HH4AI: A Methodological Framework for AI Human Rights Impact Assessment under the EU AI Act

Paolo Ceravolo[1], Ernesto Damiani[1], Maria Elisa D'Amico[1], Bianca de Teffé Erb[2], Simone Favaro[2], Samira Maghool[1], Simone La Porta[2], Lorenzo Maria Ratto Vaquer[2], Lara Mauri[1], Nannerel Fiano[1], Paolo Gambatesa[1], Niccolò Panigada[1] and Marta A. Tamborini[1]

[1]*Università degli Studi di Milano, Milan, Italy*
[2]*Deloitte Financial Advisory S.r.l. S.B., Milan, Italy*

Abstract

This paper introduces the HH4AI Methodology, a structured approach to assessing the impact of AI systems on human rights, focusing on compliance with the EU AI Act and addressing technical, ethical and regulatory challenges. The paper highlights AI's transformative nature, driven by autonomy, data and goal-oriented design, and how the EU AI Act promotes transparency, accountability and safety. A key challenge is defining and assessing "high-risk" AI systems across industries, complicated by the lack of universally accepted standards and AI's rapid evolution.

To address these challenges, the paper explores the relevance of ISO/IEC and IEEE standards, focusing on risk management, data quality, bias mitigation and governance. It proposes a Fundamental Rights Impact Assessment (FRIA) methodology, a gate-based framework designed to isolate and assess risks through phases including an AI system overview, a human rights checklist, an impact assessment and a final output phase. A filtering mechanism tailors the assessment to the system's characteristics, targeting specific areas like accountability, AI literacy, data governance and transparency.

The structured approach enables systematic filtering, comprehensive risk assessment and mitigation planning, effectively prioritizing critical risks and providing clear remediation strategies. This promotes better alignment with human rights principles and enhances regulatory compliance.

Keywords

Artificial Intelligence, Fundamental Rights, Impact Assessment, EU AI Act, AI Governance, AI Ethics

1. Introduction

Artificial Intelligence (AI) encompasses technologies perorming tasks such as reasoning, learning, decision-making and perception. The EU AI Act, defined in Article 3(1), describes AI systems as technologies operating autonomously to process inputs and generate outputs impacting various environments. This definition emphasizes autonomy, data-driven learning and adaptability.

The Act's broad scope encompasses methodologies like machine learning and symbolic reasoning, reflecting AI's evolving nature. Assessing high-risk systems involves analyzing technological and contextual factors, but compliance remains challenging due to AI's rapid evolution, methodological diversity and the absence of universally accepted standards.

AI assessment complexity arises from the interdependence of models, data and external variables that create unpredictable interactions. Continuous updates can alter system behavior without transparency, while inconsistent frameworks and differing regulatory priorities across jurisdictions hinder alignment. Ensuring fairness, transparency and accountability is particularly challenging for opaque models. Effective global governance requires harmonizing EU regulations with international frameworks to avoid trade barriers and encourage innovation.

Resource constraints, especially affecting SMEs, complicate compliance efforts. A structured methodology is essential for effective risk assessment, compliance and promoting trustworthy, human-rights-aligned AI systems.

HHAI-WS 2025: Workshops at the Fourth International Conference on Hybrid Human-Artificial Intelligence (HHAI), June 9–13, 2025, Pisa, Italy

2. Legal and Regulatory Background

2.1. The Challenges of AI Assessment

The EU AI Act [1] establishes a comprehensive framework for regulating AI systems within the EU, emphasizing their autonomy, data-driven nature and adaptive capabilities (Article 3). It encompasses diverse methodologies such as planning, reasoning, knowledge representation and learning, as noted in Recital 12. Risk management procedures (Article 9) are required for high-risk systems, involving risk identification, assessment and mitigation, while low-risk systems may implement these voluntarily. Data governance and reporting requirements emphasize GDPR compliance (Article 10), cybersecurity (Article 15) and data quality (Articles 10 and 15) [2]. Systems must include quality control mechanisms (Article 17), maintain technical documentation (Article 11), log activities (Article 12) and high-risk systems must be registered in a public database (Article 13) to ensure transparency and accountability. Additional provisions require transparency and human oversight (Articles 13 and 14).

The Act outlines conformity assessment processes, distinguishing between internal conformity assessments (Articles 16 and 43) and independent evaluations for biometric systems (Article 43). Compliance with harmonized standards published by the European Commission presumes alignment with the Act (Article 40) [1].

Compliance requirements include procedural frameworks for risk management, documentation, control and conformity assessment, as well as technical adherence to harmonized standards, codes of conduct and best practices. While harmonized standards, codes of practice [3, 4] and codes of conduct [5] are crucial for structured risk management, data governance and transparency, their expected release in 2-3 years leaves organizations without definitive benchmarks for compliance.

The lack of harmonized standards creates ambiguity in interpreting requirements, requiring organizations to rely on existing frameworks such as ISO/IEC 23894, which aligns with the Act's objectives. Industry guidelines and research institutions also offer practical compliance references. However, organizations must remain adaptable, ensuring that current strategies can align with forthcoming standards and codes of practice. Cross-industry collaboration is essential to share insights and prepare for standardized frameworks.

Implementation challenges persist due to the absence of a universally accepted reference framework, making compliance efforts inconsistent and context-dependent. Assessments vary based on application rather than technology, complicating replication and consistency. Additionally, evolving standards driven by technological advancements create a shifting compliance target and the appropriate detail level for assessments remains unclear, especially when balancing self-assessment with empirical validation.

These challenges highlight the need for a structured, flexible approach to AI risk management that aligns with evolving standards and best practices while fostering transparency, accountability and compliance.

2.2. Human Rights and Ethical Considerations

AI systems impact human rights across national, European and international levels, raising ethical and legal concerns. Achieving a balance between technological innovation and fundamental rights protection requires navigating a multi-level legal framework involving constitutions, judicial rulings, rights charters and other regulatory sources.

The EU AI Act aims to establish a uniform framework prioritizing human-centric AI aligned with fundamental rights as outlined in the Charter of Fundamental Rights of the European Union [1]. It seeks to promote trustworthy AI that safeguards health, safety, democracy, rule of law and environmental protection while fostering innovation.

From a human rights perspective, key concerns include equality, privacy, transparency and environmental protection. Biases introduced during AI training and testing can perpetuate discrimination, reflecting societal inequities. Ensuring fairness requires eliminating biases at the design stage. Privacy concerns arise from AI systems processing personal or biometric data, enabling extensive surveillance that threatens personal safety and state security. Transparency is essential for fairness, bias detection

and privacy protection, requiring users to understand AI processes, data sources and decision-making logic. Additionally, while AI can enhance sustainability efforts, its energy consumption can adversely impact the environment, conflicting with sustainable development principles.

Ethical considerations intersect with human rights through transparency, accountability and continuous monitoring of AI systems to prevent inequalities. Establishing a clear regulatory framework that addresses liability for harm caused by AI systems while promoting human-centered AI governance remains crucial. Balancing safety, innovation and human rights protection requires prioritizing transparency, accountability and education to ensure AI systems enhance rather than undermine fundamental rights.

2.3. International Frameworks

The European Union's AI Act represents a stringent regulatory model categorizing AI systems by risk level, with strict obligations on high-risk applications and prohibitions on unacceptable ones. Its extraterritorial reach ensures compliance with standards of transparency, human oversight and accountability for systems impacting the EU market [1].

In contrast, international soft-law frameworks like the OECD AI Principles, UNESCO's Recommendation on the Ethics of AI and the Council of Europe's Framework Convention on AI emphasize voluntary principles of fairness, accountability and responsible governance. While influential in shaping global AI policy, these frameworks lack direct enforcement mechanisms.

The United States follows a decentralized, sector-specific approach, lacking a comprehensive federal AI law. Instead, it relies on existing statutes, agency guidance and state-level regulations. Notably, the National Institute of Standards and Technology (NIST) has issued a non-binding AI Risk Management Framework, promoting voluntary risk assessment principles for AI systems [6]. This fragmented landscape leads to inconsistencies and debates about the need for a cohesive federal strategy.

The divergence between the EU's legally binding approach and the U.S.'s market-driven, self-regulatory model reflects broader tensions in global AI governance. While international bodies push for regulatory alignment through high-level principles, differing enforcement strategies and legal traditions hinder cross-border interoperability.

The EU AI Act's influence is evident in regulatory discussions in Canada, Japan and Brazil, which are exploring risk-based models. However, global harmonization remains elusive due to differences in enforcement mechanisms and legal frameworks. As AI technologies advance, the interplay between binding regulations, voluntary principles and sector-specific guidelines will shape future governance, emphasizing the need for continued international cooperation to address AI's risks and benefits effectively.

This analysis highlights the fragmented nature of current AI governance and underscores the need for a comprehensive, interdisciplinary approach to AI impact assessment centered on fundamental human rights.

3. Standards and Guidelines

3.1. Standards for AI Assessment

The assessment of AI systems relies on established standards and frameworks providing guidance on risk management, transparency and accountability. Key standards include ISO/IEC [7], IEEE [8, 9] and frameworks developed by the National Institute of Standards and Technology (NIST) [10].

ISO/IEC 23894 [11] addresses risk management, aligning closely with the AI Act's regulatory requirements, by providing structured methods for risk identification, assessment and mitigation. ISO/IEC 25012 [12] focuses on data quality, emphasizing accuracy, completeness and consistency, essential for high-quality datasets used in AI training and operation. ISO/IEC TR 24027 [13] targets bias identification and mitigation to ensure fairness. Governance frameworks such as ISO/IEC 38507 [14] provide guidance on integrating AI into organizational structures to enhance accountability and oversight.

Further complementing these standards, ISO/IEC 42001 [15] and ISO/IEC 42005 [16] offer frameworks for managing AI systems throughout their lifecycle. ISO/IEC 42001 defines requirements for AI management systems, supporting continual monitoring, evaluation and ethical alignment. ISO/IEC 42005, still under development, aims to standardize AI impact assessments across social, environmental and economic dimensions, with guidance for integrating these assessments into risk management processes and maintaining transparency and accountability.

NIST frameworks also play a critical role. The AI Risk Management Framework (AI RMF) [6] serves as a flexible, voluntary guide for managing AI-related risks through a comprehensive and iterative process. The AI 600-1 standard [17] focuses specifically on the risks associated with generative AI technologies, including harmful content creation, bias and misuse of generated data. Additionally, the NIST Privacy Framework offers insights into managing privacy risks, a critical concern for AI systems handling sensitive or personal data.

IEEE standards, part of the IEEE Global Initiative on Ethics of Autonomous and Intelligent Systems, emphasize ethical AI development. IEEE 7002-2022 [18] provides guidelines for accountability, transparency, fairness and safety in AI systems, promoting responsible decision-making. IEEE 7010-2020 [8] offers a framework for assessing the impact of AI systems on human well-being, particularly in sensitive domains like healthcare and data privacy.

Collectively, these standards and frameworks address key aspects of AI assessment such as risk management, data quality, bias mitigation, governance and ethical considerations. However, they lack specificity for assessing compliance with the AI Act [1], requiring organizations to adapt and combine these guidelines to their unique use cases. Consequently, a tailored approach integrating multiple standards is essential to bridge the gaps and ensure comprehensive compliance with technical and regulatory requirements.

3.2. Guidelines from Research and Industry

Recent advancements in AI have led various institutions and stakeholders to establish frameworks for AI assessment and evaluation. The **Alan Turing Institute** proposes a robust framework prioritizing *transparency*, *accountability* and *robustness*, advocating for rigorous testing against adversarial scenarios and utilizing explainability tools such as SHAP (SHapley Additive exPlanations) and LIME (Local Interpretable Model-agnostic Explanations) for interpretability [19].

The **European Union Agency for Cybersecurity (ENISA)** focuses on *security* and *resilience*, recommending continuous monitoring, risk assessment and standardized metrics to assess AI system performance under various conditions [20].

The **Partnership on AI** emphasizes *fairness* and *bias mitigation*, advocating for fairness-aware algorithms and diverse datasets to minimize discriminatory outcomes [21].

These guidelines highlight essential aspects of AI assessment, including explainability, robustness, security and fairness, providing valuable insights that complement formal standards and regulatory frameworks.

3.3. Overview of Tools

As AI systems increasingly permeate critical domains, the potential for human rights violations arising from misuse or ethical misalignment grows. To address these challenges, various organizations have developed tools aimed at assessing and mitigating AI-related risks. Below, we compare some of the most prominent tools, highlighting their strengths and limitations.

Microsoft's Responsible AI Impact Assessment (RAIIA) provides a structured framework for ensuring responsible AI development and deployment. It offers templates and guidance for evaluating AI systems against principles such as fairness, reliability, transparency, privacy and inclusiveness. However, its reliance on qualitative assessments can limit consistency.

Google's AI Toolkit encompasses a suite of tools for responsible AI development, including Explainable AI (XAI), fairness indicators and model cards. While comprehensive, its effectiveness depends heavily

on the technical expertise of users and is not designed for regulatory compliance.

IBM's AI Fairness 360 (AIF360) is an open-source toolkit that offers metrics, algorithms and visualization tools to detect and mitigate bias in machine learning models. Its strength lies in its transparency and accessibility; however, its focus is mainly on fairness, lacking broader governance and ethical considerations.

OpenAI's Guidelines emphasize responsible use of large language models and generative AI systems. While offering valuable best practices, these guidelines remain high-level and are not directly applicable to compliance with specific regulatory frameworks.

Ethical AI Toolkit by the Montreal AI Ethics Institute focuses on societal impact, providing worksheets for ethical impact assessments. While promoting a holistic approach, it lacks technical depth and automation, making it less practical for large-scale AI deployment.

Hugging Face's Model Evaluation Tools offer insights into performance and fairness for pre-trained NLP models. Although effective in enhancing explainability, their applicability is limited to specific model types and lacks comprehensive governance features.

These tools highlight diverse approaches to AI assessment, from fairness-focused toolkits to broader ethical frameworks. However, many lack integration with formal regulatory requirements, underscoring the need for more comprehensive and adaptable assessment methodologies.

3.4. Comparison and Insights

Comparing existing standards, guidelines and tools reveals varying strengths and limitations in AI impact assessment. While standards like ISO/IEC 23894 and 42001 provide structured risk management frameworks, they often lack concrete metrics for ethical assessment and broader societal impacts. ISO/IEC 25012 focuses on data quality but is not tailored for comprehensive AI assessment. NIST frameworks (AI RMF, AI 600-1) offer robust technical guidance but may be resource-intensive for smaller organizations and insufficient for addressing ethical and human rights concerns. IEEE standards, particularly IEEE 7002-2022 and 7010-2020, emphasize ethics and societal impacts but remain high-level and lack practical implementation steps.

Key insights from Table 1:

- **Transparency and Accountability:** standards like ISO/IEC 42001 and Microsoft's RAIIA emphasize structured governance and accountability mechanisms.
- **Technical Guidance:** NIST frameworks provide comprehensive guidance for managing AI-related risks, though with a strong focus on technical implementation.
- **Ethics and Human-Centricity:** IEEE 7002-2022 and 7010-2020 highlight ethical considerations but lack practical guidelines for real-world deployment.
- **Bias Mitigation:** IBM's AI Fairness 360 offers concrete tools for addressing fairness, but with limited scope for broader AI governance.
- **Scalability Issues:** tools like Microsoft's RAIIA require substantial resources and expertise, making them difficult to implement for smaller organizations.

To address the limitations identified in Table 1, it is essential to integrate multiple frameworks and tools, leveraging their strengths while mitigating their weaknesses. Future efforts should focus on enhancing interdisciplinary collaboration, improving accessibility and developing comprehensive assessment methodologies that align with both ethical and regulatory standards.

4. Proposed Methodology for AI Assessment

4.1. Overview of the Methodology

This chapter introduces the *Fundamental Rights Impact Assessment* (FRIA) methodology by HH4AI, specifically developed to assess and mitigate the potential impacts of systems on fundamental rights. The current methodology is designed for organizations seeking compliance with the AI Act while ensuring

Standard / Tool	Strengths	Limitations
ISO/IEC 23894	Comprehensive risk management framework; Emphasizes continuous monitoring and improvement; Aligned with regulatory requirements like the AI Act.	Limited scope beyond safety, security and ethical implications; Lacks guidelines for societal impacts and non-technical applications.
ISO/IEC 25012	Defines essential data quality requirements; Provides structured criteria for evaluating data accuracy, completeness and consistency.	Limited focus on AI-specific concerns; Primarily addresses data quality, not ethical or societal impacts.
ISO/IEC 42001	Provides governance principles for AI systems; Focuses on accountability, transparency and continuous assessment; Promotes ongoing monitoring and ethical alignment.	Lacks prescriptive metrics for specific AI systems; Complex to implement without established governance structures.
NIST AI RMF	Comprehensive, risk-based framework; Covers governance, transparency and performance evaluation; Useful for varied sectors.	Resource-intensive and complex to implement; Lacks prescriptive guidance for ethical and societal concerns.
NIST AI 600-1	Specialized for generative AI systems; Strong focus on security, privacy and risk assessment.	Limited focus on broader societal, ethical and human rights concerns; Primarily technical and security-oriented.
IEEE 7002-2022	Emphasizes ethical design, human-centric approaches and accountability; Useful for guiding responsible AI decision-making.	High-level guidance with limited operationalization; Requires adaptation to specific use cases.
IEEE 7010-2020	Focuses on assessing societal impacts, particularly well-being; Encourages long-term, human-centric AI design.	Limited applicability to broader ethical and governance concerns; Challenges in quantifying societal impacts across various domains.
Microsoft's RAIIA	Comprehensive templates for assessing fairness, reliability, transparency, privacy and inclusiveness; Scalability across various sectors.	Resource-intensive; Limited automation; Proprietary nature tied to Microsoft's ecosystem.
Google's AI Toolkit	Includes explainability tools, fairness indicators and model cards; Open-source availability promotes accessibility.	Primarily technical focus; Lacks comprehensive societal risk assessment; High learning curve for effective use.
IBM's AI Fairness 360	Specialized in detecting and mitigating bias; Open-source framework with comprehensive fairness metrics.	Limited applicability beyond fairness; Requires advanced knowledge of machine learning; Not optimized for large-scale systems.
OpenAI's Use Case Guidelines	Tailored recommendations for specific AI applications; Provides ethical considerations and safety guidelines.	High-level conceptual guidance; Lacks detailed implementation strategies; No dedicated software or automated frameworks.

Table 1
Comparison of AI-related standards, guidelines and tools: strengths and limitations.

that their systems adhere to fundamental human rights principles. By employing a *gate-based* structure with three main phases plus a concluding output stage (see Figure 1), the methodology streamlines the analysis process and ensures that only the most relevant impact progress to detailed evaluation.

At the core of the methodology is a structured assessment framework based on well-defined **impact domains** and **guiding criteria**. The impact domains cover key dimensions of AI-related impacts, including *Data Governance*, *Human Oversight and Control* and *Fairness and Non-Discrimination*. These guiding criteria serve as reference points for assessing AI systems' alignment with fundamental rights and regulatory requirements.

To ensure relevance and efficiency, the methodology employs a **filtering mechanism** driven by key factors, referred to as "drivers", such as the type of system, its life cycle stage and its domain of application. This structured filtering ensures that only applicable impacts and evaluation criteria are considered, avoiding unnecessary assessments. The Human Rights Checklist in Phase 1 serves as the primary tool for this evaluation, presenting targeted questions that assess whether an AI system's functionalities pose impacts warranting deeper analysis. Based on the results of this phase, the methodology identifies which impacts need further examination through defined **impact scenarios**.

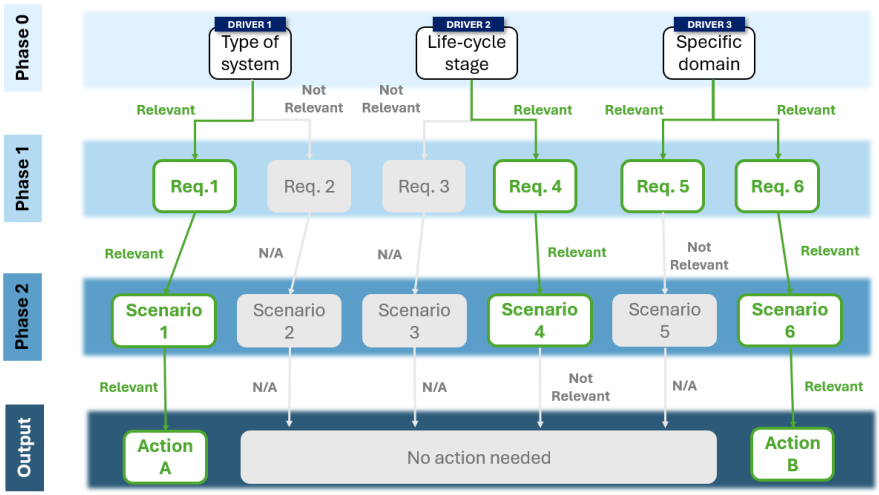

Figure 1: Overview of the FRIA Methodology: a gate-based impact assessment framework.

Impact scenarios play a crucial role in the methodology, illustrating concrete situations where an AI system could compromise fundamental rights. Each scenario undergoes a structured **self-evaluation**, assessing its relevance, severity and the effectiveness of existing impact mitigation measures. This evaluation considers multiple dimensions, including the impact on individuals and society, the difficulty of reversing potential harm and the duration of the consequences. Scenarios classified as relevant trigger specific remediation actions to mitigate impacts.

Building on this structured foundation, the methodology advances through three progressive phases, introduced at a high level earlier, which are described in detail in Section 4.2. Upon completion of the assessment, the methodology generates a comprehensive **final output**, as explained in Section 4.3. This output consolidates the assessment findings in both graphical and tabular form, summarizing identified impacts, the effectiveness of existing controls and recommended mitigation actions. In doing so, it provides decision-makers with a clear, actionable overview of the AI system's impact, thereby facilitating effective impact management and regulatory compliance.

A key differentiator of this methodology is its *gate-based* approach, ensuring efficiency by progressively refining the analysis and focusing only on the most relevant impacts. This stepwise refinement prevents unnecessary assessments, optimizes resource allocation and enhances the clarity of impact evaluation. The methodology's structured yet flexible design allows it to adapt to various AI applications while maintaining a rigorous human rights framework. The benefits of this approach extend beyond compliance; by embedding ethical considerations and proactive impact management into the AI life cycle, it enhances transparency, accountability and trust in AI systems. These aspects, along with other key advantages, are explored in Section 4.4, where the methodology's innovations and benefits are analyzed in detail.

Finally, Section 4.5 presents concluding reflections on the methodology's strengths, particularly its structured adaptability and role in reinforcing human rights protections throughout the AI system's life cycle. This final discussion underscores how the methodology ensures a systematic and effective approach to human rights impact assessment, supporting both regulatory compliance and ethical AI governance.

4.2. Phases of the Methodology

We present here a detailed explanation of each phase of the methodology, describing the key elements that compose each phase, their interactions, the specific outputs they produce and their connection to the subsequent phase.

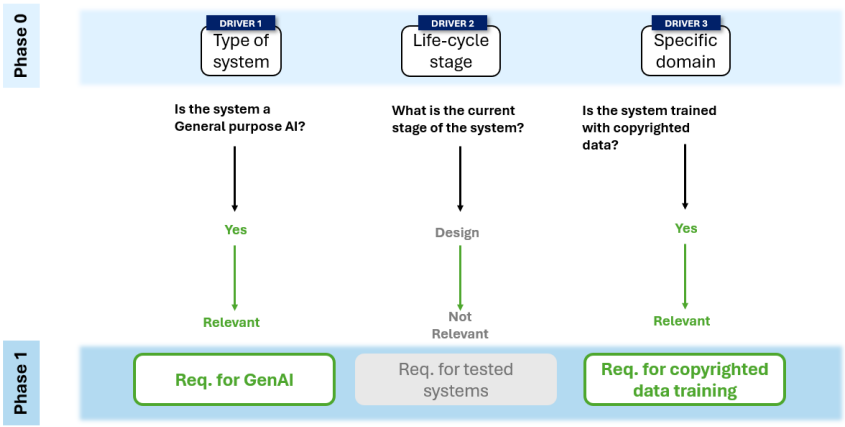

Figure 2: Transition from Phase 0 to Phase 1: identifying relevant requirements.

4.2.1. Phase 0 - AI System Overview

Phase 0 establishes the foundation for the impact assessment process by gathering essential information about the AI system. It defines the system's purpose, identifies key stakeholders and outlines the operational context. Additionally, it includes domain applicability questions to determine whether the system operates in sensitive areas, such as biometric data, processing or critical decision-making, which influence the selection of checklist questions in Phase 1. Similarly, it defines the system's life cycle stage (e.g., development, deployment or post-deployment), ensuring that the subsequent assessment is tailored to its current state.

Another crucial aspect of this phase is establishing a dedicated process for maintaining and updating the AI System Overview, including clear accountability for the individuals responsible. This ensures that the assessment remains accurate and reflects any changes to the system over time. By setting out these responsibilities and procedures from the outset, the output of Phase 0 provides a clear and well-defined scope for the assessment, laying the groundwork for identifying potential impacts in the following phase.

As shown in Figure 2, the transition from Phase 0 to Phase 1 follows a structured filtering process. This ensures that only the most relevant requirements proceed for further evaluation, optimizing the efficiency of the assessment.

4.2.2. Phase 1 - Human Rights Checklist

Phase 1 systematically identifies potential human rights impacts through a structured Human Rights Checklist. This checklist is designed to assess the AI system's impact by linking each evaluation question to *guiding criteria*, which are directly mapped to fundamental rights.

To ensure contextual relevance, the checklist questions are dynamically filtered based on two key factors: the system's life cycle stage and its domain applicability. This tailored approach ensures that only questions relevant to the specific AI system under evaluation are considered. Each checklist item is also assigned to specific internal stakeholders, ensuring that subject-matter experts evaluate the areas where they have direct oversight and expertise.

The relevance of each criterion is determined through the responses to the checklist. If a criterion receives a high relevance score, indicating a potentially significant impact on fundamental rights in the context of the specific AI system under evaluation, then the assessment proceeds to Phase 2, where a more detailed analysis is conducted. This transition from Phase 1 to Phase 2 follows a structured filtering process, as illustrated in Figure 3, ensuring that only the most critical impacts advance to deeper evaluation while optimizing efficiency.

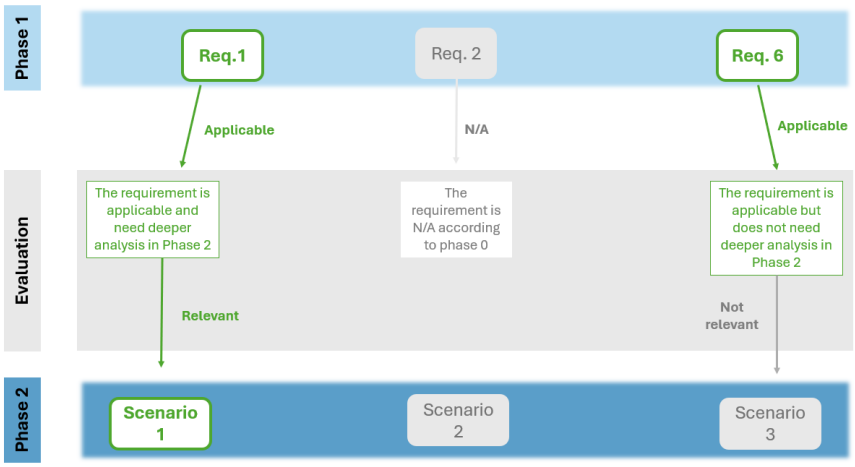

Figure 3: Transition from Phase 1 to Phase 2: identifying relevant impact scenarios.

4.2.3. Phase 2 - Impact Assessment

Phase 2 involves a detailed evaluation of the impacts identified in Phase 1, focusing on *multiple impact scenarios* for each guiding criterion. These scenarios are designed to assess a wide range of potential impacts to fundamental rights, including ethical, legal and social implications. The internal stakeholder responsible for each criterion conducts this assessment, determining whether effective controls exist within the organization to mitigate the identified impacts. Stakeholders are required to provide documentation or other evidence demonstrating the effectiveness of these controls, as well as to specify the individual or department responsible for maintaining and overseeing them.

The impact assessment considers multiple evaluation dimensions to ensure a comprehensive understanding of each impact scenario. Stakeholders assess:

- The **effect on individuals**, analyzing the potential impact on individual rights (e.g., privacy violations, discrimination).
- The **effect on society**, considering broader societal implications (e.g., increased inequality, biases in decision-making).
- The **effort required to mitigate or reverse the impact**, evaluating how difficult it would be to address the issue once it has occurred.
- The **duration of the effect**, estimating whether the impact is short-term, long-term or potentially irreversible.

The evaluation process is structured around a three-level self-evaluation scale, where each impact scenario is classified as:

- **Relevant**: the scenario poses a significant impact to fundamental rights and requires immediate action.
- **Partially Relevant**: the scenario presents moderate impacts that may require intervention but are not immediately critical.
- **Irrelevant**: the scenario does not apply or has no meaningful impact on fundamental rights.

For each scenario assessed as Relevant or Partially Relevant, a remedial action is proposed to mitigate the identified impact. The remediation process includes:

- **Action Type**: the category of intervention (e.g., policy revision, additional control implementation, training or awareness programs).
- **Action Description**: a detailed explanation of the corrective measure and how it will mitigate the identified impact.

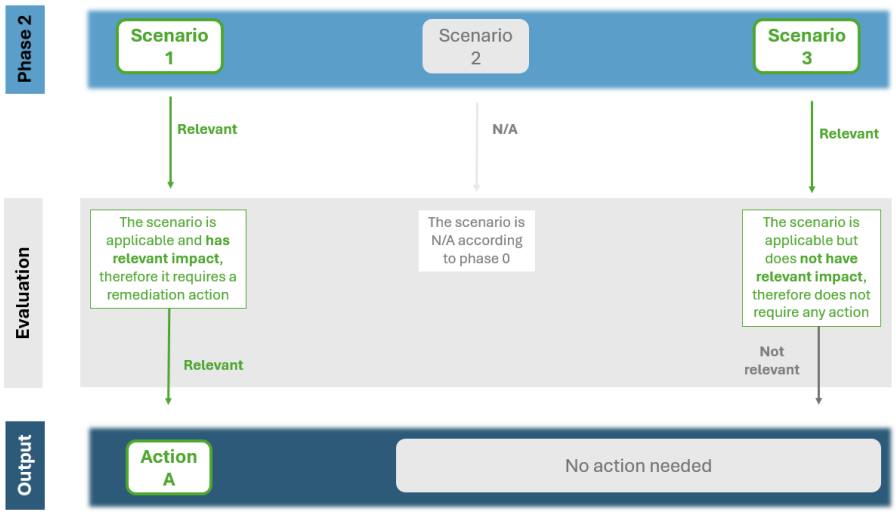

Figure 4: Transition from Phase 2 to Output: identifying required remediation actions.

- **Action Owner**: the responsible individual, team or department ensuring the implementation and effectiveness of the corrective action.

Once all impact scenarios have been evaluated and appropriate remedial actions suggested, the final classification of the impact on fundamental rights is determined for each guiding criterion. If multiple relevant impact scenarios are identified, additional mitigation strategies may be necessary to ensure compliance and impact reduction. However, if most scenarios are classified as Irrelevant, no further action or in-depth analysis is required for that specific criterion.

This structured, multi-dimensional approach ensures that AI-related impacts to fundamental rights are systematically identified, assessed and mitigated, while maintaining accountability and transparency throughout the process.

As illustrated in Figure 4, the transition from Phase 2 to the Output stage ensures that only scenarios classified as relevant and having a significant impact require corrective actions. If a scenario is deemed relevant but without a significant impact, no further action is required. Scenarios classified as not relevant are excluded from the final output. This structured filtering approach ensures that remediation efforts are targeted, efficient and aligned with the identified impacts, maintaining an effective and accountable impact assessment process.

4.3. Final Output

The final output of the Fundamental Rights Impact Assessment provides a comprehensive summary of the assessment results across all phases. This output consists of both graphical and tabular representations to facilitate a clear and structured interpretation of the evaluation process.

The tabular overview presents a structured breakdown of the assessments and evaluations conducted in Phase 1 and Phase 2, detailing relevance scores, stakeholder responses and identified impacts. The graphical overview complements this by offering a visual representation of key insights, ensuring an intuitive and easily digestible format for decision-makers.

The final output is structured into two primary components:

- An overview of results, which includes both the graphical and tabular representations of the assessment conducted in Phase 1 (requirements analysis) and Phase 2 (impact scenario evaluation).
- A remediation actions section, detailing the list of required actions, their types and the responsible stakeholders for implementation.

The final output ensures that all identified impacts and corresponding remediation actions are documented in a structured manner. The graphical and tabular overviews provide a clear impact profile, while the remediation section ensures accountability by assigning ownership to corrective actions. This comprehensive output enables decision-makers to track, evaluate and implement impact mitigation strategies effectively, ensuring that fundamental rights considerations are addressed throughout the AI system's life cycle.

4.4. Innovation and Benefits

The FRIA methodology introduces several key innovations and benefits, enhancing the effectiveness and applicability of AI impact assessments while ensuring a structured and actionable approach to impact mitigation.

- **Detailed impact Scenario Analysis**: by defining multiple scenarios for each guiding criterion, the methodology enables a comprehensive evaluation of potential impacts. This granular approach ensures a thorough understanding of how an AI system may impact fundamental rights and allows for the development of precise, targeted mitigation strategies.
- **Stakeholder-Driven Evaluation**: the assessment process integrates the expertise of internal stakeholders, leveraging their real-world insights into system design, deployment and governance. This ensures that impact identification and mitigation strategies are based on practical knowledge of existing controls and operational impacts.
- **Self-Evaluation Scale**: a standardized three-level scale (Relevant, Partially Relevant or Irrelevant) quantifies the significance of each identified impact. This structured approach facilitates clear decision-making and ensures that only substantial impacts advance to deeper analysis and remediation.
- **Human Rights Mapping**: impacts and scenarios are systematically categorized based on guiding criteria linked to fundamental rights. This structured alignment provides organizations with a transparent, legally grounded understanding of how AI functionalities may affect individual rights.
- **Flexibility and Context-Specific Adaptation**: the methodology adapts to different AI use cases by tailoring the assessment based on the system's domain and life cycle stage. This ensures that organizations focus on relevant impacts without performing unnecessary evaluations.
- **Proactive impact Mitigation**: beyond identifying impacts, the methodology prescribes concrete remedial actions for scenarios deemed Relevant or Partially Relevant. These interventions, ranging from policy revisions to technical controls and training programs, ensure that the assessment process is solution-oriented, actively supporting organizations in enhancing compliance and minimizing potential harm.

4.5. Final Remarks

The FRIA methodology provides a structured, systematic and scalable framework for assessing and mitigating the impact of AI systems on fundamental rights. By following a gate-based approach, it ensures that only the most relevant impacts undergo detailed evaluation, optimizing resources while maintaining a high level of scrutiny. This structured assessment process enables organizations to integrate ethical considerations, regulatory compliance and impact management into AI development and deployment strategies.

The methodology not only identifies and evaluates impacts but also assesses the effectiveness of existing safeguards and establishes accountability for their continuous monitoring. The final output offers a comprehensive overview of impact levels and required remediation actions, ensuring that decision-makers have a clear understanding of potential impacts and the necessary steps to mitigate them. This structured approach enhances transparency in AI governance, making impact assessment results both accessible and actionable.

Beyond regulatory compliance, the methodology fosters a proactive approach to responsible AI development by embedding fundamental rights considerations throughout the AI system life cycle. This allows organizations to move beyond a reactive compliance mindset toward continuous improvement in AI ethics and governance. The structured remediation process ensures that identified impacts are not only acknowledged but also addressed through concrete actions, reinforcing accountability and fostering trust in AI systems.

By systematically aligning AI impact assessment with human rights principles and governance best practices, the HH4AI FRIA methodology supports organizations in achieving AI accountability, regulatory alignment and ethical governance. It provides a robust framework for mitigating AI-related impacts while promoting sustainable and responsible AI development, ensuring that fundamental rights remain a priority in the design, deployment and operation of AI systems.

5. Conclusion

The proposed gate-based framework offers a structured and scalable approach to assessing AI systems' impacts on fundamental rights. Through its phased structure and filtering mechanism, it prioritizes critical risks, enhancing compliance with emerging regulations while promoting transparency, accountability and ethical governance across AI life cycles.

The methodology achieves a balance between flexibility and rigor by adapting to diverse AI applications while ensuring that accountability, literacy and data governance are systematically addressed. Its scenario-based approach allows targeted scrutiny of high-risk functionalities, optimizing resource allocation and providing clear remediation processes.

However, challenges persist, particularly in adapting the framework to various regulatory environments and rapidly evolving AI technologies. Effective implementation relies on organizational maturity, access to specialized personnel and robust governance structures. Furthermore, the framework's applicability across sectors may require tailored adaptations to accommodate specific regulatory or ethical requirements.

Future work aims to enhance the methodology by integrating quantitative metrics within Phase 2, particularly for evaluating fairness, reliability and transparency. Incorporating numerical indicators will sharpen risk estimation, facilitate benchmarking across AI systems and provide a more comprehensive basis for evidence-based remediation. Continued refinement of assessment techniques, coupled with broader stakeholder engagement, will further improve the framework's adaptability, rigor and relevance.

Ultimately, the methodology offers a practical tool for aligning technical measures with ethical principles and regulatory requirements. By promoting transparency, accountability and trust, it supports responsible AI development and deployment that prioritizes fundamental rights. Its structured approach provides a foundation for future enhancements, ensuring that AI systems remain compliant, ethical and beneficial in diverse application domains.

Acknowledgments

The work reported in this paper has been partly funded by the European Union - NextGenerationEU, under the National Recovery and Resilience Plan (NRRP) Mission 4 Component 2 Investment Line 1.5: Strengthening of research structures and creation of R&D "innovation ecosystems", set up of "territorial leaders in R&D", within the project "MUSA - Multilayered Urban Sustainability Action" (contract n. ECS 00000037).

Declaration on Generative AI

The authors have not employed any Generative AI tools in the preparation of this paper.

References

[1] E. Union, Regulation (eu) 2024/1689 of the european parliament and of the council, 2024.

[2] E. Union, Regulation (eu) 2016/679 of the european parliament and of the council, 2016.

[3] I. Bartle, P. Vass, Self-regulation and the regulatory state: A survey of policy and practice, Citeseer, 2005.

[4] M. Fichter, Voluntary regulation: codes of practice and framework agreements, in: Comparative Employment Relations in the Global Economy, Routledge, 2013, pp. 414–430.

[5] M. C. Antonucci, N. Scocchi, Codes of conduct and practical recommendations as tools for self-regulation and soft regulation in eu public affairs, Journal of Public Affairs 18 (2018) e1850.

[6] NIST, Artificial Intelligence Risk Management Framework - Technical Report NIST AI 100-1, 2023. URL: https://www.nist.gov/news-events/news/2023/01/nist-releases-draft-artificial-intelligence-risk-management-framework.

[7] J. Oviedo, M. Rodriguez, A. Trenta, D. Cannas, D. Natale, M. Piattini, Iso/iec quality standards for ai engineering, Computer Science Review 54 (2024) 100681.

[8] D. Schiff, A. Ayesh, L. Musikanski, J. C. Havens, Ieee 7010: A new standard for assessing the well-being implications of artificial intelligence, in: 2020 IEEE international conference on systems, man, and cybernetics (SMC), Ieee, 2020, pp. 2746–2753.

[9] A. F. Winfield, S. Booth, L. A. Dennis, T. Egawa, H. Hastie, N. Jacobs, R. I. Muttram, J. I. Olszewska, F. Rajabiyazdi, A. Theodorou, et al., Ieee p7001: A proposed standard on transparency, Frontiers in Robotics and AI 8 (2021) 665729.

[10] J. Dunietz, E. Tabassi, M. Latonero, K. Roberts, A plan for global engagement on ai standards, 2024. URL: https://tsapps.nist.gov/publication/get_pdf.cfm?pub_id=958389. doi:https://doi.org/10.6028/NIST.AI.100-5.

[11] L. Floridi, On the brussels-washington consensus about the legal definition of artificial intelligence, Philosophy & technology 36 (2023) 87.

[12] F. Gualo, M. Rodríguez, J. Verdugo, I. Caballero, M. Piattini, Data quality certification using iso/iec 25012: Industrial experiences, Journal of Systems and Software 176 (2021) 110938.

[13] ISO/IEC, Iso/iec tr 24027: 2021 information technology–artificial intelligence (ai)–bias in ai systems and ai aided decision making, 2021.

[14] ISO/IEC, Iso/iec tr 38507: 2022 information technology — governance of it — governance implications of the use of artificial intelligence by organizations, 2022.

[15] ISO/IEC, Iso/iec 42001:2023 information technology — artificial intelligence — management system, 2023.

[16] ISO/IEC FDIS 42005: Information technology — Artificial intelligence — AI system impact assessment - Draft, 2025. URL: https://www.iso.org/standard/44545.html.

[17] C. Autio, R. Schwartz, J. Dunietz, S. Jain, M. Stanley, E. Tabassi, P. Hall, K. Roberts, NIST Trustworthy and Responsible AI - 600-1. Artificial Intelligence Risk Management Framework: Generative Artificial Intelligence Profile, 2024. URL: https://tsapps.nist.gov/publication/get_pdf.cfm?pub_id=958388. doi:https://doi.org/10.6028/NIST.AI.600-1.

[18] J. I. Olszewska, D. C. S. Committee, et al., Ieee standard for data privacy process: Ieee standard 7002-2022 (2022).

[19] D. Leslie, Understanding artificial intelligence ethics and safety, arXiv preprint arXiv:1906.05684 (2019).

[20] S. Ntalampiras, G. Misuraca, P. Rossel, Artificial intelligence and cybersecurity research, ENISA (2023).

[21] J. Heer, The partnership on ai, AI Matters 4 (2018) 25–26.