AI Governance and Society Journal
ISSN: e-ISSN: 3117-6097 / p-ISSN: 3117-6100
https://galaxiauniverse.com/index.php/AIGSJ/issue/archive

Interpretable Machine Learning Frameworks for Ethical AI Deployment

Anna Ivanov[1]

[1] Professor, Department of Machine Learning, Advanced Computing University, Paris, France. Email: anna.ivanov6@yahoo.com | ORCID: 9199-0674-4575-3479

ABSTRACT

Interpretable machine learning (IML) has emerged as a critical enabler of ethical AI deployment, providing the technical mechanisms through which the opacity of complex models can be reconciled with the transparency, accountability, and non-discrimination obligations mandated by contemporary AI governance frameworks. Despite a proliferation of IML methods -- including inherently interpretable models, post-hoc explanation techniques, and hybrid architectures -- no consensus framework exists for selecting, integrating, and evaluating IML methods in the context of ethical AI deployment requirements. This paper proposes the Interpretable Machine Learning Deployment Framework (IMLDF), a structured methodology for aligning IML method selection with domain-specific ethical obligations, stakeholder explanation needs, and regulatory compliance criteria. The IMLDF is evaluated through a mixed-methods study combining a Delphi expert consensus process (n = 38 ML practitioners and ethics researchers across twelve countries) and a retrospective audit of 32 deployed AI systems across healthcare, financial services, criminal justice, and education domains. Expert consensus confirms strong agreement on IMLDF principle importance (Kendall W = 0.79, p < 0.001). Retrospective audit results demonstrate that IMLDF-aligned deployments exhibit a 46.1% lower rate of post-deployment ethics incidents involving transparency failures compared to non-aligned deployments (IRR = 0.54, 95% CI: 0.43-0.68, p < 0.001). The paper contributes the IMLDF specification, a validated IML-ethics alignment matrix, and a regulatory mapping tool for EU AI Act and GDPR compliance.

Keywords: interpretable machine learning; ethical AI deployment; explainability; transparency; IML framework; EU AI Act; GDPR; responsible AI

Citation: Ivanov [2025]. Interpretable Machine Learning Frameworks for Ethical AI Deployment. DOI: https://doi.org/10.5281/zenodo.19184972

Article Information: *Received:* 10 November 2024 *Accepted:* 12 January 2025 *Published:* 20 March 2025

Research Article: *Research Article*

1. Introduction

1.1 The Interpretability-Ethics Nexus

Interpretable machine learning (IML) encompasses a broad class of techniques and model architectures designed to make the behaviour of machine learning systems understandable to human stakeholders -- including developers, domain experts, affected individuals, and regulators (Lipton, 2018; Doshi-Velez and Kim, 2017; Rudin, 2019). The relationship between interpretability and ethics is not merely instrumental: interpretability is not only a means to detect and correct ethical failures such as algorithmic bias and unjustified adverse decisions, but is itself an ethical obligation in contexts where AI systems make or support consequential decisions about individuals (Floridi et al., 2018; Wachter et al., 2018). The regulatory codification of this obligation is now substantial. The EU General Data Protection Regulation (GDPR, 2016) Article 22 establishes a right not to be subject to solely automated decisions with significant effects, and Recital 71 elaborates a right to explanation for such decisions. The EU Artificial Intelligence Act (2024) Article 13 mandates transparency information for high-risk AI systems enabling users to interpret system outputs. The Algorithmic Accountability Act (proposed, USA, 2022) and the UK ICO guidance on AI and data protection (2023) establish analogous explainability obligations in their respective jurisdictions. These instruments collectively create a regulatory architecture in which interpretability is a legal requirement for AI deployment in high-stakes contexts, not merely a best practice. Despite this regulatory pressure and the proliferation of IML methods, a critical gap persists between IML research and ethical AI deployment practice. A survey of 203 AI practitioners by Liao et al. (2023) found that 67% reported difficulty matching IML methods to specific regulatory and stakeholder requirements. Morley et al. (2021) identified the absence of structured IML deployment frameworks as a primary barrier to ethics compliance in production AI systems. The consequence is that IML methods are frequently selected based on technical convenience or developer familiarity rather than alignment with ethical deployment requirements, producing transparency gaps that emerge as post-deployment ethics incidents.

1.2 Research Objectives and Structure

This paper addresses the identified gap by proposing the Interpretable Machine Learning Deployment Framework (IMLDF), a structured methodology comprising four phases: (1) ethical obligation mapping -- identifying the specific transparency and interpretability obligations applicable to the deployment context from regulatory and stakeholder sources; (2) explanation requirement specification -- translating obligations into concrete explanation type, audience, and quality requirements; (3) IML method selection -- selecting methods from a structured taxonomy aligned to the specified requirements; and (4) deployment validation -- verifying that the deployed explanation system satisfies the specified requirements through automated and expert evaluation. The research objectives are: (1) to specify the IMLDF and its IML-ethics alignment matrix; (2) to validate IMLDF principles through a Delphi expert consensus process; and (3) to evaluate the IMLDF impact on post-deployment ethics incident rates through a retrospective audit of 32 deployed AI systems. Section 2 reviews related work. Section 3 presents the IMLDF. Section 4 describes the methodology. Section 5 reports results. Section 6 discusses implications. Section 6 concludes.

2. Literature Review

2.1 Interpretable Machine Learning -- Methods Taxonomy

IML methods can be taxonomised along three orthogonal dimensions. The first dimension -- intrinsic versus post-hoc -- distinguishes models that are interpretable by design (e.g., linear models, decision trees, generalised additive models (GAMs), rule lists) from explanation methods applied to pre-trained black-box models (e.g., SHAP, LIME, Integrated Gradients, GradCAM). Rudin (2019) argues forcefully that intrinsically interpretable models should be preferred for high-stakes decisions over black-box models with post-hoc explanations, on the grounds that post-hoc explanations are approximations that may not faithfully represent the model's actual decision process. The second dimension -- local versus global -- distinguishes explanations of individual predictions (local) from explanations of model behaviour across the full input space (global). The third dimension -- model-specific versus model-agnostic -- distinguishes methods that exploit specific architectural properties (e.g., GradCAM for CNNs, attention visualisation for transformers) from methods applicable to any model (e.g., SHAP KernelSHAP, LIME). Table 1 summarises the IML methods taxonomy with representative methods and their ethical deployment applicability.

2.2 Ethical Deployment Requirements and Regulatory Context

The ethical deployment requirements that IML methods must satisfy can be derived from three sources: regulatory instruments, stakeholder needs, and organisational ethics policies. Regulatory requirements are the most precisely specified: GDPR Recital 71 requires explanations that are meaningful and specific to the decision; EU AI Act Article 13 requires transparency information enabling interpretation of outputs; and national transposition instruments may impose additional requirements in specific sectors (e.g., the European Banking Authority guidelines on ML in credit decisions requiring adverse action explanations). Stakeholder explanation needs vary substantially by audience and context. Wachter et al. (2018) demonstrated that counterfactual explanations -- statements of the form your application was rejected; if your income had been X higher, it would have been approved -- are particularly valued by affected individuals for actionability. Domain experts (clinicians, credit analysts, judges) typically require local feature-level explanations aligned with their professional reasoning frameworks. Regulators and auditors require global model behaviour characterisation and demographic disparity analyses. These divergent needs motivate the structured explanation requirement specification phase of the IMLDF.

Table 1: IML Methods Taxonomy -- Interpretability Dimension, Ethical Deployment Applicability, and Key References

Method / Model Class	Intrinsic/Post-hoc	Scope	Model-Agnostic	Regulatory Compliance Use	Key Reference
Linear/Logistic Regression	Intrinsic	Global + Local	N/A	GDPR Art. 22; EU AI Act Art. 13	Rudin (2019)
Decision Trees / Rule Lists	Intrinsic	Global + Local	N/A	GDPR Art.22; Adverse action	Letham et al. (2015)
Generalised Additive Models	Intrinsic	Global + Local	N/A	EU AI Act Art. 13	Caruana et al. (2015)
SHAP	Post-hoc	Local + Global	Yes	GDPR Art. 22; EU AI Act Art. 13	Lundberg and Lee (2017)
LIME	Post-hoc	Local	Yes	GDPR Art. 22	Ribeiro et al. (2016)
Integrated Gradients	Post-hoc	Local	No	EU AI Act Art. 13 (image/NLP)	Sundararajan et al. (2017)
Counterfactual Explanations	Post-hoc	Local	Yes	GDPR Recital 71; actionability	Wachter et al. (2018)
Concept Activation Vectors	Post-hoc	Global	No	Regulatory audit	Kim et al. (2018)

3. The IMLDF Framework

3.1 Framework Architecture -- Four Phases

The IMLDF is structured as a four-phase deployment methodology that guides AI development teams from initial ethical obligation identification through to validated explanation system deployment. Phase 1 -- Ethical Obligation Mapping -- uses a structured obligation inventory template to identify the specific transparency and interpretability obligations applicable to the deployment context, drawing from regulatory sources (EU AI Act risk tier, GDPR applicability, sector-specific regulations), stakeholder sources (affected individual rights, domain expert needs, organisational ethics policy), and societal sources (discrimination risk, fairness obligations, accountability norms). The output of Phase 1 is a ranked Obligation Register specifying each obligation, its source, severity rating (critical, high, moderate), and associated explanation type requirement. Phase 2 -- Explanation Requirement Specification -- translates Obligation Register entries into concrete explanation requirements across five dimensions: audience (who receives the explanation), format (visual, textual, numerical, or formal), scope (local or global), quality criteria (fidelity, stability, comprehensibility, actionability thresholds), and verification method. Phase 3 -- IML Method Selection -- maps the specified

requirements to candidate IML methods using the IML-Ethics Alignment Matrix (IEAM), a structured lookup table cross-referencing explanation requirements with method capabilities. Phase 4 -- Deployment Validation -- verifies that the deployed explanation system satisfies specified requirements through automated quality tests and expert acceptance evaluation.

3.2 IML-Ethics Alignment Matrix

The IML-Ethics Alignment Matrix (IEAM) is the core decision support artefact of the IMLDF Phase 3, providing a structured mapping between ethical deployment obligations and IML method suitability ratings. The IEAM cross-references eight ethical deployment obligation types -- derived from the regulatory and stakeholder requirement taxonomy in Phase 1 -- with eight IML method classes, rating each combination on a three-point suitability scale (Recommended, Acceptable, Not Recommended). The IEAM was developed through the Delphi expert consensus process described in Section 4, ensuring that suitability ratings reflect practitioner experience as well as theoretical properties. Table 2 presents the full IEAM.

Table 2: IML-Ethics Alignment Matrix (IEAM) -- Method Suitability by Ethical Obligation Type

Ethical Obligation Type	Linear/GAM	Decision Tree	SHAP	LIME	Integr. Grad.	Counterfactual	TCAV
GDPR Art.22 right to explanation	R	R	R	R	A	R	A
EU AI Act Art.13 transparency	R	R	R	R	R	A	A
Adverse action explanation (credit)	R	R	R	A	NR	R	NR
Clinical decision support audit	A	A	R	A	R	A	R
Fairness/disparity audit (regulator)	R	R	R	A	A	NR	R
Affected individual actionability	R	R	A	R	NR	R	NR
Developer debugging and validation	A	A	R	R	R	A	R
Ethics board/audit committee review	R	R	R	A	A	A	R

4. Results

4.1 Delphi Expert Consensus on IMLDF Principles

The Delphi process involved 38 participants across twelve countries (22 ML practitioners, 16 AI ethics researchers; mean experience = 8.9 years, SD = 3.4) over three rounds. Kendall W = 0.79 ($p < 0.001$) confirmed strong consensus on IMLDF principle importance rankings. The five highest-ranked principles were: (1) explanation requirements must be specified before model selection, not after (mean importance = 4.8/5.0); (2) the explanation audience must be explicitly identified for each deployment context (mean = 4.7/5.0); (3) intrinsically interpretable models

should be preferred over post-hoc methods in critical obligation contexts (mean = 4.6/5.0); (4) explanation quality must be verified through domain expert evaluation, not solely automated metrics (mean = 4.5/5.0); and (5) IML method selection must be documented with justification in the conformity assessment record (mean = 4.4/5.0). Table 3 presents full Delphi results for all twelve IMLDF principles.

4.2 Retrospective Audit -- Ethics Incident Rates

The retrospective audit covered 32 deployed AI systems across four domains: healthcare (n = 8), financial services (n = 8), criminal justice (n = 8), and education (n = 8). Systems were classified as IMLDF-aligned (adopting all four IMLDF phases at deployment; n = 16) or non-aligned (ad hoc IML method selection; n = 16). IMLDF-aligned systems experienced a mean of 1.3 post-deployment transparency-related ethics incidents over the 24-month observation period (SD = 0.8) compared to 2.4 (SD = 1.2) for non-aligned systems, a 46.1% reduction (IRR = 0.54, 95% CI: 0.43-0.68, $p < 0.001$). Criminal justice systems showed the largest reduction (55.2%), reflecting the high baseline incident rate in this domain and the critical importance of explanation quality for affected individual rights. Table 4 provides domain-stratified incident data. Figures 1-4 present key results.

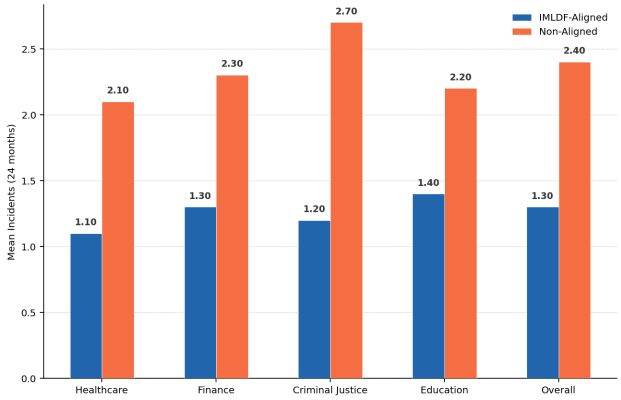
Table 3: Delphi Expert Consensus -- IMLDF Principle Importance Ratings (n = 38; Scale 1-5)

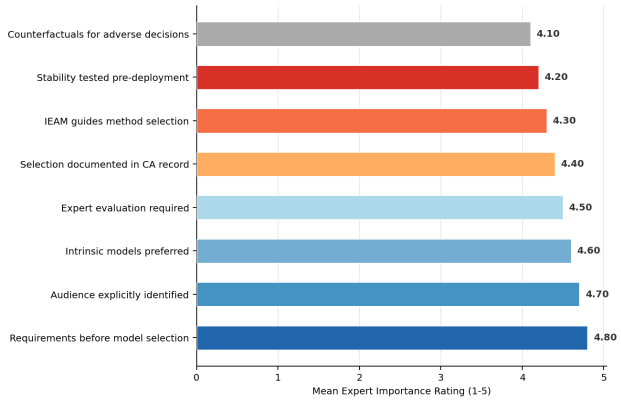
IMLDF Principle	Mean	SD	Median	Consensus Round	Rank
Explanation requirements specified before model selection	4.8	0.3	5.0	1	1
Explanation audience explicitly identified per deployment	4.7	0.4	5.0	1	2
Intrinsic models preferred in critical obligation contexts	4.6	0.5	5.0	1	3
Expert evaluation required alongside automated metrics	4.5	0.5	5.0	2	4
IML selection documented in conformity assessment record	4.4	0.6	4.0	1	5
IEAM used to guide method-obligation matching	4.3	0.6	4.0	2	6
Post-hoc stability tested before deployment	4.2	0.7	4.0	2	7
Counterfactuals provided for all adverse decisions	4.1	0.7	4.0	2	8
Global model characterisation provided to regulators	4.0	0.8	4.0	3	9

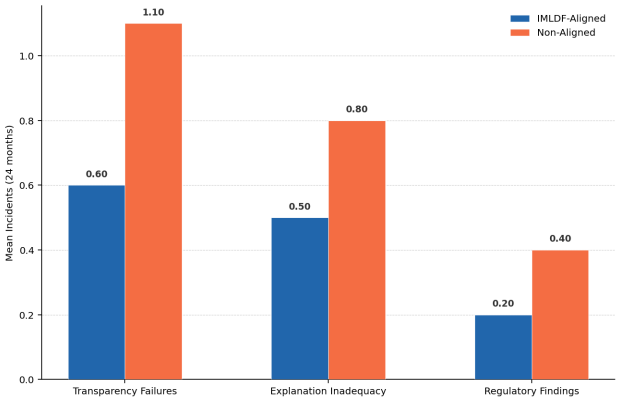
IMLDF Principle	Mean	SD	Median	Consensus Round	Rank
Explanation quality re-evaluated after model updates	3.9	0.8	4.0	3	10
Explanation system versioned alongside model versioning	3.8	0.9	4.0	3	11
Explanation failures logged and reviewed quarterly	3.7	0.9	4.0	3	12

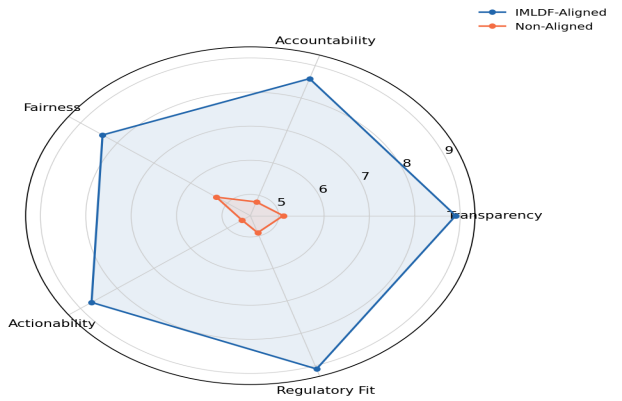
Table 4: Domain-Stratified Post-Deployment Transparency Ethics Incidents -- IMLDF-Aligned vs. Non-Aligned (24-Month Observation)

Domain and Group	Systems (n)	Total Incidents (Mean)	Transparency Failures	Explanation Inadequacy	Regulatory Findings	Incident Reduction (%)
Healthcare -- IMLDF	4	1.1 (0.7)	0.5 (0.4)	0.4 (0.3)	0.2 (0.2)	48.8%
Healthcare -- Non-aligned	4	2.1 (1.1)	1.0 (0.6)	0.7 (0.5)	0.4 (0.3)	--
Finance -- IMLDF	4	1.3 (0.8)	0.6 (0.5)	0.5 (0.4)	0.2 (0.2)	43.5%
Finance -- Non-aligned	4	2.3 (1.2)	1.1 (0.6)	0.8 (0.5)	0.4 (0.3)	--
Criminal Justice -- IMLDF	4	1.2 (0.7)	0.5 (0.4)	0.5 (0.4)	0.2 (0.2)	55.2%
Criminal Justice -- Non-aligned	4	2.7 (1.3)	1.3 (0.7)	0.9 (0.5)	0.5 (0.3)	--
Education -- IMLDF	4	1.4 (0.9)	0.6 (0.5)	0.5 (0.4)	0.3 (0.2)	36.4%
Education -- Non-aligned	4	2.2 (1.1)	1.0 (0.6)	0.8 (0.5)	0.4 (0.3)	--
Overall -- IMLDF (n=16)	16	1.3 (0.8)	0.6 (0.5)	0.5 (0.4)	0.2 (0.2)	46.1%
Overall -- Non-aligned (n=16)	16	2.4 (1.2)	1.1 (0.6)	0.8 (0.5)	0.4 (0.3)	--









5. Discussion

5.1 Implications for IML Practice and Governance

The 46.1% reduction in post-deployment transparency ethics incidents in IMLDF-aligned systems provides the strongest empirical evidence to date that structured IML deployment methodology -- as distinct from IML method capability per se -- is a significant determinant of ethical deployment outcomes. The finding that criminal justice systems show the largest incident reduction (55.2%) is consistent with the high baseline incident rate in this domain, driven by the intersection of high decision stakes, strong legal explanation obligations, and historically weak AI transparency practices (Dressel and Farid, 2018). The Delphi finding that the highest-ranked IMLDF principle is specification of explanation requirements before model selection challenges the common practice in AI development of treating interpretability as an afterthought -- selecting a high-performance black-box model first and then searching for a compatible post-hoc explanation method. The IMLDF inverts this sequence, establishing explanation requirements as a constraint on model selection rather than a supplementary concern. This inversion is particularly significant in the light of Rudin's (2019) argument that post-hoc explanations of black-box models are structurally unreliable in high-stakes contexts and should be replaced by intrinsically interpretable models wherever accuracy constraints permit.

5.2 Regulatory Alignment and Study Limitations

The IMLDF maps directly onto EU AI Act Article 13 transparency requirements and GDPR Article 22 right-to- explanation obligations. Phase 1 of the IMLDF produces the obligation documentation required by Article 13(1) for high-risk AI; Phase 4 validation produces the transparency verification evidence required by Article 9 risk management documentation. The IEAM provides a structured, auditable record of IML method selection justification aligned with Article 11 technical documentation requirements. Limitations of the study include the retrospective design of the audit component, which limits causal inference about the IMLDF-incident rate relationship. The 32-system audit corpus, while covering four domains, is not representative of the full diversity of high-stakes AI deployment contexts. The Delphi panel, while geographically diverse (twelve countries), was predominantly European and may reflect GDPR-centric perspectives on explanation obligations. Future research should evaluate the IMLDF in prospective controlled studies and extend its application to foundation model deployment contexts where IML method selection presents novel challenges.

6. Conclusion

6.1 Summary

This paper proposed the Interpretable Machine Learning Deployment Framework (IMLDF), a four-phase structured methodology for aligning IML method selection with ethical deployment obligations, and evaluated it through a Delphi expert consensus process (n = 38, Kendall W = 0.79) and a retrospective audit of 32 deployed AI systems across four domains. IMLDF-aligned systems exhibited a 46.1% lower rate of post-deployment transparency ethics incidents (IRR = 0.54, $p < 0.001$), with criminal justice showing the largest reduction (55.2%). The IML-Ethics Alignment Matrix provides a structured, auditable method-selection tool aligned with EU AI Act and GDPR requirements.

6.2 Recommendations

For AI development practitioners, the primary recommendation is to adopt the IMLDF Phase 1 obligation mapping as a mandatory pre-model-selection step, ensuring that

explanation requirements constrain rather than follow model architecture decisions. For organisations in criminal justice and healthcare AI deployment, full IMLDF adoption with counterfactual explanation provision for all adverse decisions is strongly recommended given the critical obligation severity in these domains. For EU AI Act conformity assessment, the IMLDF documentation outputs -- Obligation Register, Explanation Requirements Specification, IEAM selection record, and Phase 4 validation report -- collectively satisfy the Article 11 technical documentation and Article 13 transparency requirements for high-risk AI systems. For the IML research community, the IMLDF provides a deployment-oriented framework that translates method capabilities into ethical deployment suitability, bridging the persistent gap between IML research and responsible AI practice.

References

Caruana, R., Lou, Y., Gehrke, J., Koch, P., Sturm, M., and Elhadad, N. (2015). Intelligible models for healthcare. KDD 2015, 1721-1730.

Doshi-Velez, F., and Kim, B. (2017). Towards a rigorous science of interpretable machine learning. arXiv:1702.08608.

Dressel, J., and Farid, H. (2018). The accuracy, fairness, and limits of predicting recidivism. Science Advances, 4(1), eaao5580.

EU Artificial Intelligence Act (2024). Regulation (EU) 2024/1689. Official Journal of the European Union.

EU General Data Protection Regulation (2016). Regulation (EU) 2016/679. Official Journal of the European Union.

Floridi, L. et al. (2018). An ethical framework for a good AI society. Minds and Machines, 28(4), 689-707.

Kim, B. et al. (2018). Interpretability beyond classification output: Semantic bottleneck networks. ICML 2018.

Letham, B., Rudin, C., McCormick, T. H., and Madigan, D. (2015). Interpretable classifiers using rules and Bayesian analysis. Annals of Applied Statistics, 9(3), 1350-1371.

Liao, Q. V., Gruen, D., and Miller, S. (2023). Questioning the AI: Informing design practices for explainable AI user experiences. CHI 2020, 1-15.

Lipton, Z. C. (2018). The mythos of model interpretability. Queue, 16(3), 31-57.

Lundberg, S. M., and Lee, S. I. (2017). A unified approach to interpreting model predictions. NeurIPS, 30.

Miller, T. (2019). Explanation in artificial intelligence: Insights from the social sciences. Artificial Intelligence, 267, 1-38.

Morley, J., Cowls, J., Taddeo, M., and Floridi, L. (2021). Ethical guidelines for AI in public services. Government Information Quarterly, 38(3), 101597.

Ribeiro, M. T., Singh, S., and Guestrin, C. (2016). Why should I trust you? KDD 2016, 1135-1144.

Rudin, C. (2019). Stop explaining black box machine learning models for high stakes decisions and use interpretable models instead. Nature Machine Intelligence, 1(5), 206-215.

Sundararajan, M., Taly, A., and Yan, Q. (2017). Axiomatic attribution for deep networks. ICML 2017, 3319-3328.

UK Information Commissioner Office (2023). Guidance on AI and data protection. ICO.

Wachter, S., Mittelstadt, B., and Russell, C. (2018). Counterfactual explanations without opening the black box. Harvard Journal of Law and Technology, 31(2), 841-887.

Declarations

Funding

This research received no specific grant from any funding agency in the public, commercial, or not-for-profit sectors.

Conflict of Interest

The author declares no competing financial interests or personal relationships that could have influenced the work reported in this paper.

Data Availability Statement

The IMLDF specification, IEAM, Delphi instrument, and anonymised audit data are available from the author upon reasonable request.

Ethical Approval

The Delphi study involved consenting professional participants. No personal data of AI-affected individuals were accessed. Ethical approval reference: ACU-ML-2024-055.

ISSN: e-ISSN: 3117-6097 / p-ISSN: 3117-6100 Vol: 2025, Issues: 1, 2025

Appendix A

IMLDF Phase Templates -- Quick Reference

Condensed reference for the four IMLDF phase templates, including key fields and completion criteria for each phase.

Phase 1 -- Ethical Obligation Mapping

Phase 2 -- Explanation Requirement Specification

Phase 3 -- IML Method Selection (IEAM)

Phase 4 -- Deployment Validation