Operationalizing Explainable Artificial Intelligence in the European Union Regulatory Ecosystem

Luca Nannini, *Minsait – Indra Sistemas, 28108, Madrid, Spain*

Jose Maria Alonso-Moral, Alejandro Catalá, Manuel Lama, and Senén Barro, *University of Santiago de Compostela, 15782, Santiago de Compostela, Spain*

The European Union's (EU's) regulatory ecosystem presents challenges with balancing legal and sociotechnical drivers for explainable artificial intelligence (XAI) systems. Core tensions emerge on dimensions of oversight, user needs, and litigation. This article maps provisions on algorithmic transparency and explainability across major EU data, AI, and platform policies using qualitative analysis. We characterize the involved stakeholders and organizational implementation targets. Constraints become visible between useful transparency for accountability and confidentiality protections. Through an AI hiring system example, we explore the complications with operationalizing explainability. Customization is required to satisfy explainability desires within confidentiality and proportionality bounds. The findings advise technologists on prudent XAI technique selection given multidimensional tensions. The outcomes recommend that policy makers balance worthy transparency goals with cohesive legislation, enabling equitable dispute resolution.

The member states of the European Union (EU) aim to advance civil rights and national interests in developing and deploying information and communication technologies (ICT). The regulatory ecosystem set in place by the European Commission (EC) includes interconnected policies related to foundational drivers like data economy, and key enablers such as artificial intelligence (AI) and platform services, as evidenced by definitive planning documents like the EU Rolling Plan 2022[a] and 2030 Digital Compass.[b] Attention directed at the dignity, freedom, and justice of EU citizens[c] is increasingly being translated into explainability: the ability to understand digital systems, especially those with sophisticated AI architectures.

Conventionally, explainable AI (XAI) techniques enhance the transparency of systems by focusing on making specific AI outputs comprehensible.[1] Yet, being able to deliver explanations regarding this understanding to several users, with different intentions and expertise, is an open-ended challenge. This is particularly relevant to convey enhanced transparency of sophisticated AI systems, encompassing their decision-making process, input data, and other internal variables as well as their design rationale.

In practice, it is often hard to determine which kind of information is required. An explanation could be reproved as not useful, or even jeopardize a third-party's privacy and intellectual property (IP). Thus, XAI applications shall favor AI innovation through major comprehension under the EU rule of law.

To the best of our knowledge, academic literature lacks a perspective that examines AI explainability requirements across multiple EU regulations on data, AI, and platforms. This article maps explainability requirements from EU regulations to operationalization dimensions and involved stakeholders. The former refers to the various functional targets related to implementing

[a]https://joinup.ec.europa.eu/collection/rolling-plan-ict-standardisation/rolling-plan-2022
[b]https://eur-lex.europa.eu/legal-content/en/TXT/?uri=CELEX3A52021DC0118
[c]Encoded in the Charter of fundamental rights (EU2012/c 326/02) https://eur-lex.europa.eu/legal-content/EN/TXT/PDF/?uri=CELEX:C2012/326/02

Digital Object Identifier 10.1109/MIS.2024.3383155
Date of publication 3 April 2024; date of current version 30 July 2024.

explainability in real-world systems.[2] The latter refers to entities like developers, users, and auditors across the AI system lifecycle.[1,3]

As a key remark, we note the lack of a single denotation of explainability for AI systems—which is intended to serve as a proxy to either ensure regulatory oversight or serve AI systems' end users, and also a burden of proof under legal dispute—within the EU ecosystem. This ambiguity should not be interpreted as a lack of consensus or legislative expertise on AI systems. Instead, it reflects the need to balance EU citizen rights, business secrecy, and innovation across diverse contexts from system compliance to user services and litigation. Yet, we observe how the EU regulatory approach may limit end users' legal recourse through AI system explanations.

The article is organized as follows. The "Background" section provides a background on operationalization of XAI techniques and legal considerations. The "Explainability in the EU Regulations" section provides a policy content analysis of EU regulations on algorithmic explainability for data, AI, and platforms. The "Mapping" section maps explainability desiderata in regulations informed by previous work that defines dimensions and stakeholders. The "Discussion" section highlights contexts of use and constraints around the mapped dimensions. The "Conclusion" section concludes with final considerations.

BACKGROUND

Explainability does not exist in a vacuum. Cognitive, behavioral, and sociotechnical factors constrain explainer and recipient heuristics.[4] Early XAI research focused on developer-centric solutions, often tied to mentions of AI principles like *fairness* and *trust*.

Initial research on XAI techniques[1] focused on algorithmic interpretability through local approximations and feature-relevance mappings. However, it largely overlooked factors like cognitive limitations, biases, and real-world contextual deployment.[3,5] Failure to incorporate contextual factors could jeopardize the effectiveness of AI explanations. Qualitative research in XAI has underscored the need for safeguards to confirm the robustness of explanations.[3,6] Despite that, robustness still does not guarantee that explanations are useful or beneficial for recipients.[7,8] The context of use for an explanation is often diverse, influencing user acceptance, understanding, and task performance.[5,9]

Thus, explanations should not be solely perceived as mechanisms that confirm an AI system's abilities or promise added value for users. Even well-articulated, factual explanations might be overshadowed by the system's own accuracy and informativeness if not effectively designed.[10] Furthermore, user acceptance of an explanation does not automatically equate to improved comprehension or functional efficiency. Efforts to bolster transparency via explanations may inadvertently confuse or reinforce undue beliefs or reliance on the AI system, as defined by the concepts of information overload paired to automation and confirmation bias.[5,8]

Thus, safeguards like rigorous testing can help balance transparency to build appropriate trust and avoid pitfalls. Explainability should be integrated into business organizations in ways that account for cognitive and social biases that shape human judgment (heuristic layers), while also balancing transparency demands for accountability and oversight with confidentiality of user data and proprietary models (transparency trade-offs). As for AI ethics principles, their implementation shall be conceived as a top-down procedure across an organization to quantify the potential barriers and risks that mitigate their efficacy.[2]

In this vein, growing attention has been redirected toward AI policy regulation to balance the rule of law and AI-driven business innovation.[11] Such attention denoted how ensuring compliance over explanation generation is not just a technical endeavor, but rather a legal and political one.[12]

Given the heterogeneity and timing of the EU agenda on digital transformation fostering ICT services such as data, AI, and online platforms, single or partial approaches are now widening the lens from the presumption of a "*right to an explanation*" from the General Data Protection Regulation (GDPR)[d] to other regulatory drafts.[12,13,14] Yet, no contributions have mapped algorithmic explainability regulations with the current EU legal ecosystem.

EXPLAINABILITY IN THE EU REGULATIONS

To identify relevant legal explainability requirements, we conducted a structured search using the *EurLex* database of EU legislation. We focused on binding policies related to data, AI, and digital platforms that shape emerging explainability mandates.

The initial candidate texts were screened to determine inclusion based on containing provisions on algorithmic transparency or explainability. From the final selected regulations, we extracted and qualitatively

[d]For updated legal document versions, we report in footnotes links to EurLex official repository: https://eur-lex.europa.eu/eli/reg/2016/679/oj

analyzed articles and recitals with explicit mandates regarding AI system interpretability, transparency, or explainability.

Through an iterative coding process among multiple researchers, we categorized these legal provisions into broader explainability themes, dimensions, and involved stakeholders. This allowed systematic identification and synthesis of legal explainability requirements across the EU regulatory ecosystem into a summarized mapping. The interested reader can consult the methodology and codebook, which is available in the supplemental material[15] at http://doi.org/10.1109/MIS.2024.3383155.

Data Regulations

GDPR

A discussion on the explainability dimension of algorithmic decision-making (ADM) systems and their legal obligations had already been triggered before the GDPR's enactment in 2016.[16] The debate concerned the legitimacy of a "*right to an explanation*" based on the interpretation of Recital 71 and Article 22(1) on automated individual decision making, including profiling. Alongside that, Article 13(2)(f), Article 14(2)(g), and Article 15(1)(h) regard the information to be delivered to a data subject[e] whenever personal data are collected, have not been obtained, or to confirm data processing procedures.

The articles mention the right of the data controller to "*meaningful information about the logic involved, as well as the significance and the envisaged consequences of such processing for the data subject*" in the automated decision. A direction mentioned in Article 22, where modalities are defined to appeal to this decision and contest it, empowers the data subject to not be exposed "*to a decision based solely on automated processing, including profiling, which produces legal effects*."

Researchers have explored ambiguities around the requirement to provide information on "*the logic involved*" in ADM system outcomes that profile and affect individuals. The discussion has centered on whether this requirement pertains to data processing rationale or fuller transparency, including design choices. As expounded on by legal scholars,[12,13] the term "*right*" to an explanation finds a single mention, within a recital, thus holding no binding power. What makes it problematic in its implementation, however, is also the casuistry to which it refers.

The context in which an explanation may be required is based on cases where an ADM system operates decisions with legal effects (or the like) on an end user without *any* human intervention ("*solely*"). This delineation likely invalidates every possible case in which a human operator contributed to the decision about the user, whether or not AI systems were used as decision-making support services.

Regarding the limitations that affect the efficacy of Article 22, the principle of information proportionality adopted by the EU is relevant. In Article 15(4), there is a mention of how an adversary context (i.e., litigation) jeopardizes this right to obtain information even affecting "*the rights of freedom of others*." Furthermore, norms on business secrecy[f] offer a disincentive to disclose information potentially connected to the interests of the controller.

In other words, information obtained by an end user presumably regards only personal data, excluding ADM systems' rationale and the related business design choices conceived. Also, note the lack of legislative precedents for the GDPR as ruled by the Court of Justice of the EU.[12,13] Constitutional courts can establish whether an algorithmic system determined a violation in a specific individual case, yet this will not constitute a general legally binding standard.

Data Act and Data Governance Act

Although the GDPR lays out harmonized rules regarding foremost the elaboration of personal data, the Data Governance Act (DGA)[g] enables the provision of common interoperable data spaces in strategic ICT sectors as designed by the EU Digital Compass. Close to its final draft approval after being proposed in 2020, the DGA was paired in 2022 with a further bill advanced by the EC, defined as the Data Act (DA).[h]

The latter outlines data transfer processes and addresses potential information abuses that could arise from contractual imbalances, potentially obstructing fair access to shared databases. Although the DGA and the DA primarily ensure legislative norms for data

[e]To note, 'data subjects' and 'data controllers' are common terms in EU data protection law - the former referring to users whose personal data is processed, the latter to entities like developers responsible for data processing.

[f]References to Art. 16 of the EU Charter of Fundamental Rights; the Directive (EU) 2016/943 of the European Parliament and of the Council of 8 June 2016 on the protection of undisclosed know-how and business information (trade secrets) against their unlawful acquisition, use, and disclosure: https://eur-lex.europa.eu/legal-content/EN/TXT/?uri=CELEX3A32016L0943; and the European Parliament Resolution of 20 October 2020 on IP rights for the development of artificial intelligence technologies (2020/2015(ini): https://www.europarl.europa.eu/doceo/document/TA-9-2020-0277 EN.html

[g]http://data.europa.eu/eli/reg/2022/868/oj

[h]https://eur-lex.europa.eu/legal-content/EN/TXT/?uri=COM:2022:68:FIN

sharing, their direct relevance to ADM systems is less pronounced, and they could be considered peripheral to our analysis. Nevertheless, this normative framework provides insights into what constitutes transparency of public data, once again emphasizing the principle of informational proportionality as a cornerstone, as seen in the DGA's Article 5 on data reuse conditions.

AI Regulations

Artificial Intelligence Act

The major regulatory instance on AI was proposed with the initial Artificial Intelligence Act (AIA) draft in April 2021. This subsequently went through extensive debate and revisions, with the final compromise text officially approved in December 2023[i] following trilogue negotiations among EU institutions. The approved Act is informed by the principle of proportionality of regulatory intervention for its horizontal approach to risk, i.e., being applied broadly across industrial sectors rather than being domain specific. Aside from prohibited ones, this categorization spans three levels of risk—low, medium, and high—with different application limits, transparency requirements, and oversight mechanisms.

AI stakeholders are introduced alongside definitions of risk assessment (Article 9), data quality requirements (Article 10), and extensive technical documentation for authorities (Article 11, Annex IV); further, for high-risk systems, key provisions requiring measures that facilitate user interpretability and human oversight (Article 14).

For interpretability, Article 13 requires instructions that explain system capabilities, limitations, and accuracy for high-risk systems, while requiring the attachment of instructions for use and "*concise, complete, correct, and clear*" information to users with respect to "*characteristics, capabilities, and limitations of performance*" [Article 13(3)(b)(ii)].

Since the first draft in 2021, scholars have inquired what entails the development of "*sufficiently*" transparent systems that allow end-user interpretation of the outputs and enable their appropriate use.[12,13,14] To wit, the criterion of "*appropriate*" [Article 13(1),(2)] with regard to the type of transparency hinted at an understanding directed for legal sufficiency rather than for a generic end user. The major issue was balancing model complexity, end-user expertise, and business and legal constraints. Indeed, Article 13 does not provide either standards or procedures for evaluating such balances,[14] allegedly leaving the transparency assessment to the discretion of the provider of the high-risk system.[13]

As for the GDPR, objections to fully interpretable design criteria have found appeal in the EU directive that establishes trade secret integrity and IP rights. Although transparency is desirable to respect EU user rights, it is problematic to unilaterally maximize an AI system's transparency when the same users are not controlled over possible misuses.[17,18] In this regard, end-user liability can be ascertained from the need, expressed in Article 13 and Recital(47), for instructions from the provider to the end user on the intended use. In addition, Article 13(1) disregarded the possibility of using interpretability as a burden of proof by third parties affected by the AI system.

After the final draft delivered by the Permanent Representative Committee in November 2022, the European Parliament voted in mid-June 2023 to approve its negotiating position,[j] including notable amendments for guaranteeing algorithmic transparency and individual explainability rights. Among the changes kept in the approved text in February 2024, Article 52[k] requires disclosing AI use in human interactions. Even if some amendments of June 2023 were not later retained,[l] most significantly, Article 68(c) remained. It creates an individual right for those affected by high-risk AI decisions to receive clear explanations about the system's role, main parameters, and inputs with an explicit mention of the GDPR. However, exemptions to Article 68(c) remain possible to protect confidentiality.

Artificial Intelligence Liability Directive

In September 2022, the proposed directive on the adoption of a system of civil liability for the Artificial

[i]We analyze the text version later released by the EU Council in February 2024: https://data.consilium.europa.eu/doc/document/ST-5662-2024-INIT/en/pdf, https://europa.eu/!hjMd9Q

[j]https://www.europarl.europa.eu/doceo/document/TA-9-2023-0236 EN.pdf

[k]General purpose AI models (GPAI) are defined as systems capable of competently performing a wide range of tasks. Article 52 spans both high-risk as well as GPAI systems interacting with humans or generating synthetic audio/video/text content. For GPAI – classified in Art.52a – requirements mandate transparency documentation disclosing acceptable uses, data/training processes, model capabilities/limitations, and content used in Art.52c. But specifics differ and remain less extensive presently compared to comprehensive rules for high-risk systems [15].

[l]We refer to: (i.) the removed Amend.300 in Art.13 to strengthen transparency – to "reasonably understand" the system's functioning and data used to providers and users, not just interpret outputs – moved now to Art.14(4)(c) and Recital (9b) and (47); (ii.) transparency extensions in Art.29 covered explaining AI logic, capabilities and limitations (Amend.308, 411, 767), with a new recital highlighting transparency and explainability to counter digital asymmetry and "dark patterns"(Amend.84).

Intelligence Liability Directive (AILD)[m] was issued to better define those areas of legislative uncertainty during litigation for AI. The proposed AILD is directly relevant to real-world explainability requirements as it aims to address the liability gaps that could affect obtaining explanations or recourse.

It makes it easier to bring claims by lowering evidentiary requirements and allowing representative actions. A core provision introduces a rebuttable presumption of causation if three conditions are met: 1) the defendant failed to meet a relevant duty of care, such as AIA obligations; 2) their fault likely influenced the AI output or failure to output; and 3) the output or lack of output caused the damage.

This presumption facilitates establishing liability in AI-related cases—only within fault-based scenarios—by connecting the defendant's actions to the AI system's role in the damage.

Article 3 allows potential claimants and courts to request the disclosure of relevant documentation and evidence about specific high-risk AI systems from providers or users. This can help claimants access necessary information to determine whether an AI system caused damage. Recitals 16–21 further explain how this allows claimants to identify liable persons and supports the plausibility of contemplated claims.

The court is empowered to access AI system documentation and design, as stated in Article 4, to validate what is referred to at the time of the inspection as a *presumption of causality* between the damage produced and the system's design. Interestingly, Article 4(2)(b),(c) refers to Article 13 of the AIA if providers do not meet transparency and oversight requirements for AI system design and development.

As Recitals 22–30 describe, proving causation can be difficult in AI-enabled damage scenarios given complexity and opacity. Recital 28 specifies how explanations relate to proving causation for opaque AI systems. The duty of care of an AI provider focuses on demonstrating to a mandated supervisory body that the system was compliant with only its instructions for use and documentation. This likely invalidates the possibility for end users to interpret system outputs or receive explanations as a burden of proof under litigation.[19]

Platform Services Regulations

Digital Service Act

The Digital Service Act (DSA)[n] defines intermediary due diligence obligations and conditions for liability exemptions related to digital online services, including platforms for online shopping, content sharing, cloud and messaging services, and marketplaces. The DSA distinguishes among intermediary services, hosting services like online platforms and very large online platforms (VLOPs).

To enhance transparency in content moderation, the DSA requires all intermediary services to designate a legal representative and describe their methods, including the use of ADM systems. Providers must also offer notice mechanisms for allegedly illegal information (Articles 9 and 15) and clearly explain terms and conditions for managing third-party content (Article 14).

For advertising, Recital 68 enhances transparency for users in platform services through "*meaningful explanations*" around the ad and such profiling referring to the GDPR. Article 26 requires online platforms to provide transparency into how users are advertised to; VLOPs using recommender systems (Article 27) are subject to audits on activities like profiling and targeting recipients, having to explain "*design, the logic, the functioning and the testing of their algorithmic systems*" [Article 40(3)].

Similarly, Recital 141 defines the Commission to request documentation and explanations about algorithmic systems from all providers. Article 69(2)(d) and (5) grants the Commission authority during inspections to examine algorithms and require platforms to explain system functionality, data practices, and business conducts. Additionally, Article 72(1) enables the Commission to monitor compliance of VLOPs and search engines by ordering access to databases and algorithms, and requiring related explanations via appointed experts [Article 72(2)].

These provisions focus on oversight and compliance rather than mandating interpretability, yet provide authorities with the tools to request explanations about VLOPs' algorithmic and data systems, enabling investigation of AI systems even if it does not prescribe specific explainability standards.

Digital Markets Act

Although the DSA is focused on regulating online platforms, the Digital Markets Act (DMA)[o] aims to regulate the access of companies to digital markets and services. Specifically, it seeks to prevent companies from abusing their position in the market by imposing unfair conditions on other companies in terms of gatekeeper access. In addition, the DMA puts requirements on companies to share data with competitors, which could lead to increased competition. In regard to

[m]https://eur-lex.europa.eu/legal-content/EN/TXT/?uri=CELEX:52022PC0496

[n]http://data.europa.eu/eli/reg/2022/2065/oj

[o]https://eur-lex.europa.eu/eli/reg/2022/1925/oj

TABLE 1. Mapping of articles referring explicitly to interpretability and explainability dimensions in the EU regulations for AI and ADM systems. For the AIA, only high-risk AI systems are considered. The relevant recitals related to an article's provisions are reported. The explainability dimension(s) covered in the article's provision are signaled with an "[x]" if explicitly mentioned, and an "[∼]" to design potential desiderata. Stakeholder(s) are marked with [G]: giver of an explanation; [R]: recipient of an explanation; and [I]: interpreter. GDPR: GDPR (2016/679); AIA: AIA Draft (2021/0106); AILD: AILD (2022/0303); DSA: DSA (2022/2065); DMA: DMA (2022/1925).

		Explainability dimension(s)			Stakeholder(s)		
	Recital	**Data**	**Process**	**Business**	**Auditor**	**Provider**	**User**
GDPR	—	—	—	—	—	—	—
Article 13(2)(f), 14(2)(g), 15(1)(h)	[60]	[∼]	[∼]	[x]	—	[G]	[R]
Article 22(3)	[71]	[∼]	[x]	—	—	[G]	[R]
AIA	—	—	—	—	—	—	—
Article 13(1),(3)	[14a], [47]	[x]	[x]	—	—	[I]	—
Article 14(4)(c)	[48]	[x]	[∼]	—	—	[I]	—
Article 52(1)	[48], [70a]	—	[∼]	[x]	—	[G]	[R]
Article 68(c)	[84b]	[x]	[x]	[x]	—	[G]	[R]
AILD	—	—	—	—	—	—	—
Article 3(1)	[16]–[21]	—	[∼]	[x]	[R]	[G]	—
Article 4(4),(5)	[22]–[30]	—	[x]	[∼]	—	[G]	[R]
DSA	—	—	—	—	—	—	—
Article 27(1),(2)	[68]	[∼]	—	[x]	—	[G]	[R]
Article 40(3)	[141]	—	[x]	[x]	[R]	[G]	—
Article 69(2)(d),(5)	[146]	[∼]	[x]	[x]	[R]	[G]	—
Article 72(1)	[93], [141]	[x]	[x]	[∼]	[R]	[G]	—
DMA	—	—	—	—	—	—	—
Article 21(1),(2)	[81]	[x]	[x]	[∼]	[R]	[G]	—
Article 23(2)(d),(4)	[83]	[∼]	[∼]	[x]	[R]	[G]	—

explainability, Article 21(1),(2) and Article 23 empower the EC to request information and conduct inspections to access data, algorithms, testing information, and business practices from gatekeepers. This enables transparency and oversight into the technical processes behind gatekeepers' AI systems as well as their organizational operations.

MAPPING

To present a comprehensive taxonomy of the different functions that correspond to an explanation, we report specific articles and their related recitals in Table 1. The table categorizes regulations by the explainability dimensions and stakeholders they address, providing a useful overview of existing regulatory requirements related to the interpretability and explainability of AI systems arising from the policy content analysis.

The table is divided into two main sections: "Explainability Dimensions" and "Stakeholders." The former refers to the different aspects of AI and ADM systems that require explanation, while the latter refers to the entities involved in the AI lifecycle who either provide or receive them.

For dimensions, we consult previous mapping work on the operationalization of ethical principles in the AI lifecycle by Georgieva et al.[2] Their classification is divided into explainability targets such as *technical* for traceability and system description, *process* for decision-making process criteria, and *business* for organizational decision-making processes and system design criteria.

We build on these distinctions to account for a contextual understanding during the due diligence phase or litigation associated with the stakeholders involved in the AI lifecycle. Explainability requirements are also considered in the sociotechnical context of data and platforms use where an AI system can be implemented.

Under *data*, we intend both the temporal side of accessibility and fruition in the ex-ante coordinates (i.e., access to databases) and after processing of the AI system (i.e., data output), but also their topology (i.e., user profiling or a reference-sampled group). Under *process*, we refer to the architecture and capabilities of an AI system. Under *business*, we refer to the business choices that designed the AI system, the acknowledgment of the user interacting with it, and also the social effects of an AI system that shapes an organization's business policy and affects third parties.

As previously illustrated in the "Explainability in the EU Regulations" section, articles and recitals may ambiguously define interpretability or explainability targets. For example, the "*logic involved*" concept in Article 22 of the GDPR varies case by case, allegedly including only the processing of personal data to system design and the related business choices.

As a precaution, we use the *tilde* symbol to designate legislative ambiguity on *what* (target) can be explained. In addition to the table, we define *how* (type) something can be explained. We draw on the typology advanced by Cabitza et al.[20] for their degree of sufficiency in covering the major casuistry for AI explanations.

Their work differentiates explainability types as *computational* (i.e., how the algorithm produced any output), *mechanistic* (i.e., why so), *justificatory* (i.e., why the output is deemed right), *causal* (i.e., which factors and how the output was caused), *informative* (i.e., implications of the output), and *cautionary* (i.e., the uncertainty behind an output). In line with these considerations, we view types of explainability desiderata as neither mutually exclusive nor rigidly defined.

However, we do identify certain interpretative tendencies. Mechanistic and computational types potentially encompass the enhancement of technical interpretability via XAI methodologies to describe an AI system's data output and process. Justificatory and causal explanations are desirable when a business explanation needs to be provided and supported by technical explanations, establishing a heuristic ground truth of the system being explained. These explanations could also serve as evidence subject to cross examination under litigation. Informative and cautionary explanations, on the other hand, add layers of semantic granularity, i.e., they advance epistemological knowledge about the system analyzed. Their teleological nature addresses future uncertainty and is more closely related to business explainability, informing users or organizations about remedies to achieve compliance or improve user services.

For stakeholders, we advance three macro categories emerging from the analysis of regulations. By *user*, we mean service clients, data subjects, or claimants; by *provider*, we refer to both AI providers and general deployers of ADM systems, also in platform services; with *auditor*, we intend oversight bodies, including legal persona delegated to conduct audits.

DISCUSSION

Building on the mapping of explainability dimensions and provisions in the "Explainability in the EU Regulations" and "Mapping" sections, this discussion explores the competing tensions around explainability situated across various contexts of use, from oversight procedures to user services and litigation dispute, as shown in Figure 1. Balancing transparency for accountability with confidentiality concerns emerges as a central challenge. Regulators seek explainability to audit systems, while providers shield proprietary details. Users impacted by AI decisions desire recourse through understanding, but legal and technical ambiguities persist around useful explainability. We propose a functional view of explanation types[2,20] based on these scenarios.

To illustrate core explainability issues independent of a specific regulatory context, we present an example of an AI hiring system in an online platform, e.g., for job advertising and screening, to be deemed high risk under the AIA. Through this example, we explore cross-sector tensions between explainability desires and transparency barriers. Rather than focus narrowly on hiring, we aim to highlight key considerations for trustworthy AI broadly. By anchoring analysis in this scenario, we strengthen the principles of ethical AI with the practical challenges organizations encounter.

Explainability for Controller Oversight

AI explainability can be leveraged as a risk management tool in oversight procedures. Auditors, related to the EC or national legal bodies, request explainability from AI/ADM service providers or deployers to ensure legislative compliance.

Requests for explanations can occur either as a standard oversight procedure (e.g., for job platforms that advertise in the DSA) or as an exceptional case due to litigation (e.g., for an alleged presumption of

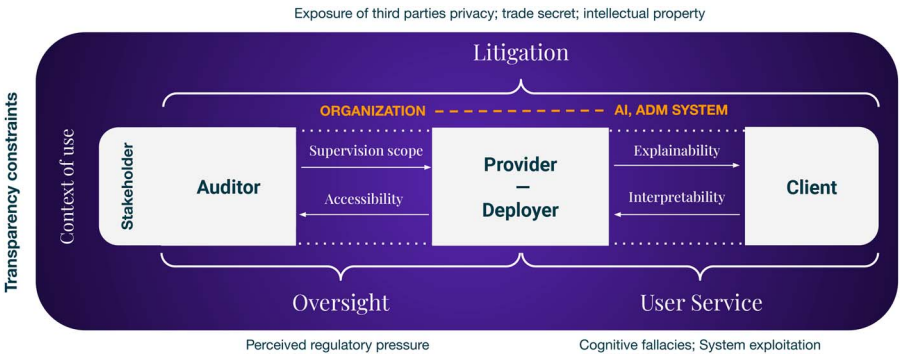

FIGURE 1. Contexts of use highlighted by discrepancies within different stakeholders involved in the generation and reception of explanations. On an outside layer, transparency constraints are found relative to different contexts of use. An additional axis is proposed to locate the explainability area from AI or ADM systems, respectively, to the business organization.

causality on the harms produced by faults in an AI system over a claimant in the AILD).

Explainability, during an oversight procedure, is addressed transversely in an organization. Even if explanations are generated automatically by an XAI technique, liability spans across an organization's managerial-empowered figures (e.g., data protection officers or C-suite) or technical ones (e.g., data scientists and engineers). Legal dispositions generally interpret explainability as granting access to data, algorithms, and business models. Maintained confidentially within supervision bounds, transparency constraints during disclosure should be minimal given the binding power of regulations.

During oversight, auditors are expected to specify the scope of supervision, outlining the process, objectives, and involved systems. They can request explanations be given to providers for specific remedies over preliminary noncompliance, e.g., whether an AI or ADM system in a provider's platform (DSA—Articles 69 and 72) might bias results in recommending job announcements to certain demographics of applicants. The companies that use these high-risk AI systems must provide clear and meaningful explanations of their decision-making processes, including the AI system's role, main decision parameters, and related input data [AIA—Article 68(c)].

Despite the principle of proportionality, the sufficiency level needed to satisfy an audit scope seems to be at the auditor's discretion, leaving room for interpretation regarding the target and type of explanation.

Transparency Tradeoffs

The regulations also allow for exceptions or restrictions to full transparency under certain conditions.

According to Article 68(c)(2) of the AIA, the obligation to provide explanations does not apply to the use of AI systems for which exceptions or restrictions are provided under EU or national law, as long as they respect fundamental rights and are deemed necessary and proportionate.

At the organizational level, exposing additional information might endanger business secrecy as well as third parties. Aside from users maliciously exploiting the system, public exposure of its features might disincentivize market competitiveness of AI-empowered services offered by a provider. This could discourage providers' business drivers due to enhanced EU regulatory pressure. Yet, regulatory prototyping could reduce compliance costs, informing the debate on standardization practices.[11]

Additionally, measures to shield business secrecy and IP are defined in the DMA to protect against unfair digital market competition. In this vein, as an example, Article 3(4) of the AILD mentions the Trade Secret Directive (EU) 2016/943 as a baseline for evaluating proportionality and confidentiality of information being disclosed.

Aside from directly affected organizations, increasing transparency may endanger the privacy of third parties if differential measures are not set in place. This might result in a GDPR violation to implement appropriate technical and organizational measures to ensure a level of security appropriate to the risk (Article 32). This type of breach primarily regards interpretability and explainability on target data, even before algorithms and business.

Viable XAI Solutions

When implementing explainability for regulatory oversight of the AI hiring system, several XAI approaches can help balance compliance needs with other constraints.

Context-aware, participatory design processes are imperative to reconcile transparency for accountability with confidentiality of user data and business practices.

Inspecting globally interpretable models like generalized additive models (GAMs),[1] despite their limited applicability to high-risk systems if constituted of complex deep neural networks, could provide regulators with insights into feature impacts on hiring decisions in a globally interpretable way while protecting raw parameters. Interactive counterfactual scenarios tailored to the hiring context could also demonstrate how changes to applicant profiles alter outcomes—without exposing personal information if properly aggregated—while still considering their reliance on assumptions about feature independence and locality that may not fully hold.

However, inherent tensions remain between openness for avoiding direct and indirect discrimination and protecting legitimate trade secrets. Advanced techniques like multiparty computation through federated learning could enable collective auditing without exposing raw data, but incur substantial coordination costs, especially for smaller organizations.

Overall, the precise XAI techniques used in the hiring system must be selected based on legal guidance under EU nondiscrimination law and extensive consultation across affected stakeholders. The explanations that are sufficient for regulatory compliance differ from those that are useful for applicants or the business itself.

Explainability for User Services

Although explainability serves accountability purposes for regulators, additional considerations arise around useful explainability for the end-user groups affected by AI systems.

In the AIA, Article 68(c) grants individuals affected by high-risk AI decisions the right to request clear explanations about the system's role, main parameters, and related input data. In a similar vein, the GDPR and Recital 68 in the DSA hint to such recourse mechanism.

Additionally, Articles 13 and 52 in the AIA require that high-risk AI systems be designed and developed to ensure sufficient transparency for providers, users, and affected individuals to reasonably comprehend and interpret the system's functioning.

Applied to a hiring context [high-risk scenario for the AIA, (Annex III)], this means that the applicants not selected by an AI screening system could seek understandable explanations about how the system contributed to the decision, which key factors and applicant profile features influenced the outcome, and which personal and training data were utilized.

Together, these provisions aim to make the opaque AI systems used in impactful contexts, like hiring, more transparent and explainable to those affected and obliged to deploy them responsibly. However, tensions around useful explainability, accountable oversight, and legitimate secrecy persist.

The degree of sufficiency for the target and type of explanation remains undefined. For instance, it is unclear what constitutes *relevant* input data [AIA—Article 13 and Article 68(c)] and whether it refers to personal information or general training and fine-tuning data. Similarly, the term *parameters involved* [Article 68(c)] could be interpreted as referring to feature relevance, but this is not explicitly stated. The overall design of the high-risk AI system is also to be explained, but the level of granularity required is not specified.

In terms of XAI methods,[1] local interpretable model-agnostic explanations (LIME) provides intuitive local explanations by constructing linear surrogate models to approximate the behavior of complex machine learning models like deep neural networks. Despite nonlinearity in the full model, LIME allows for interpreting individual predictions locally. Alternative techniques like shapely additive explanations can explain nonlinear models but require disclosure of global model mechanics, which could reveal IP. Interactive counterfactual interfaces enable applicants to tweak inputs and visualize outcomes, empowering them to improve future job applications.

Yet, safeguards like differential privacy and access controls are imperative when exposing counterfactual features as excessive transparency risks of users unfairly exploiting it.[11,17] Multimodal explanations that combine saliency maps, partial dependence plots, and natural language could provide complementary views for end users with different backgrounds. But robustness to perturbation, cognitive load, and user fatigue require careful, human-centered design.[5]

Therefore, considering evaluation metrics and conducting empirical studies to validate explanation quality across user groups is critical.[8,9] Beyond picking convenient XAI techniques, responsible implementation demands extensive testing to ensure that explanations truly empower users without introducing harms.[4,17]

Explainability for Litigation

When alleged harms prompt legal action, explanations face renewed scrutiny in litigation contexts, subject to evidentiary standards.

Under the GDPR and the AIA, an explanation would likely focus on the processing of end users' personal data, without revealing the decision-making process and design choices made by the AI provider. The choice of XAI methods is left to the provider's discretion, and such a provision likely holds an *exemplificative* value [AIA—Article 14(4)(c)]. This discretion does not ensure reference standards or evaluation criteria regarding the suitability or accuracy of the provided method.

Therefore, beyond the context of oversight, end users are likely exempted from any legal recourse within their interpretation or explanations as explanation types are upon providers' discretion.[12,13] This could foster a context of information asymmetry where providers have no real incentive to use interpretable models to safeguard business secrecy, nor are considered liable once generic instructions or explanations are generated.

In a class-action lawsuit alleging a discriminatory effect of the AI hiring system, the company would need to reveal the model's inner workings to an auditor, including the features it considers and their assigned weights. Under the proposed AILD, if claimants provide sufficient evidence of system opacity or complexity, they can appeal for a court-ordered audit process. The court could access the AI system's documentation and design to validate a claimant's allegations. Failure to meet this interpretability requirement (AIA Article 13) could be used against the company in court. Conversely, the company could argue a lack of system compliance with its instructions for use and documentation, shifting liability to the end user.

In this hiring-litigation scenario, automated explanatory systems could enforce presumptive causality, disincentivizing providers' XAI use because it risks becoming a burden of proof against them. Yet, interpretable models like GAMs offer explanatory compliance evidence without full model exposure. The counterfactual scenarios tailored to hiring contexts may also showcase nondiscrimination through localized tweaking inputs, with proper anonymity protections.

Overall, selective XAI techniques are beneficial if compliance can be demonstrated without requiring full model disclosure. Informative explanations could reflect adopted design criteria, while cautionary ones could outline uncertainties around instructions and intended use.[20] Companies shall proactively adopt suitable explainability to verify fairness and mitigate litigation risks, with awareness of the tradeoffs between court-mandated transparency and legitimate secrecy around proprietary models.

CONCLUSION

Explainability principles endorsed through regulations must balance transparency for accountability with confidentiality barriers and recourse obstacles. This analysis highlighted the open challenges that transition ideals into organizational processes.

Central findings have emerged for managing tensions. Theoretically, first by balancing transparency for accountability against confidentiality protections. This translates to an explanation typology that spans technical functions justifying system logic, to informative ones guiding business practices. Practically, organizations must proactively adopt lawful, customized transparency that balances compliance risks and constraints. Standardization efforts should acknowledge multidimensional tensions between principles and organizational processes. Technologists, specifically, must assess explanations to empower users without introducing new barriers or potential harms.

Further empirical research can provide nuanced insights into realizing explainability amid constraints, directing policy and enforcement. Overall, explainability requires a context-aware customization that balances demands and protections. But legal compliance alone does not guarantee ethical deployment absent remedies that enable individuals, especially vulnerable ones, to challenge the algorithmic decisions that affect them.

ACKNOWLEDGMENTS

The authors acknowledge a contribution from the Innovative Training Networks project NL4XAI, which has received funding from the EU's Horizon 2020 research and innovation program under Marie Sklodowska-Curie Grant 860621. This work was also supported by Grant PID2021-123152OB-C21 funded by MCIN/AEI/10.13039/501100011033, and by the European Regional Development Fund's (ERDF's) "A way of making Europe." This work was also supported by the Galician Ministry of Culture, Education, Professional Training and University [Grant ED431G2019/04 and Grant ED431C2022/19, cofunded by the ERDF/Fondo Europeo de Desarrollo Regional (FEDER) program].

REFERENCES

1. A. B. Arrieta et al., "Explainable artificial intelligence (XAI): Concepts, taxonomies, opportunities and challenges toward responsible AI," *Inf. Fusion*, vol. 58, pp. 82–115, Jun. 2020, doi: 10.1016/j.inffus.2019.12.012.
2. I. Georgieva, C. Lazo, T. Timan, and A. F. van Veenstra, "From AI ethics principles to data science practice:

A reflection and a gap analysis based on recent frameworks and practical experience," *AI Ethics*, vol. 2, no. 4, pp. 697–711, Nov. 2022, doi: 10.1007/s43681-021-00127-3.

3. M. Langer et al., "What do we want from Explainable Artificial Intelligence (XAI)? – A stakeholder perspective on XAI and a conceptual model guiding interdisciplinary XAI research," *Artif. Intell.*, vol. 296, Jul. 2021, Art. no. 103473, doi: 10.1016/j.artint.2021.103473.
4. A. Páez, "The pragmatic turn in explainable artificial intelligence (XAI)," *Minds Mach.*, vol. 29, no. 3, pp. 441–459, Sep. 2019, doi: 10.1007/s11023-019-09502-w.
5. A. Bertrand, R. Belloum, J. R. Eagan, and W. Maxwell, "How cognitive biases affect XAI-assisted decision-making: A systematic review," in *Proc. AAAI/ACM Conf. AI, Ethics, Soc.*, New York, NY, USA: ACM, 2022, pp. 78–91, doi: 10.1145/3514094.3534164.
6. D. C. Elton, "Common pitfalls when explaining AI and why mechanistic explanation is a hard problem," in *Proc. 6th Int. Congr. Inf. Commun. Technol.*, X.-S. Yang, S. Sherratt, N. Dey, and A. Joshi, Eds., Singapore: Springer, 2022, vol. 235, pp. 401–408, doi: 10.1007/978-981-16-2377-6_38.
7. T. Miller, "Explanation in artificial intelligence: Insights from the social sciences," *Artif. Intell.*, vol. 267, pp. 1–38, Feb. 2019, doi: 10.1016/j.artint.2018.07.007.
8. G. Bansal et al., "Does the whole exceed its parts? The effect of AI explanations on complementary team performance," in *Proc. CHI Conf. Human Factors Comput. Syst.*, New York, NY, USA: ACM, 2021, pp. 1–16, doi: 10.1145/3411764.3445717.
9. R. R. Hoffman, S. T. Mueller, G. Klein, and J. Litman, "Metrics for explainable AI: Challenges and prospects," 2018, *arXiv:1812.04608*.
10. A. Papenmeier, D. Kern, G. Englebienne, and C. Seifert, "It's complicated: The relationship between user trust, model accuracy and explanations in AI," *ACM Trans. Comput.-Human Interact.*, vol. 29, no. 4, pp. 1–33, Mar. 2022, doi: 10.1145/3495013.
11. M. Fenwick, E. P. M. Vermeulen, and M. Corrales, "Business and regulatory responses to artificial intelligence: Dynamic regulation, innovation ecosystems and the strategic management of disruptive technology," in *Robotics, AI and the Future of Law* (Perspectives in Law, Business and Innovation), M. Corrales, M. Fenwick, and N. Forgó, Eds., Singapore: Springer, 2018, pp. 81–103, doi: 10.1007/978-981-13-2874-9_4.
12. P. Hacker and J.-H. Passoth, "Varieties of AI explanations under the law. From the GDPR to the AIA, and beyond," in *Proc. Int. Workshop Extending Explainable AI Beyond Deep Models Classifiers*, A. Holzinger, R. Goebel, R. Fong, T. Moon, K.-R. Müller, and W. Samek, Eds., Cham, Switzerland: Springer International Publishing, 2022, pp. 343–373, doi: 10.1007/978-3-031-04083-2_17.
13. M. Ebers, "Regulating explainable AI in the European Union. An overview of the current legal framework(s)," in *Nordic Yearbook of Law and Informatics 2020–2021: Law in the Era of Artificial Intelligence*, L. Colonna and S. Greenstein, Eds., Stockholm, Sweden: The Swedish Law and Informatics Research Institute, Aug. 2021, pp. 1–20.
14. F. Sovrano, S. Sapienza, M. Palmirani, and F. Vitali, "Metrics, explainability and the European AI act proposal," *J*, vol. 5, no. 1, pp. 126–138, Feb. 2022, doi: 10.3390/j5010010.
15. L. Nannini, J. M. Alonso-Moral, A. Catala, M. Lama, and S. Barro, 2024, "Supplemental material—Operationalizing explainable AI in the EU regulatory ecosystem (1.0)," Zenodo, doi: 10.5281/zenodo.10792499.
16. B. Goodman and S. Flaxman, "European union regulations on algorithmic decision-making and a 'right to explanation'," *AI Mag.*, vol. 38, no. 3, pp. 50–57, Fall 2017, doi: 10.1609/aimag.v38i3.2741.
17. M. Ananny and K. Crawford, "Seeing without knowing: Limitations of the transparency ideal and its application to algorithmic accountability," *New Media Soc.*, vol. 20, no. 3, pp. 973–989, 2018, doi: 10.1177/1461444816676645.
18. A. Bringas Colmenarejo et al., "Fairness in agreement with European values: An interdisciplinary perspective on AI regulation," in *Proc. AAAI/ACM Conf. AI, Ethics, Soc.*, New York, NY, USA: ACM, 2022, pp. 107–118, doi: 10.1145/3514094.3534158.
19. S. Bordt, M. Finck, E. Raidl, and U. von Luxburg, "Post-hoc explanations fail to achieve their purpose in adversarial contexts," in *Proc. ACM Conf. Fairness, Accountability, Transparency*, New York, NY, USA: ACM, 2022, pp. 891–905, doi: 10.1145/3531146.3533153.
20. F. Cabitza et al., "Quod erat demonstrandum? Towards a typology of the concept of explanation for the design of explainable AI," *Expert Syst. Appl.*, vol. 213, Mar. 2023, Art. no. 118888, doi: 10.1016/j.eswa.2022.118888.

LUCA NANNINI is an industrial Ph.D. student in research in information technologies at Minsait by Indra Sistemas, 28108, Madrid, Spain, and the Research Center in Intelligent Technologies, University of Santiago de Compostela, Santiago, Spain. His research interests include explainable artificial intelligence (AI), process mining, and AI governance. Nannini received his M.A. degree in cognitive semiotics from Aarhus University. Contact him at l.nannini@usc.es.

JOSE MARIA ALONSO-MORAL is an associate professor at the Research Center in Intelligent Technologies, University of Santiago de Compostela, 15782, Santiago de Compostela, Spain. Alonso-Moral received his Ph.D. degree in telecommunication engineering from the Technical University of Madrid. He is a Member of IEEE. Contact him at josemaria.alonso.moral@usc.es.

ALEJANDRO CATALÁ is an assistant professor at the Research Center in Intelligent Technologies, University of Santiago de Compostela, 15782, Santiago de Compostela, Spain. His research interests include intelligent user interfaces and smart multimodal technologies involving conversational, robotic, and tangible interactions. Catalá received his Ph.D. degree in computer science from Polytechnic University of Valencia. Contact him at alejandro.catala@usc.es.

MANUEL LAMA is an associate professor at the Research Center in Intelligent Technologies, University of Santiago de Compostela (USC), 15782, Santiago de Compostela, Spain. His research interests include machine learning, process mining, and business process management. Lama received his Ph.D. degree in physics from USC. Contact him at manuel.lama@usc.es.

SENÉN BARRO is director of the Research Center in Intelligent Technologies, University of Santiago de Compostela, 15782, Santiago de Compostela, Spain, where he is a full professor of computer science. His research interests include machine learning and the social and economic impact of artificial intelligence. Barro received his Ph.D. in physics from USC. Contact him at senen.barro@usc.es.