From Explainability to Accountability: XAI and Fundamental Rights Impact Assessments under the EU AI Act

Lovro Bobnar
Department of International Law
University of Ljubljana, Faculty of Law
Institute for Comparative Law at the Faculty of Law in Ljubljana
Ljubljana, Slovenia
lovro.bobnar@pf.uni-lj.si

Abstract

This paper examines the relationship between Explainable Artificial Intelligence (XAI) and Fundamental Rights Impact Assessment (FRIA) under the European Union Artificial Intelligence Act (AI Act), focusing on whether technical explainability can meaningfully contribute to legal accountability and the protection of fundamental rights. While Article 13 of the AI Act requires sufficient transparency for high-risk artificial intelligence (AI) systems and XAI techniques can support interpretability and human oversight, technical explanations alone cannot establish whether AI-driven decisions are lawful, fair, or non-discriminatory. Article 27 therefore introduces the FRIA as a broader mechanism for identifying and mitigating fundamental rights risks arising from the deployment of high-risk AI systems in specific contexts. However, existing impact assessment approaches may inadequately capture systemic and interconnected harms where rights are assessed in isolation. The paper proposes a three-layer framework connecting technical explanation, operational use and legal justification, informed by intersectionality and vulnerability. A qualitative application to published credit-risk explanations illustrates how XAI findings can inform FRIA enquiries, evidence requests, safeguards and deployment decisions.

Keywords

Explainable Artificial Intelligence; Fundamental Rights Impact Assessment; EU Artificial Intelligence Act; fundamental rights; algorithmic transparency; accountability.

1 Introduction

Regulation (EU) 2024/1689, commonly referred to as the AI Act, establishes a comprehensive framework for the development, placing on the market and use of AI systems in the European Union (EU). Its objectives extend beyond technological safety. The Regulation also seeks to ensure a high level of protection of health, safety and fundamental rights, while supporting innovation and the functioning of the internal market [1].

The AI Act follows a risk-based approach. It prohibits specifically defined practices that conflict with fundamental EU values, including human dignity, freedom, equality, democracy and the rule of law. Examples include certain forms of social scoring and harmful exploitation of vulnerabilities related to age, disability or a person's social or economic situation. High-risk systems are instead subject to the requirements laid down in Articles 8 to 15, including risk management under Article 9, data governance under Article 10, technical documentation and record-keeping under Articles 11 and 12, transparency under Article 13, human oversight under Article 14, and accuracy, robustness and cybersecurity under Article 15. Under Articles 43, 47 and 48, the provider must complete the applicable conformity assessment, draw up an EU declaration of conformity and affix the CE marking as a visible declaration of compliance. The marking does not, however, guarantee that every deployment or AI-assisted decision will comply with fundamental rights. A technically compliant system may still produce unacceptable consequences in a particular institutional setting or when used on a population that differs from the one on which it was tested.

This distinction explains the relationship between Articles 13 and 27 of the AI Act. Article 13(1) requires high-risk AI systems to be sufficiently transparent to enable deployers to interpret their outputs and use them appropriately. Article 27(1), in turn, requires specified deployers to assess the effects that the use of a high-risk AI system may have on fundamental rights. The connection between these obligations is made explicit in Article 27(1)(d), which requires the deployer to identify specific risks of harm while considering the information supplied by the provider under Article 13. XAI may therefore support a FRIA by making the functioning or outputs of a system more intelligible. Nevertheless, explainability and accountability are not equivalent. XAI primarily addresses a descriptive question: how did the system produce a particular output? A legal assessment of the deployment must ask broader questions: was the use of that output lawful and justified, who is responsible for the resulting decision, and what remedies are available if the decision is unlawful or harmful? This paper argues that XAI should serve as a source of technical evidence within a FRIA. It combines legal analysis with a structured qualitative application of a three-layer framework comprising technical explanation, operational explanation and legal justification. Drawing on published credit-

Permission to make digital or hard copies of part or all of this work for personal or classroom use is granted without fee provided that copies are not made or distributed for profit or commercial advantage and that copies bear this notice and the full citation on the first page. Copyrights for third-party components of this work must be honored. For all other uses, contact the owner/author(s).
Information Society 2026, 5–9 October 2026, Ljubljana, Slovenia
© 2026 Copyright held by the owner/author(s).
https://doi.org/10.70314/is.2026.sikdd.120

risk explanations, the application shows how an XAI finding can be connected to its evidential limitations, potentially affected rights, further information requirements and a deployment decision. Published technical findings are distinguished from hypothetical deployment conditions and the author's normative assessment.

2 The Limits of Technical Explainability

Article 13(2) requires the instructions for use accompanying a high-risk AI system to contain concise, complete, correct and clear information that is relevant, accessible and comprehensible to deployers. Article 13(3) further requires information about the system's intended purpose, capabilities, limitations, expected accuracy, foreseeable risks and performance concerning particular persons or groups. Where applicable, the instructions must also include information relevant to explaining and interpreting the system's outputs under Article 13(3)(b)(iv) and (vii), as well as the human-oversight measures referred to in Article 13(3)(d). The provision does not prescribe a particular XAI method or establish a universal metric for explainability. Indeed, the AI Act neither generally mandates the use of XAI nor prohibits black-box models. XAI is therefore one possible means of supporting transparency and human oversight rather than a legally prescribed solution [2], [3].

The appropriate explanation depends on its audience and purpose. Developers may need a global account of the variables and patterns on which a model relies. A doctor, civil servant or credit officer may require an explanation of an individual recommendation and its uncertainty. An affected person may instead need to know which information was decisive, how the output influenced the final decision and how that decision can be challenged. Research therefore supports an audience-specific and lifecycle-based approach [4], [5].

Feature-attribution methods describe the contribution assigned to input features in explaining a model's prediction. SHAP (SHapley Additive exPlanations), for example, indicates how individual input features contribute to a particular prediction relative to a reference prediction [14]. An explanation concerning one individual should be distinguished from a summary of explanations across many individuals: the latter may reveal broader patterns but does not explain every individual decision [15]. For a FRIA, these outputs can help identify features requiring further scrutiny. They do not, by themselves, establish whether reliance on those features is legally permissible or whether the resulting treatment constitutes discrimination [7].

Explanations produced after a model has generated an output, commonly described as *post-hoc* explanations, may approximate its behaviour rather than reveal its actual logic. Their results may vary depending on the method, data or parameters selected. In high-stakes contexts, an inherently interpretable model may therefore be preferable to adding an explanation layer to an unnecessarily opaque system [6]. Explanations should at least be evaluated for fidelity, stability, comprehensibility and their capacity to communicate uncertainty.

Explainability does not independently guarantee non-discrimination. Removing protected characteristics from a dataset does not prevent other variables from functioning as proxies, while historical inequality may already be encoded in labels, institutional practices and definitions of success. Statistical fairness metrics can help identify disparities, but they cannot determine compliance with EU non-discrimination law without contextual legal analysis [7]. More broadly, treating fairness as a purely technical property can obscure the institutional and social environment in which an AI system operates [8].

XAI may even increase risk when a persuasive explanation creates unjustified confidence in an inaccurate output. Article 14(4)(b) expressly requires persons responsible for human oversight to remain aware of the risk of automatically relying or over-relying on an AI system's output. Articles 14(4)(c) and (d) further require them to be able to interpret the output and to decide not to use it or to disregard, override or reverse it. Article 26(2) complements these requirements by obliging deployers to assign human oversight to persons with the necessary competence, training, authority and support. Effective oversight therefore requires more than the nominal presence of a human in the decision-making process [2].

3 The Fundamental Rights Impact Assessment as an Accountability Mechanism

Article 27 shifts attention from the AI system in isolation to its specific context of use. Under Article 27(1), the obligation to conduct a FRIA applies to deployers that are bodies governed by public law, private entities providing public services, and deployers of high-risk systems referred to in points 5(b) and (c) of Annex III. High-risk systems used in the area listed in point 2 of Annex III are excluded from this obligation. Article 27(2) clarifies that the assessment is required for the first use of the system. Following the 2026 amendment to Article 113, third paragraph, point (c)(i), these provisions will apply from 2 December 2027 to systems classified as high-risk under Article 6(2) and Annex III [9]. Private employers are therefore generally outside the scope of Article 27 unless they fall within one of the specifically listed categories.

Article 27(1)(a) to (f) directly defines the required content of a FRIA. Points (a) and (b) require a description of the deployer's processes and the intended duration and frequency of use. Points (c) and (d) require the identification of affected persons and groups and the specific risks of harm, considering the information supplied by the provider under Article 13. Points (e) and (f) require a description of human-oversight measures and the measures to be taken if risks materialise, including internal governance and complaint mechanisms. Under Article 27(2), the assessment must be updated where any of these elements changes or is no longer current.

Article 27(4) permits relevant parts of a DPIA to be incorporated into, or cross-referenced in, the FRIA where they already satisfy the corresponding obligations [9]. The distinction between the assessments should not be reduced to privacy versus other fundamental rights. DPIAs address risks arising from personal-data processing and may also concern rights such as non-discrimination [16]. A FRIA addresses the deployment of the AI system within the scope of Article 27. Existing DPIA findings can therefore be reused, while any outstanding requirements concerning affected groups, deployment-specific harms, human oversight and complaint arrangements must still be addressed.

The principal value of a FRIA is its recognition that rights are affected by a complete socio-technical process. The same score

may create different risks depending on whether it is advisory, presumptive or practically decisive. Its impact also depends on input data, operator incentives, available alternatives and opportunities to challenge the result [8], [10].

Article 27 nevertheless has structural limitations. Its restricted scope may exclude systems that influence important decisions. Providers may understand the model but lack knowledge of its context, while deployers may understand their processes but lack access to technical evidence. The provision also lists assessment elements without establishing a complete methodology or generally requiring the participation of affected persons [10], [11].

A conventional matrix based only on the probability and severity of harm is therefore insufficient. Risks involving absolute rights cannot be accepted merely because a deployment offers economic or administrative benefits. Where limitations of other rights are possible, Article 52(1) of the EU Charter requires respect for the essence of the right, a legitimate objective, necessity and proportionality [12]. The FRIA should consequently combine technical risk analysis with substantive legal assessment. Article 85 complements the preventive function of the FRIA with an *ex post* accountability mechanism by allowing any natural or legal person having grounds to consider that the Regulation has been infringed to lodge a complaint with the relevant market surveillance authority. In addition, under the conditions set out in Article 86(1), affected persons may obtain a clear and meaningful explanation of the role of an Annex III high-risk AI system in a decision producing legal or similarly significant effects and of the main elements of that decision.

4 Intersectionality and Vulnerability

Article 27(1)(c) expressly requires the identification of "the categories of natural persons and groups likely to be affected" in the specific context of use. The provision therefore supplies a legal basis for examining differentiated impacts, but it does not specify how deployers should account for intersecting forms of disadvantage. Assessments organised around isolated categories such as sex, race, age or disability may overlook harms experienced at the intersection of several characteristics. The AI Act also recognises vulnerability in Article 5(1)(b), which prohibits certain AI practices that exploit vulnerabilities related to age, disability or a specific social or economic situation. However, this provision addresses vulnerability only within the narrowly defined context of prohibited exploitative practices and does not provide a general methodology for identifying vulnerability within a FRIA.

A recruitment system might produce similar aggregate outcomes for women and men while disproportionately disadvantaging older migrant women. An education system may achieve high average accuracy but perform poorly for children who combine a linguistic minority background with a learning disability. Small intersectional groups can disappear within aggregate statistics even where the consequences for their members are serious.

Recent scholarship therefore proposes intersectionality and vulnerability as correctives to FRIA models organised around isolated rights and fixed categories [11]. Vulnerability should not be understood only as a permanent personal characteristic. It may also arise from institutional dependency, limited digital literacy, language barriers, economic disadvantage or the absence of realistic alternatives and remedies. The relevant inquiry is not merely which group a person belongs to, but how a particular deployment distributes decision-making power, exposure to harm and the capacity to challenge decisions.

Intersectional analysis should combine disaggregated quantitative testing with qualitative evidence and, where appropriate, participation by affected persons. The Council of Europe's HUDERIA methodology provides a useful model by combining context-based risk analysis, stakeholder engagement, substantive assessment, mitigation planning and iterative review [13]. Although HUDERIA is non-legally binding and cannot replace Article 27, it can help operationalise its relatively open requirements.

5 From Explainability to Accountability

The gap between XAI and accountability cannot be resolved through a single explanation technique. This paper proposes that an effective FRIA should integrate three complementary layers, each anchored in different provisions of the AI Act. Technical explanation, informed particularly by Articles 13 and 15, concerns how the system operates, which variables influence its outputs, how accurate it is, where uncertainty exists and whether performance differs across groups. Operational explanation, reflected in Articles 14 and 26, concerns how the system is integrated into the deployer's processes, what weight its output carries, who reviews it and when it may be overridden. Legal justification builds on Articles 27, 85 and 86 but extends beyond the AI Act's minimum requirements by examining the legal basis and objective of the deployment, the rights affected, less restrictive alternatives, the allocation of responsibility and the availability of complaints, explanations and other remedies [5], [10].

To apply these layers, the deployer should first define the intended use and consider whether the proposed system is an appropriate means of achieving it. Each relevant XAI finding should then be documented together with its scope, evidential limitations, potential rights implications, further information requirements, responsible actors and consequences for deployment. This proposed workflow makes the connection between explanation and institutional action explicit.

A published credit-risk study provides a concrete starting point. Bone-Winkel and Reichenbach used SHAP to interpret models developed using lending data from Bondora. Their analysis identified the combined country-and-language feature as the most influential demographic feature and reported different contributions for Estonian-speaking and Russian-speaking Estonian borrowers [17, Section 4.2, Figure 6]. This is a finding about the models examined in that study; it does not establish unlawful discrimination or describe the platform's actual lending decisions.

For the present illustration, suppose that an EU lender proposes to use a high-risk system exhibiting this explanatory pattern to assess natural persons' creditworthiness, within Annex III, point 5(b), and Article 27. Assume that an adverse assessment normally leads to refusal unless a loan officer intervenes. The

deployment procedure and the FRIA response below are hypothetical.

At the technical layer, the finding identifies a feature requiring scrutiny. Because country and language are combined, the explanation does not by itself isolate their separate contributions. Moreover, a pattern across explanations cannot establish the reason for a particular refusal. The deployer should obtain documentation explaining the feature's construction, the scope of the explanations and the evidence supporting their reliability in the intended context [15]. The published study supplies a relevant example of explanatory information, rather than validation of the proposed deployment.

At the operational layer, the assessment should examine how the adverse score becomes a lending decision. Relevant questions include whether the loan officer understands the explanation, can correct inaccurate information, has authority to depart from the recommendation and records reasons for doing so. The proposed procedure should also enable applicants to obtain understandable reasons and challenge the decision. Human involvement should therefore be assessed through the reviewer's actual powers and working conditions.

At the legal layer, the explanation raises questions about potentially unequal access to credit. The assessment should identify the applicable non-discrimination rules, examine any connection with protected characteristics and consider any legally available justification and less discriminatory alternatives [7]. It should also investigate whether language-related disadvantage interacts with other circumstances, such as age or economic vulnerability [11]. These are questions for further assessment, rather than conclusions that follow automatically from feature importance.

Assume, finally, that the proposed deployer lacks evidence about how the identified pattern affects lending outcomes for the relevant groups. Under the framework proposed here, this material gap should lead to a recorded decision to defer deployment pending further assessment. Table 1 illustrates how the explanation is translated into a FRIA record.

Table 1. Illustrative FRIA record: the author's proposed application of the published finding in [17].

FRIA element	**Illustrative record**
Affected persons and potential harm — Article 27(1)(c)–(d)	Applicants belonging to the language groups represented by the feature; possible unjustified disadvantage in access to credit.
Evidential limitation	The explanation concerns model behaviour. Evidence connecting the pattern to lending outcomes and its legal significance is still required.
Oversight and complaints — Article 27(1)(e)–(f)	Specify the reviewer's authority, procedures for correcting information, recorded reasons and a route for contesting adverse decisions.
Action and responsibility	The provider supplies technical documentation; the lender's assessment team examines deployment effects and legal implications. The lender retains responsibility for the deployment decision.
Proposed decision	Defer deployment until the material evidence gap has been addressed and the resulting safeguards and residual risks have been assessed.

The three-layer framework complements existing assessment approaches. HUDERIA already connects technical characteristics with deployment context, stakeholder engagement and human-rights impacts [13]. The contribution proposed here is a focused procedure for following an XAI finding through evidential scrutiny, operational use and a reasoned FRIA decision. Relevant DPIA findings can inform this procedure, subject to the requirements discussed in Section 3. The framework therefore provides an analytical structure that can be used within these broader assessment processes.

Affected persons, representative organisations and relevant experts should inform the assessment where appropriate. The proposed documentation and deployment decision process constitutes a good-practice model extending beyond the minimum wording of Article 27. The assessment must subsequently be updated when the elements identified in Article 27(1) change or are no longer current.

6 Conclusion

XAI is an important instrument for transparency, auditing and human oversight, but it cannot independently guarantee compliance with fundamental rights. Technical explanations describe the operation or outputs of an AI system; they do not determine whether its use is lawful, necessary, proportionate or non-discriminatory.

A meaningful FRIA should therefore integrate technical evidence with contextual legal analysis, intersectional identification of affected persons, stakeholder participation and effective remedies. Explainability asks how an AI system produced an output. Accountability asks whether the system should have been used, whether reliance on its output was justified, who is responsible for the decision and what happens when the process causes harm.

The worked example illustrates a procedure for translating published XAI findings into documented FRIA questions and institutional responsibilities. Its contribution is analytical and procedural; whether this approach improves assessments in practice remains a question for empirical evaluation.

References

[1] European Parliament and Council, "Regulation (EU) 2024/1689 laying down harmonized rules on artificial intelligence (Artificial Intelligence Act)," *Official Journal of the European Union*, L 2024/1689, 12 July 2024.

[2] C. Panigutti et al., "The Role of Explainable AI in the Context of the AI Act," in *Proceedings of the 2023 ACM Conference on Fairness, Accountability, and Transparency (FAccT '23)*, 2023, pp. 1139–1150.

[3] F. Walke, L. Bennek and T. J. Winkler, "Artificial Intelligence Explainability Requirements of the AI Act and Metrics for Measuring Compliance," in *Solutions and Technologies for Responsible Digitalization*, D. Beverungen, C. Lehrer and M. Trier, Eds., vol. 75. Cham: Springer, 2025, pp. 113–129.

[4] T. Miller, "Explanation in Artificial Intelligence: Insights from the Social Sciences," *Artificial Intelligence*, vol. 267, pp. 1–38, 2019.

[5] E. H. van Beem, "Explainable throughout the AI Lifecycle: The Evolution of the Explainability Provisions in the EU AI Act," *International Review of Law, Computers & Technology*, pp. 1–35, 2026.

[6] C. Rudin, "Stop Explaining Black Box Machine Learning Models for High Stakes Decisions and Use Interpretable Models Instead," *Nature Machine Intelligence*, vol. 1, pp. 206–215, 2019.

[7] S. Wachter, B. Mittelstadt and C. Russell, "Why Fairness Cannot Be Automated: Bridging the Gap between EU Non-discrimination Law and AI," *Computer Law & Security Review*, vol. 41, Art. 105567, 2021.

[8] A. D. Selbst, D. Boyd, S. A. Friedler, S. Venkatasubramanian and J. Vertesi, "Fairness and Abstraction in Sociotechnical Systems," in *Proceedings of the Conference on Fairness, Accountability, and Transparency (FAT '19)*, 2019, pp. 59–68.

[9] European Parliament and Council, "Regulation (EU) 2026/1744 amending Regulation (EU) 2024/1689 as regards the simplification of the implementation of harmonised rules on artificial intelligence," *Official Journal of the European Union*, L 2026/1744, 24 July 2026.

[10] A. Mantelero, "The Fundamental Rights Impact Assessment (FRIA) in the AI Act: Roots, Legal Obligations and Key Elements for a Model Template," *Computer Law & Security Review*, vol. 54, Art. 106020, 2024.

[11] M. Lasek-Markey and L. Hogan, "Delivering on the Promise of the Fundamental Rights Impact Assessments in the EU AI Act: Intersectionality and Vulnerability," *European Journal of Law and Technology*, vol. 16, no. 2, pp. 1–23, 2025.

[12] European Union, "Charter of Fundamental Rights of the European Union," *Official Journal of the European Union*, C 202, pp. 389–405, 7 June 2016.

[13] Council of Europe, Committee on Artificial Intelligence, "Methodology for the Risk and Impact Assessment of Artificial Intelligence Systems from the Point of View of Human Rights, Democracy and the Rule of Law (HUDERIA Methodology)," CAI(2024)16rev2, Strasbourg, 28 November 2024.

[14] S. M. Lundberg and S.-I. Lee, "A Unified Approach to Interpreting Model Predictions," *Advances in Neural Information Processing Systems*, vol. 30, 2017.

[15] K. Sokol and P. Flach, "Explainability Fact Sheets: A Framework for Systematic Assessment of Explainable Approaches," in *Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency*, 2020, pp. 56–67.

[16] Article 29 Data Protection Working Party, "Guidelines on Data Protection Impact Assessment (DPIA) and determining whether processing is 'likely to result in a high risk' for the purposes of Regulation 2016/679," WP 248 rev.01, revised and adopted 4 October 2017.

[17] G. F. Bone-Winkel and F. Reichenbach, "Improving credit risk assessment in P2P lending with explainable machine learning survival analysis," Digital Finance, vol. 6, pp. 501–542, 2024.