s Time to Use Reskilling to Unlock Women’s STEM Potential,” World Economic Forum, January 13, 2025, https://www.weforum.org/stories/2025/01/why-it-s-time-to-use-reskilling-to-unlock-women-s-stem-potential/
[34] European Institute for Gender Equality, "Gender impact assessment: Gender mainstreaming toolkit," 2016. [Online]. Available: https://eige.europa.eu/publications/gender-impact-assessment-gender-mainstreaming-toolkit