Habemus a Right to an Explanation: So What? – A Framework on Transparency-Explainability Functionality and Tensions in the EU AI Act

Luca Nannini[1, 2]

[1]Centro Singular de Investigación en Tecnoloxías Intelixentes da USC, Santiago de Compostela, Spain
[2]Minsait by Indra Sistemas SA, Madrid, Spain
l.nannini@usc.es, lnannini@minsait.com

Abstract

The European Union's Artificial Intelligence Act (AI Act), finalized in February 2024, mandates transparency and explainability requirements for AI systems to enable effective oversight and safeguard fundamental rights. Yet the practical implementation of these requirements faces challenges due to tensions between the need for meaningful explanations and their potential risks. This research proposes the Transparency-Explainability Functionality and Tensions (TEFT) framework to analyze the interplay of legal, technical, and socio-ethical factors shaping the realization of algorithmic transparency and explainability in the EU context. Through a two-pronged approach combining a focused literature review and an in-depth examination of the AI Act's provisions, we identify key friction points and challenges in operationalizing the right to explanation. The TEFT framework maps the interests and incentives of various stakeholders, including AI providers & deployers, oversight bodies, and affected individuals, while considering their goals, expected benefits, risks, possible negative impacts, and context to algorithmic explainability.

1 Introduction

Artificial intelligence (AI) systems are being deployed to automate decisions across crucial social domains. Yet their current complexity, especially for Machine Learning (ML) based Deep Learning models (DL), has elicited calls for transparency to enable effective human oversight. Governmental policies are now attempting to tackle such exigence, yet it is unclear to what extent published communications, regulations, and standards adopt an informed perspective over AI explainability to support research, industry, and civil interests (Nannini, Balayn, and Smith 2023).

Particularly for European Union (EU) policy, the Artificial Intelligence Act (AI Act) draft endorsed in February 2024 (European Parliament 2024a) contains provisions requiring AI providers to ensure explanations of system operations and individual outputs, especially for "*high-risk*" applications with significant impacts on citizen rights. Nevertheless, this policy foregrounds the tensions between satisfying explainability obligations for human oversight and providers' concerns around disclosing commercially sensitive intellectual property and competitive intelligence.

Copyright © 2024, Association for the Advancement of Artificial Intelligence (www.aaai.org). All rights reserved.

Prior scholarly work has examined the "*right to explanation*" (henceforth RTE) introduced in the EU General Data Protection Regulation (GDPR) (European Parliament and Council of the European Union 2016) and algorithmic transparency from legal, technical and socio-ethical perspectives among also interdisciplinary analyses. However, a critical gap remains in synthesizing these diverse viewpoints and identified tensions through the lens of the EU AI Act's final provisions on explainability ratified in April 2024 (European Parliament 2024b). Providing a cohesive mapping of the complex interplay of factors catalyzed by these new regulatory requirements demands novel conceptual grounding, also investigating if and how debated tensions persist. To tackle that, we define the following research questions:

RQ1 *How has the RTE scholar discourse evolved from the EU GDPR until the latest version of the AI Act?*

RQ2 *What are the key transparency and explainability requirements in the EU AI Act, and what are the main challenges in implementing them for different stakeholders?*

RQ3 *How can we conceptualize and address the tensions between transparency, explainability, and competing interests to effectively govern AI systems in the EU context?*

Our research aims to synthetize and move forward this discourse by proposing a holistic "*Transparency-Explainability Functionality and Tensions*" (TEFT) framework analyzing the interplay of legal, technical, and socio-ethical factors shaping the practical implementation of the EU AI Act's transparency requirements. Specifically, our contributions are twofold: first providing a comprehensive, multidisciplinary literature review synthesizing legal analyses, technical perspectives, and socio-ethical debates surrounding the RTE underpinning algorithmic transparency, with a focused examination of this concept's evolution from the GDPR to the definitive EU AI Act. From that, introducing a novel framework to consider key tensions and map stakeholder motivations, constraints, and potential equilibria to achieve transparency.

The paper is structured as follow. Section 2 details our methodology, where its first phase informs Section 3 on literature review and the second Section 4 the analysis of the final EU AI Act provisions. From these analyses, Section 5 presents the framework, while Section 6 considers current limitations and Section 7 the future outlook.

2 Methods and Scope

To respond to those RQs and clearly position our work, this research conducts a two-phase analysis[1].

I. Literature Selection and Coding – We began by identifying a relevant set of papers that discuss the legal, technical, and socio-ethical aspects of algorithmic transparency and explainability, with a specific focus over the RTE on the European Union's regulatory landscape, including the GDPR and the EU AI Act. The selection process was guided by the following inclusion criteria: (a) papers published from 2016 onwards, coinciding with the enactment of the GDPR; (b) papers that directly address the legal, technical, or socio-ethical aspects of the RTE or algorithmic transparency/explainability in the EU context; and (c) papers published in peer-reviewed journals or conference proceedings. As an additional quality criteria, pre-screened papers were excluded if resulted focusing on non-EU contexts or mentioning the RTE without engaging with prior scholarly discourse, thus not being substantially novel in the debate. We used Scopus (April 2024), expanding the pool through citation chaining.

To extract and synthesize insights from the selected literature corpus, we developed a coding scheme tailored to capture key perspectives and themes. The coding process involved a comprehensive analysis of each paper, systematically identifying and categorizing relevant information across several dimensions aligned with RQ1. Specifically, we captured: (1.) the paper's disciplinary lens (law, computer science, social science, or interdisciplinary, etc); (2.) a summary of its key arguments regarding the "*right to explanation*" or algorithmic transparency/explainability; (3.) any specific challenges or limitations highlighted in implementing such transparency measures; (4.) proposed frameworks or approaches to address those challenges; (5.) an assessment of the paper's relevance and potential contributions to the TEFT framework; (6.) identified "*friction points*," which we define as areas of contention or ambiguity – i.e., potential conflicts between stakeholder interests, technical feasibility, and legal enforceability – that may hinder the implementation of the AI Act's explainability requirements, informed by seminal literature on the RTE (e.g., Selbst and Powles (2017); Wachter, Mittelstadt, and Floridi (2017)).

II. Synthesis and Framework Development – For the analysis of the EU AI Act (RQ2), we focused on the explainability provisions outlined in the February 2024 finalized amendments, while we consulted the *Corrigendum* text released in April 2024 (European Parliament 2024b). We systematically examined the relevant articles and recitals, paying particular attention to the requirements for high-risk AI systems, the obligations of providers and users, and the rights of affected individuals. Building on the coded literature, we evaluated if the identified themes, arguments, and friction points were persisting in the final version of the Act. This comparison (RQ3) helped us to refine the TEFT framework and its integration of multiple perspectives, including stakeholder interests, technical feasibility, legal enforceability, socio-ethical implications, and contextual factors.

[1]An expanded version of the manuscript - with codebook description and additional material – is available as Nannini (2024).

3 Literature Review

The RTE for automated decisions has emerged as a focal concept in broader discussions surrounding algorithmic accountability, transparency, and ethical governance. Reported in legal frameworks like the European Union's GDPR, this notion aims to empower individuals impacted by consequential algorithmic decisions and foster trustworthy AI that respects core human values. Scholarly discourse spanning disciplines such as law, computer science, philosophy, ethics, and social science has revealed intricate tensions and challenges in interpreting the scope of this right, assessing its legal enforceability, navigating technical feasibility constraints, and aligning it with diverse stakeholder interests and socio-ethical considerations.

3.1 Unpacking the GDPR's "Right to Explanation"

At the heart of this debate lies a fundamental disagreement over whether the GDPR, through provisions like Articles 13-15 and 22, indeed establishes a legally binding and enforceable right for automated decisions significantly impacting individuals. On one side of this interpretive divide, scholars such as Selbst and Powles (2017) and Goodman and Flaxman (2017) contend that the regulation's requirements to provide "*meaningful information about the logic involved*" in such decisions, coupled with mandated "*suitable safeguards*," create an obligatory RTE. Conversely, Wachter, Mittelstadt, and Floridi (2017) and Brkan and Bonnet (2020) argue the GDPR lacks precise, enforceable language establishing such a right. They point to ambiguities surrounding key terms like "*solely automated decisions*" and "*legal or similarly significant effects*," as well as the notable absence of an explicit mention of a "*right*" in the regulation's legally binding articles, appearing only in non-binding recitals. This group of scholars contends the GDPR merely creates a narrower "*right to be informed*" about the general functionality and existence of automated decision-making systems, rather than a right to rationales behind specific decisions.

Attempts to resolve this interpretive rift have spawned divergent proposals for operationalizing the principle of algorithmic explainability under the GDPR. Malgieri and Comandé (2017) advocate for a robust "*right to algorithmic legibility*" combining transparency over system logic and architecture with comprehensible, tailored explanations of individual decisions. This holistic "*legibility*" framework aspires to provide meaningful information enabling individuals to understand and potentially challenge algorithmic decisions or outcomes. Kaminski (2019)'s analysis further elaborates on the concept of "*qualified transparency*," arguing that the GDPR establishes a system of layered transparency, with individuals receiving meaningful information about automated decision-making, while regulators and expert third parties gain deeper access for systemic oversight. In contrast, Wachter, Mittelstadt, and Floridi (2017)'s interpretation suggests the GDPR's provisions point towards a limited approach to transparency, reserving in-depth information (e.g., source code and technical details) for regulators and auditors serving a systemic oversight role.

3.2 Tensions in Realizing the Right to Explanation

Despite the merits of these varied interpretive lenses, scholarly analysis has illuminated significant friction points and obstacles complicating the realization of a legally cognizable and technically feasible RTE departing from the GDPR's provisions[2].

Stakeholder Conflicts and Socio-Ethical Implications A recurring tension identified relates to the conflicting interests of individuals seeking meaningful transparency to understand and contest automated decisions impacting their lives, versus the commercial priorities of companies developing AI/ML systems (Wachter, Mittelstadt, and Floridi 2017; Mylly 2023; Grochowski et al. 2021).

Data controllers have strong incentives to safeguard intellectual property rights, trade secrets, and competitive advantages by limiting disclosure requirements that could compromise valuable algorithms and datasets (Hacker et al. 2020; Busuioc, Curtin, and Almada 2023; de Laat 2022). This underlying conflict must be carefully balanced against individuals' rights and ability to exercise self-advocacy through informed scrutiny of automated decisions, as scholars like Vredenburgh (2021) compellingly argue is a moral imperative for legitimizing non-voluntary governance systems. Indeed Vredenburgh (2021); Munch, Bjerring, and Mainz (2024) further ground the RTE in the fundamental interest of "*informed self-advocacy*," contending that explanations are necessary for individuals to represent their interests, conform to rules, and hold decision-makers accountable in hierarchical, non-voluntary governance systems.

On top of that, the quest for algorithmic explainability intersects with broader socio-ethical considerations around ensuring automated decision-making systems respect principles of fairness and human autonomy. Goodman and Flaxman (2017) point that simply removing sensitive variables like race or gender from algorithmic models is insufficient to prevent discriminatory outcomes, as other non-sensitive variables can serve as proxies perpetuating demographic biases present in historical training data. Addressing such "*uncertainty bias*" (Goodman and Flaxman 2017) favouring well-represented groups underscores the need for robust auditing and remediation mechanisms enabled by algorithmic transparency.

From a philosophical perspective, scholars like Jongepier and Keymolen (2022) have challenged the commonly assumed human-machine dichotomy underlying many conversations about "*rights*" to explanation. They argue for a "*symmetry thesis*" focusing not on the source of decisions (human or algorithmic), but rather on their justification and alignment with core human deliberative agency. Under this view, the same expectations for transparency and reason-giving should apply universally – automated decisions warrant no special normative status over decisions made exclusively by humans. Running contrary to this argument, other scholars like Laux, Wachter, and Mittelstadt (2024) ground the RTE as an essential mechanism for fostering public trust and accountability in automated decision-making systems increasingly governing aspects of human activities. Explainability is thus framed as a *prerequisite* for safeguarding human autonomy and empowerment in the face of powerful, opaque algorithmic systems. Munch, Bjerring, and Mainz (2024) further underscore the importance of considering the stakes associated with algorithmic decisions, arguing that high-stakes decisions warrant a RTE, while low-stakes decisions may not.

Technical Feasibility Constraints Compounding these socio-ethical tensions are well-documented technical hurdles impeding the provision of meaningful, intelligible, and actionable explanations for the logic and rationale underlying many contemporary AI systems, particularly those leveraging advanced ML techniques like deep neural networks.

A core issue, as reported by Sovrano, Vitali, and Palmirani (2021); Brkan and Bonnet (2020); Gryz and Rojszczak (2021) is the inscrutability and opacity of complex ML models to human comprehension - the "*black box*" problem. These systems' decision-making processes can involve high-dimensional representations and non-linear transformations of data, defying straightforward translation into intuitive concepts and logic flows understandable to lay individuals (Kim and Routledge 2022). Even if debated (Rudin 2019), inherent trade-offs could arise between the performance and accuracy of ML models and their interpretability, wherein constraints to enhance the latter can inadvertently undermine the former (Goodman and Flaxman 2017; Sovrano, Vitali, and Palmirani 2021). Selbst and Powles (2017) suggest that infeasible systems simply cannot be legally deployed under the GDPR, reinforcing the idea that technical limitations should not be used as an excuse to undermine the RTE.

While the field of explainable AI (XAI) (Adadi and Berrada 2018) has made strides in developing techniques to improve the transparency of such opaque models, researchers like Sovrano et al. (Sovrano, Vitali, and Palmirani 2020; Sovrano et al. 2022) have noted challenges in aligning available XAI methods with GDPR requirements. Specifically, they argue existing XAI tools struggle to generate the types of explanations legally mandated, such as contrastive, counterfactual rationales and justifications conveying the logic involved step-by-step. A parallel issue is the sheer computational and cognitive complexity of attempting to deliver customized, contextualized explanations for each specific decision made by AI/ML systems operating at scale across different domains (Edwards and Veale 2018; Bertrand

[2]The significance of this broad interpretation of "*decision*" under Article 22(1) of the GDPR was recently underscored by the Court of Justice of the European Union (CJEU) in the SCHUFA case (Court of Justice of the European Union 2023a,b). The court in December 2023 held that the automated establishment of a probability value concerning a person's ability to meet payment commitments constitutes a *decision* where a third party draws strongly on that value to make decisions. This expansive reading reinforces the effective protection intended by Article 22 and helps prevent circumvention and gaps in legal protection. The court also clarified that such automated decision-making is prohibited unless one of the exceptions in Article 22(2) applies and the specific requirements in Article 22(3) and (4) are met.

et al. 2022). Doing so may rapidly become intractable, suggesting the need for carefully curated frameworks that provide universal, ex ante explanations of general system functionality complemented by tailored rationales for decisions meeting defined risk thresholds. This "*layered*" or "*qualified*" approach to transparency has been discussed by scholars like Kaminski (2019); Nannini et al. (2024). Especially for the EU AI Act, Panigutti et al. (2023) identify several key limitations of current XAI methods, including the lack of robustness and stability of explanations, absence of standard evaluation frameworks, and the potential to increase automation bias and over-reliance on AI suggestions.

Legal Enforceability Frictions Even assuming the socio-ethical merits and technical feasibility of a RTE under the GDPR, commentators have identified additional considerable obstacles to ensuring consistent interpretation and robust enforcement of such a right across the EU's diverse national jurisdictions. A prominent issue highlighted by scholars like Malgieri (2019) is the diverging implementations and regulatory approaches taken by different EU Member States in codifying the GDPR's provisions around automated decision-making into national law. There exist discrepancies in the types of exceptions permitted, safeguards mandated, and crucially, the scope of decisions covered – some nations have broadened regulatory oversight beyond solely automated decisions with legal effects to include those with potentially detrimental impacts. This fragmentation generates legal uncertainty and raises concerns over the EU's goal of harmonizing data and algorithmic standards across their member states (Novelli et al. 2024).

Furthermore, inconsistencies in regulatory resourcing and technical capacity across Member States may impede uniform enforcement (Kaminski 2019). Edwards and Veale (2018); Wachter, Mittelstadt, and Floridi (2017) caution that even robust transparency rights may have limited practical efficacy if regulators lack expertise to meaningfully scrutinize complex algorithmic systems. Obstacles around access to justice and asserting individual rights could also hinder legal enforceability (De Gregorio and Demkova 2024). De Gregorio and Demkova (2024) further highlights the challenges posed by the intertwined remedies in the EU's digital regulations[3], emphasizing the need for clarity in the interplay between remedial designs and a focus on institutional collaboration to ensure effective enforcement.

Additional frictions identified by Ebers (2021); de Laat (2022) include the potential for conflicts between transparency mandates and legitimate needs to protect sensitive information around national security or classified systems. Concerns regard balancing GDPR obligations with safeguarding intellectual property and commercial interests, with lack of regulatory guidance creating uncertainties around permissible practices[4]. Custers and Vrabec (2024) further highlight the challenges in applying GDPR data subject rights to inferred data and profiles generated by data controllers, adding another layer of complexity to the legal enforceability of the RTE.

3.3 Proposed Governance Solutions

Recognizing the multidimensional complexities implicated in the RTE debate, scholars have advocated for holistic governance frameworks synthesizing interdisciplinary perspectives from law, computer science, and philosophy.

Systemic Accountability Models Moving beyond narrowly defined individual explanation rights, scholars like Casey, Farhangi, and Vogl (2019); Edwards and Veale (2018); Gryz and Rojszczak (2021) argue for adopting comprehensive "*algorithmic accountability by design*" methodologies. Such approaches mandate cross-cutting processes like algorithmic impact assessments, third-party auditing, and systemic safeguards to be integrated throughout the entire AI system lifecycle from early design to real-world deployment. The proposed benefits of these approaches are twofold: first proactively embeds transparency, testing, and oversight mechanisms into the development process rather than treating explanations as disconnected ex post obligations. Second, by shifting the governance burden from individuals towards empowered oversight bodies and expert auditors, it circumvents challenges around the technical infeasibility of delivering customized explanations at scale.

Furthermore, as Casey, Farhangi, and Vogl (2019) argue, the GDPR's extraterritorial scope may lead to its algorithmic accountability requirements, including its RTE, becoming a global standard for companies deploying AI systems, even beyond the EU's borders. Embracing this paradigm, Malgieri (2019) highlight novel practices emerging across EU Member States to potentially inform standardized EU-wide governance norms. These include France and Hungary's "*algorithmic legibility*" requirements mandating disclosure of key system criteria and methods, the UK's procedural transparency approach based on individual contestation rights, and Slovenia's algorithmic impact assessment obligations considering human rights implications.

Collaborative Governance and Policy Prototyping Recognizing the inevitable contextual variability in *if* and *how* transparency obligations should apply across different domains, scholars like Heymans, Gils, and Ooms (2024); Laux, Wachter, and Mittelstadt (2023) recently discussed institutionalizing participatory governance approaches. These frameworks would facilitate structured collaboration between policymakers, technical experts, impacted communities and interest groups to co-develop regulations, guide-

[3] i.e., GDPR, AI Act, but also the DSA's provisions over algorithmic transparency as illustrated by Helberger and Samuelson (2024); Söderlund et al. (2024).

[4] The recent Uber/Ola case in the Netherlands illustrates the challenges in enforcing data access rights in practice (Amsterdam Court of Appeal 2023a,c,b). The appeals court found that the platforms violated drivers' rights in several instances, including when algorithms were involved in terminating driver accounts. Importantly, the court ruled that the platforms cannot rely on trade secrets exemptions to deny drivers access to their data, although challenges remain for workers to use existing laws to get sufficient visibility into platforms' data processing. The case underscore the value of collective actions and worker organizations like the Worker Info Exchange (2023) (WIE) in supporting platform workers to exercise their rights and hold platforms accountable.

lines, and best practices tailored to the local cultural, economic, and ethical nuances surrounding different use cases for automated decision systems (Bibal et al. 2021; Casey, Farhangi, and Vogl 2019; Olsen et al. 2024). Yet, adoption of these approaches should also hinge on the forthcoming AI standards over algorithmic explainability – cognizant of the evolving nature and heterogeneity of such landscape (Nannini, Balayn, and Smith 2023; CEN-CENELEC 2020).

Rather than attempting to codify rigid, one-size-fits-all rules, policymakers would chart forward by testing innovative transparency and accountability pilots, evaluating shortcomings, and refining through recursive feedback cycles integrating technical, legal, and sociological expertise. For instance, Heymans, Gils, and Ooms (2024)'s policy prototyping experiments demonstrate the merits of stakeholder involvement in testing legal compliance documents to surface ambiguities, identify accessibility frictions, and refine transparency implementations for "*high-risk*" AI systems under the EU AI Act. Laux, Wachter, and Mittelstadt (2023) propose embedding similar participatory mechanisms into the standardization processes of the AI Act, ensuring articulated transparency norms reflect social consensus rather than technocratic imposition. Kim and Routledge (2022) further argue for a trust-based approach to informed consent, proposing a "*right to remedial explanation*" when harms occur and a "*right to updating explanation*" to allow users to reassess over time, complementing the collaborative governance and policy prototyping approaches.

4 EU AI Act and Current Ambiguities

To further complicate the debate over the GDPR's RTE and its continuity to AI policies, it shall be acknowledged that the drafting of the EU's AI Act sensibly changed its final provisions regarding algorithmic explainability and transparency[5]. While indeed the approved Act introduces several provisions to enhance transparency and explainability, particularly for high-risk AI systems, some ambiguities and friction points persist. This section explores the key transparency provisions in the EU AI Act, the remaining areas of contention, and the potential spaces for further resolution.

[5]The draft originally proposed in April 2021 by the European Commission (2021) did not comprehend any explicit article regarding a RTE. The first inclusion of this concept was proposed as *Article 69(c)* through the amendments by the European Parliament Committee on Legal Affairs (2022) (JURI) in their Opinion of September 2022. These amendments were not accepted in the common position of November that year by the Council of the European Union (2022) (COREPER). It was only in the negotiating position advanced at the end of 2023 that the RTE was adopted as *Article 68(c)* by the European Parliament (2023a). When the EU AI Act agreement reached in December 2023 between the European Parliament (2023b) and the Council was endorsed by members state in February 2024 (European Parliament 2024a), Article 68(c) was retained. Although, certain provisions were modified - e.g., Article 13 to direct interpretability requirements towards AI *providers* and not *users* anymore. Subsequently, in the Corrigendum released in April 2024 (European Parliament 2024b), the numbering of the "*right to explanation*" article was changed from 68(c) to *Article 86* and Article 52 to 50.

4.1 Transparency and Explainability Clauses

The final version of the Act mandates certain transparency provisions, notably targeting "*high-risk*" systems as legally required to ensure embedded explainability capacities by design under Title III, Chapter 2 (European Parliament 2024b). We comprehensively examine key articles as follows.

Article 13 – requires understandable AI outputs, capabilities, and instructions enabling oversight. 13(1) mandates sufficient transparency to interpret outputs for appropriate use. This explains recommendations or decisions per Article 86(1). Clause 13(2) requires clear, complete documentation aiding comprehension of characteristics. As per clause 13(3)(b)(ii), documentation must elucidate accuracy metrics and impacts - inaccurate/biased systems need greater elucidation. For relevant systems, clause 13(3)(b)(iiia) also needs transparency on explanatory abilities, while 13(3)(d) requires explanations on oversight measures under Article 14 for deployment monitoring.

Article 14 – also targets effective human oversight over AI systems via interpretability, accountability and control. Clause 14(1) requires high-risk AI systems allow for such oversight facilitating user autonomy. More specifically, clause 14(3)(c) necessitates transparency enabling accurate interpretations of outputs produced while clause 14(4) stipulates understanding capacities, constraints etc. Recital 48 supports this navigate appropriate, contextually sound AI adoption preventing risks.

Article 50 – stresses disclosures for certain AI systems directly interacting with individuals. Clause 50(1) requires clear notifications when conversing with an AI system for sound transparency norms that respect personal dignity. For biometric categorization or emotion recognition applications, clause 50(3) asks for transparent processing explanations respecting privacy rights. Beyond design phases, Article 26(5) places transparency burdens upon AI system deployers too for monitoring best practices adherence including around accuracy metrics communicated (which clause 13(3)(b)(ii) requires providers detail initially).

Article 86 – indeed named "*Right to explanation of individual decision-making*" allows individuals to contest opaque determinations by demanding understandable explanations over "*the role of the AI system in the decision-making procedure and the main elements of the decision taken*" if delegated decisions lack transparency, as recommended by recitals 107 & 171. Yet, as detailed by Figure 1, it is subject to certain limitations:

- It only applies to decisions made using high-risk AI systems listed in Annex III, with several carve-outs under Article 6(3) for systems that do not pose significant risks to health, safety, or fundamental rights.
- The RTE exists only if a decision produces legal effects or similarly significantly affects a person in a way that they consider to adversely impact their health, safety, and fundamental rights.
- Article 86(3) specifies that the RTE applies only to the extent that it is not otherwise provided for under Union law, such as the GDPR's RTE arising from Articles 15(1)(h) and 22(1).

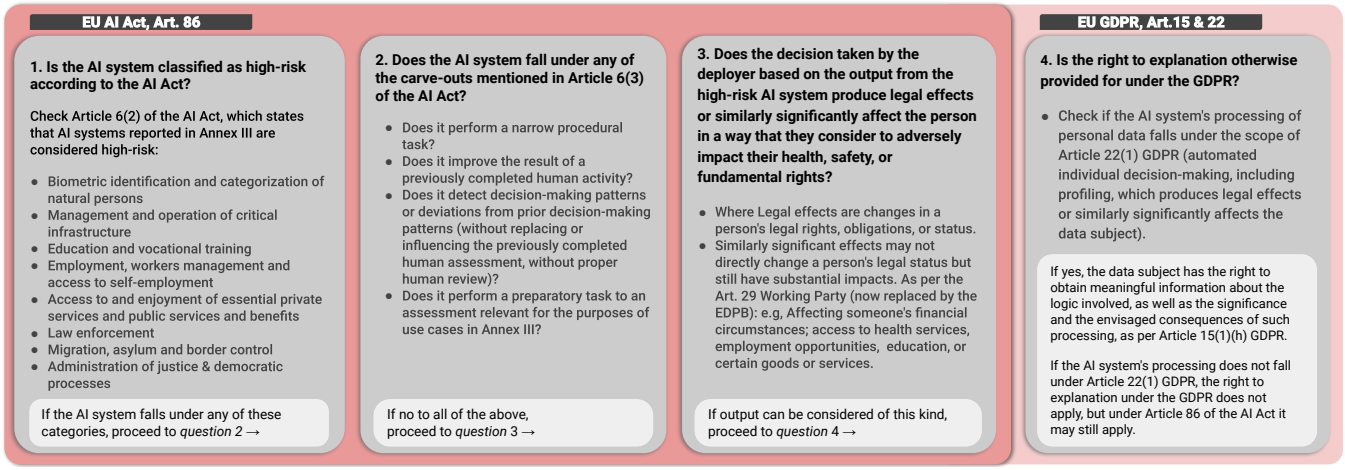

Figure 1: Prerequisites to ensure applicability of Article 86 under the EU AI Act for high-risk systems

In addition, Chapter V of the EU AI Act introduces specific transparency obligations for providers of general purpose AI models. Article 53(1) requires providers of general purpose AI models to maintain technical documentation (Annex XI), provide information and documentation to downstream providers integrating the model (Annex XII), establish policies to ensure compliance with EU copyright law, and provide a summary of the content used for training the model. In this line, Article 55 imposes additional obligations on providers of general purpose AI models with systemic risks, such as conducting model evaluations, assessing and mitigating systemic risks, reporting serious incidents, and ensuring cybersecurity protection.

4.2 Frictions and Residual Ambiguities

Despite the unprecedented regulatory provisions now legally necessitating transparency and embedded explainability capacities in potentially high-risk AI systems, there are still uncertainties regarding standards, roles, enforcement, and participatory legitimacy that could allow for persistent opacity despite nominal transparency requirements (Article 13). To ensure the Act's success, these tensions must be clarified, balancing the need for confidentiality with the demand for comprehension (Article 78). Also informed by the literature review, four main categories of ambiguities are identified:

1. *The specific content components that require technical explanations and the level of detail needed* (Article 13(3)(b)(ii));
2. *How explanatory procedures should be implemented in practice, including measures taken by providers and obligations of deployers* (Article 14);
3. *The lack of participatory diversity in oversight mechanisms* (Recital 171);
4. *The appropriate calibration of transparency requirements based on the risk level of AI systems* (Article 6).

Moreover, there is a lack of clear evaluation metrics and methods for assessing the effectiveness of explanations (i.e., no mentions under Article 72 for post-market monitoring). This ambiguity could lead to providers exploiting loopholes by providing limited transparency while maintaining strategic opacity, prioritizing their competitive advantage over the public interest (Recital 168).

Gaps in Explainability Specifications Article 13 mandates that certain AI systems enable transparency, allowing overseers to understand outputs and contest decisions. Yet the Act does not provide clear technical standards for what constitutes *sufficient* explanations (Recital 58). Beyond output interpretations, there is a lack of specificity regarding the documentation required to explain model construction, workflows, and other processes, including customization needs, security allowances, and transparency tools necessary for legitimate inspection and redress (Article 11) (Malgieri and Pasquale 2024).

A lack of procedural clarity regards the appropriate protocols for assessing the informativeness of explanations and system documentation provided by providers to deployers and consumers (Article 13(3)(b)(ii)). Metrics for evaluating explanations in terms of relevance, completeness, and actionability are missing, thus potentially favoring *token* transparency that preserves strategic opacity. Effective oversight, aided by expertise of regulators as advocated by Edwards and Veale (2018), requires comprehensive explainability specifications covering multiple aspects of system elucidation, component disclosure, and evaluation rubrics, which should contribute to fulfill the EU AI Act's constitutional promise of safeguarding rights (Recital 4). Yet Panigutti et al. (2023) demonstrate through an AI-based proctoring use case how transparency and human oversight requirements can be met via comprehensive documentation, instructions for use, and organizational measures, without needing XAI techniques or transparent-by-design models.

Confidentiality and Participation Trade-offs Regarding the content components that require technical explainability while preserving the confidentiality of privileged intellectual property (IP) (Ebers 2021; Hacker et al. 2020; Mylly 2023), a tension remains between the need for explanations

to ensure due process rights (Article 86(1)) and the extensive confidentiality exceptions that shield providers' commercial interests (Articles 53 and 55). Articles 7 and 11 particularly tolerate opacity to preserve trade secrecy for security or competitive innovation reasons (Article 78). This denotes a lack of robust public participatory consensus on the exact standards for reconciling transparency needs with confidentiality concerns (Recital 171). The risk of excessive IP protections could lead to token explanations that lack full technical transparency on model construction specifics.

Proportionality and Evaluation Uncertainty Similar coherency gaps exist in the procedures for implementing transparency, with current provisions rarely stipulating monitoring and evaluation methods for assessing the appropriateness of explanations and documentation (Article 72). Article 26 discusses the obligations of deployers of high-risk AI systems, including the need to monitor the operation of the AI system and inform providers of any serious incidents or malfunctioning. Yet, it does not provide clear guidance on *how* deployers should fulfill these obligations in relation to transparency and explainability, which may lead to varying interpretations and practices across different sectors and jurisdictions. This risks lax compliance without a structured approach addressing tensions between transparency requirements and oversight amidst constraints (Recital 168).

Likewise, again we denote how currently are deficiencies in participatory diversity within oversight mechanisms (Recital 171). While several articles place transparency obligations on AI providers directly, with partial duties on immediate deployers, the involvement of broader oversight institutions like consumer protection bodies and the inclusion of diverse public voices representing affected individuals are absent. This threatens to reproduce existing socioeconomic biases by limiting explainability interventions to a *provider-deployer-regulator* triad lacking lived social participation (Article 67). This could scale up to AI explainability being weaponized for public-relations (PR) campaigns rather than genuine public redress or participation unless affected groups have sustained channels for consultation (e.g., the Worker Info Exchange (2023) in the Uber/Ola case and their support to platform workers).

Lower-risks, thus lower explainability? Lastly, questions persist around the proportionality of transparency requirements across high, medium, and low-risk AI categories, which are currently lacking in EU regulation (Article 6). While the EU AI Act now legally mandates explainability for "*high-risk*" AI systems that significantly impact citizen rights and choices (Annex III), the precise scope requires clarification regarding its applicability to lower-risk AI systems beyond the identified high-risk categories, where explainability appears essential for oversight based on ethics and fundamental rights principles (Recital 48). For AI systems legally designated as high-risk, where potential harms could outweigh benefits and necessitate external oversight safeguards against corporate or engineering interest risks, Articles 13 and 14 apply entirely, mandating transparency procedures to ensure model processes and individual outputs remain contestable by demanding explanations from providers. This upholds constitutional priorities around rights-based governance (Recital 4).

Nonetheless lower-risk AI systems that likely escape such categorization currently have greater opacity allowances, with voluntary codes of conduct being the sole transparency instrument rather than strict mandates (Article 95). While basic disclosures on interactions with AI systems may persist in certain conversational use cases directly interfacing with individuals, such as chatbots under Article 50 for truth preservation, data dignity, and consent ethics, intensive requirements ensuring interpretable outputs or documentation of engineering processes are waived for non-high-risk AI currently exempt from close inspection. This proportionality approach, which might easing market innovation by limiting transparency burdens for lower-risk systems and small-scale providers, still carries fundamental rights risks for EU citizens, since algorithmic determinations lacking oversight might produce unfairness due to unexamined biases against minorities and other vulnerable communities (Recital 48).

Thus, while the EU AI Act introduces pioneering transparency demands for potentially impactful algorithms managing vital services, proportionality exemptions for numerous lower-risk AI systems imply that explainability remains *legally non-compulsory* based on current regulations that calibrate confidentiality assurances for providers against oversight desires for the public (Article 6).

5 Transparency in Practice: TEFT Framework

The Transparency-Explainability Functionality and Tensions (TEFT) framework detailed in Figure 2 provides a comprehensive lens for understanding the complexities surrounding algorithmic transparency and explainability in the context of the EU AI Act. The framework builds upon the challenges and proposed solutions identified in the literature, such as systemic accountability models (Casey, Farhangi, and Vogl 2019), collaborative governance (Heymans, Gils, and Ooms 2024; Laux, Wachter, and Mittelstadt 2023), and policy prototyping (Heymans, Gils, and Ooms 2024; Bibal et al. 2021), while explicitly examining the interactions between different components and addressing the friction points and ambiguities identified in the latest version of the Act. The TEFT framework is composed as follows: (I.) *Actors and their Goals*, (II.) *Benefits*, (III.) *Costs/Risks*, and (IV.) *Possible Negative Impacts*, complemented by *Context Equilibria* and additional considerations over mapping explanation facets and different phases considerations.

5.1 Components

I. Actors & Goals In line with Nannini et al. (2024); Laux, Wachter, and Mittelstadt (2024), we define three macro classes of actors holding different goals over algorithmic transparency in industrial, oversight, and social domains:

1. **Providers/Deployers**: Technology suppliers and vendors who prioritize commercial competitiveness, favouring opacity to protect their trade secrets and IP (Articles 53 and 55), embodying the concerns raised in the literature

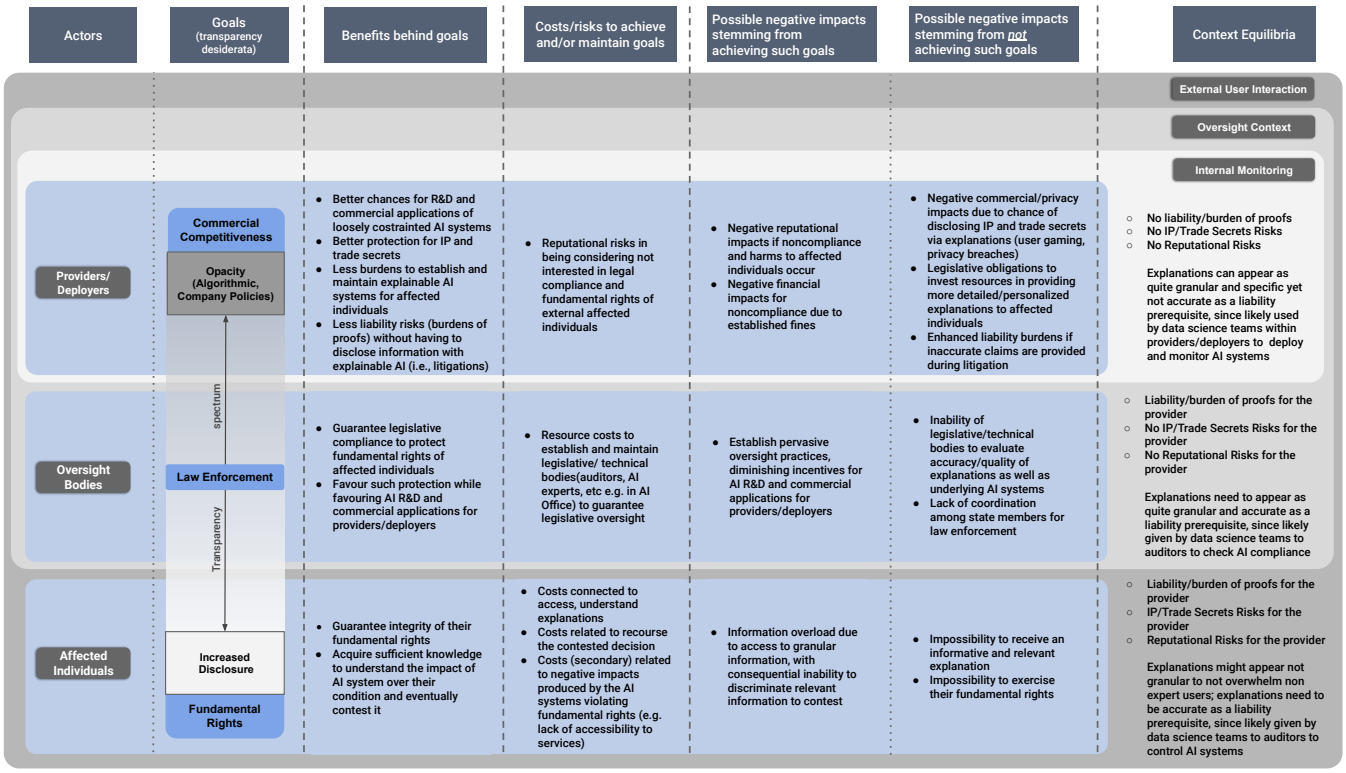

Figure 2: TEFT key actors, components, and context equilibria

about tensions between individual rights and commercial interests (Busuioc, Curtin, and Almada 2023; Grochowski et al. 2021; Hacker et al. 2020).

2. **Oversight Bodies**: Regulatory agencies, sector authorities, and standards bodies in the need to balance citizens' fundamental rights with provider-aligned interests given their economic growth mandates and moderate enforcement capacities (Article 74), reflecting the challenges in regulatory capacity and expertise identified in the literature (Kaminski 2019; Edwards and Veale 2018; Wachter, Mittelstadt, and Floridi 2017).
3. **Affected Individuals & Civil Society**: Vulnerable minorities, general citizens, digital rights groups, and consumer welfare associations who need to protect their fundamental rights by seeking personalized explanations for unfair outcomes and inclusive participation in oversight governance (Article 67, Recital 171), echoing the importance of explainability for informed self-advocacy discussed in the literature (Vredenburgh 2021).

II. Benefits Each actor may derive different benefits from achieving their goals related to algorithmic transparency. For AI system providers, key benefits include better chances for research and development (R&D) and commercial applications of loosely constrained AI systems, stronger protection for IP and trade secrets, fewer burdens to establish and maintain XAI systems for affected individuals, and reduced liability risks by limiting the disclosure of information through explainability. Oversight bodies, on the other hand, benefit from ensuring legislative compliance to protect the fundamental rights of affected individuals while also supporting AI R&D and commercial applications for providers/deployers. For affected individuals, the main benefits are safeguarding the integrity of their fundamental rights and acquiring sufficient knowledge to understand the impact of AI systems on their circumstances and potentially contest decisions.

III. Costs Each actor may face different costs and risks in pursuing their goals related to algorithmic transparency. AI Providers/Deployers may encounter reputational risks if they are perceived as uninterested in legal compliance and the fundamental rights of affected individuals. Oversight bodies may incur significant resource costs to establish and maintain legislative/technical bodies (e.g., auditors, AI experts, AI Office) to ensure effective oversight (Novelli et al. 2024; European Commission 2024a,b). Affected individuals may face costs associated with accessing and understanding explanations, contesting decisions, and dealing with the secondary impacts of AI systems violating their fundamental rights (e.g., lack of access to services). These costs are not distributed equally among actors and may have disproportionate impacts on certain groups. For example, individuals from marginalized or disadvantaged backgrounds may face higher barriers to accessing and understanding explanations or contesting decisions due to linguistic, cultural, or socioeconomic factors. Similarly, smaller AI Providers/Deployers may struggle to absorb compliance costs with transparency requirements compared to larger, well-resourced firms.

IV. Potential Negative Impacts Potential negative impacts could arise both when actors' goals are achieved and when they are not.

If goals are achieved, AI Providers/Deployers may face negative reputational impacts if non-compliance and harms to affected individuals occur, as well as financial penalties for non-compliance. Oversight Bodies may inadvertently establish pervasive oversight practices that diminish incentives for AI R&D and commercial applications by providers/deployers. Affected Individuals may experience information overload due to unconstrained access to granular information, leading to difficulty identifying relevant information for contesting decisions.

On the other hand, if goals are *not* achieved, AI Providers/Deployers may encounter negative commercial/privacy impacts due to the risk of disclosing IP and trade secrets via explanations, along with legislative obligations to invest resources in providing more detailed/personalized explanations to affected individuals and enhanced liability burdens if inaccurate explanations are considered as burden of proof during litigation. Oversight bodies may struggle with the inability to evaluate the accuracy/quality of explanations and the underlying AI systems, as well as a lack of coordination among state members for law enforcement (Malgieri 2019). Affected Individuals may be left without the ability to receive informative and relevant explanations or exercise their fundamental rights.

5.2 Contextual Factors

While the TEFT framework provides a general structure for analyzing the goals, benefits, costs, and potential negative impacts of algorithmic transparency and explainability, it is important to recognize that these factors may vary significantly depending on the specific context and application domain. For example, the transparency requirements and expectations for AI systems used in high-stakes domains like healthcare, criminal justice, or financial services (aligned with Annex III categories) may be much higher than those used in lower-stakes domains like entertainment or marketing. Aside of stakes, we define at least three possible contexts with different interactions, costs/risks, and negative impacts for the actors:

Internal Monitoring (Providers and Deployers) AI Providers/Deployers use internally explanations to debug, monitor, and improve their systems. This context primarily focuses on enhancing system performance and reliability, with minimal risks related to their IP, trade secrets, or liability.

Explanations can be highly specific – i.e., through specific XAI techniques – as they are intended for internal use by data science teams. In this context the accuracy of these explanations may not be a first priority, as they would not be subject to legal or compliance purposes. One potential drawback of this context is that Providers/Deployers may have limited incentives to prioritize transparency of their AI systems beyond what is necessary for internal development and monitoring purposes.

Oversight (Providers/Deployers and Oversight Bodies) involves AI providers and deployers providing explanations to regulators and oversight bodies to demonstrate compliance with legal requirements. This context aims to enhance accountability by ensuring that AI systems adhere to established requirements (i.e., for high-risks AI systems).

Explanations in this context need to be detailed and accurate – not necessarily grounded in XAI approaches (Panigutti et al. 2023) – as they may be used to assess compliance and identify potential violations. However, these explanations may be protected from wider disclosure by confidentiality agreements, which can help mitigate risks related to IP and trade secrets. Despite the benefits of enhanced accountability, the oversight context also introduces potential costs and risks for providers and deployers, such as increased liability exposure and compliance burdens.

External User Interaction (Affected Individuals and Providers/Deployers) AI Providers/Deployers offer explanations to affected individuals who request them, often as part of a legal right or redress process. This context shall prioritize user trust, empowerment, and access to justice by enabling individuals to understand how AI systems impact their lives and contest decisions they believe to be unfair or inaccurate.

Explanations should be meaningful and understandable to non-expert users, which may require a higher level of granularity compared to explanations provided for internal monitoring or oversight purposes. However, this context also presents significant challenges for providers and deployers, including increased liability risks, potential disclosure of IP and trade secrets, and reputational damage if explanations reveal flaws or biases in their systems.

5.3 Additional Considerations

Granularly mapping explanation facets and projecting across phases, as suggested by Nannini, Balayn, and Smith (2023), can help address the gaps in explainability specifications and proportionality concerns identified in Section 4.

Granularly Map Explanation Facets To implement balanced transparency that reconciles oversight interests with confidentiality assurances (Articles 53, 55), we clarify over the need for granular mappings of specific explanation facets. Rather than mandating total visibility across all aspects of AI systems, targeted transparency measures can be considered:

- **Selective Model Component Explanations**: Explaining specific logical elements, such as learned relationships between variables, can enable contestability of particular determination bases without compromising the entire model's competitive advantage (Article 14). For example, disclosing linking functions between variables in credit approval systems allows contesting unfair or biased outcomes for specific cohorts.
- **Validation Performance Metrics**: Releasing performance dashboards on accuracy, contestations, and revisions logs can signal system reliability and responsiveness to public concerns without disclosing sensitive train-

ing data or underlying IP (Article 15). This graded transparency approach might balance accountability with data protection concerns.

- **Anonymized Data Case Summaries**: Providing representative case data summaries that preserve individual privacy can demonstrate system limitations and support contestability channels for specific determinations assessed as unfair (Article 10).

Project Across Phases Implementing strategic transparency that reconciles innovation aims with accountability requires mapping oversight needs, confidentiality risks, and value positioning across the AI lifecycle. We further outline key considerations for different phases:

- **Research and Model Building**: Early research phases warrant secrecy shields to protect ideation independence. This should be also conduct cognizant that opacity during exploratory stages also risks accumulating harms if biased techniques propagate without participatory input guardrails.
- **Design and Development**: As applications crystallize with clearer revenue and legal risk implications, selective transparency protocols can signify intentions and gather user feedback to shape iterations. This proactive approach might help preventing entrenchment of issues while aiding proactive design.
- **Deployment and Maintenance**: Post-deployment, integrated transparency systems that are secure, contextually compatible, and participative are essential for sustaining algorithmic effectiveness, managing liability and reputational risks, and respecting privacy imperatives. Demonstrating model integrity allows value positioning and justifies investments in continuous enhancements to meet evolving public expectations.

By proactively scoping risks and possibilities across long-term product and research cycles, the TEFT framework encourages an iterative approach that aligns with the regulatory experimentation and interdisciplinarity strategies discussed in the literature (Bibal et al. 2021; Casey, Farhangi, and Vogl 2019; Sovrano et al. 2022; Heymans, Gils, and Ooms 2024) and addresses the need for ongoing monitoring and evaluation of explanations identified as a friction point in the EU AI Act analysis (Article 72).

6 Research Limitations & Directions

The TEFT framework should be seen as a foundation for ongoing dialogue and experimentation, rather than a definitive solution over AI transparency in the EU digital ecosystem. With this awareness, we acknowledge the following limitations, also delimited by our research scope.

First, our focused literature review does not intend to be exhaustive of how the RTE is encountered across the entire spectrum of the AI literature, but rather is motivated by the necessity to target scholarly work actively reflecting over the EU RTE debate. Given the lively matter hereby treated, we note that the EU AI Act was recently released in Official Journal of the European Union (OJEU) on 12 July 2024 (European Parliament and Council of the European Union 2024). The full application of the Act will be gradual in the next years (e.g., high-risk system requirements fully applicable in from August 2026): during that time AI standards for explainability and transparency should be completed and harmonized (CEN-CENELEC 2020). Additionally, the European Commission's AI Pact and the new AI Office might help in conducting prototyping activities to further explore explainability possibilities (European Commission 2024a,b). The AI Act's provisions on the AI regulatory sandboxes (Article 57) and codes of conduct (Articles 56 and 95) are likely to provide valuable insights and shape the practical implementation of transparency requirements, which future adapation of the TEFT framework will need to incorporate as these developments unfold. Moreover, as scholars suggest (Heymans, Gils, and Ooms 2024; Panigutti et al. 2023), continued research to improve the robustness and reliability of XAI methods will be crucial to address the current limitations and support the effective implementation of the AI Act's transparency requirements.

Our TEFT framework is then intended to complement and inform these ongoing developments rather than replace them. Indeed, while it is grounded in a rigorous analysis of the literature and the AI Act, it constitute a conceptual tool that would benefit from further empirical refinement and operationalization. As future research directions, we envision to develop more specific metrics and thresholds for assessing the costs, risks, and benefits of transparency in different contexts, while also exploring more targeted, in-depth reviews of specific subdomains or minority views.

7 Conclusion

This research aimed to address the critical gap in synthesizing diverse viewpoints on algorithmic transparency and explainability through the lens of the EU AI Act's provisions.

We contributed to this goal in two key ways. First, we provided a comprehensive literature review synthesizing legal, technical, and socio-ethical perspectives on the RTE, from the GDPR to the EU AI Act [RQ1 - Section 3]. This review crystallized key friction points around the scope of explainability rights, stakeholder conflicts, technical constraints, and legal enforceability. We then compared current explainability provisions [RQ2 - Section 4] and proposed a novel framework [RQ3 - Section 5] to identify and discuss the interplay of factors shaping algorithmic transparency's real-world implementation. The TEFT framework addresses key gaps found in the literature, offering a structured approach to tensions between stakeholder interests, technical feasibility, legal enforceability, and socio-ethical implications.

While acknowledging the need for further empirical validation and the evolving nature of the EU's AI governance landscape, the TEFT framework provides a conceptual foundation for future research and policy initiatives aimed at ensuring the effective implementation of the AI Act's transparency requirements. As the AI Act's provisions on regulatory sandboxes and codes of conduct unfold, the framework can adapt to incorporate these developments and continue to inform the ongoing dialogue between policymakers, researchers, and practitioners.

Acknowledgments

Contribution from the ITN project NL4XAI (*Natural Language for Explainable AI*). This project has received funding from the European Union's Horizon 2020 research and innovation programme under the Marie Skłodowska-Curie grant agreement No 860621. This document reflects the views of the author and does not necessarily reflect the views or policy of the European Commission. The REA cannot be held responsible for any use that may be made of the information this document contains.

References

Adadi, A.; and Berrada, M. 2018. Peeking Inside the Black-Box: A Survey on Explainable Artificial Intelligence (XAI). *IEEE Access*, 6: 52138–52160.

Amsterdam Court of Appeal. 2023a. Decision on request from Uber drivers to Uber for information under Article 15(1)(h) GDPR on existence of automated decision-making after account deactivations.

Amsterdam Court of Appeal. 2023b. Decision on requests from Ola drivers to Ola for access to personal data under Article 15 GDPR, information under Article 15(1)(h) GDPR on existence of automated decision-making, and data portability under Article 20 GDPR.

Amsterdam Court of Appeal. 2023c. Decision on requests from Uber drivers to Uber for access to personal data under Article 15 GDPR, information under Article 15(1)(h) GDPR on existence of automated decision-making, and data portability under Article 20 GDPR.

Bertrand, A.; Belloum, R.; Eagan, J. R.; and Maxwell, W. 2022. How Cognitive Biases Affect XAI-assisted Decision-making: A Systematic Review. In *Proceedings of the 2022 AAAI/ACM Conference on AI, Ethics, and Society*, AIES '22, 78–91. New York, NY, USA: Association for Computing Machinery. ISBN 9781450392471.

Bibal, A.; Lognoul, M.; de Streel, A.; and Frénay, B. 2021. Legal requirements on explainability in machine learning. *Artif Intell Law*, 29(2): 149–169.

Brkan, M.; and Bonnet, G. 2020. Legal and Technical Feasibility of the GDPR's Quest for Explanation of Algorithmic Decisions: of Black Boxes, White Boxes and Fata Morganas. *Eur. j. risk regul.*, 11(1): 18–50.

Busuioc, M.; Curtin, D.; and Almada, M. 2023. Reclaiming transparency: contesting the logics of secrecy within the AI Act. *European Law Open*, 2(1): 79–105.

Casey, B.; Farhangi, A.; and Vogl, R. 2019. Rethinking Explainable Machines: The GDPR s Right to Explanation Debate and the Rise of Algorithmic Audits in Enterprise. *Berkeley Technology Law Journal*, 34(1): 143–188.

CEN-CENELEC. 2020. Focus Group Report - Road Map on Artificial Intelligence (AI). Accessed on: 2024-04-30.

Council of the European Union. 2022. Proposal for a Regulation of the European Parliament and of the Council laying down harmonised rules on artificial intelligence (Artificial Intelligence Act) and amending certain Union legislative acts - General approach. https://data.consilium.europa.eu/doc/document/ST-14954-2022-INIT/en/pdf. LIMITE TELECOM 472 JAI 1494 COPEN 396 CYBER 374 DATAPROTECT 320 EJUSTICE 89 COSI 293 IXIM 267 ENFOPOL 569 RELEX 1556 MI 843 COMPET 918 CODEC 1773. From: Permanent Representatives Committee (Part 1). Interinstitutional File: 2021/0106(COD).

Court of Justice of the European Union. 2023a. Judgment of the Court (First Chamber) of 7 December 2023 - Case C-634/21 - OQ v Land Hessen. ECLI:EU:C:2023:957.

Court of Justice of the European Union. 2023b. Opinion of Advocate General Pikamäe delivered on 16 March 2023 - Case C-634/21 - OQ v Land Hesse. ECLI:EU:C:2023:220.

Custers, B.; and Vrabec, H. 2024. Tell me something new: data subject rights applied to inferred data and profiles. *Computer Law & Security Review*, 52: 105956.

De Gregorio, G.; and Demkova, S. 2024. The Constitutional Right to an Effective Remedy in the Digital Age: A Perspective from Europe. In van Oirsouw, C.; de Poorter, J.; Leijten, I.; van der Schyff, G.; Stremler, M.; and de Visser, M., eds., *European Yearbook of Constitutional Law*. Forthcoming. Available at SSRN.

de Laat, P. B. 2022. Algorithmic decision-making employing profiling: will trade secrecy protection render the right to explanation toothless? *Ethics and Information Technology*, 24(3): 17.

Ebers, M. 2021. Regulating Explainable AI in the European Union. An Overview of the Current Legal Framework(s).

Edwards, L.; and Veale, M. 2018. Enslaving the Algorithm: From a "Right to an Explanation" to a "Right to Better Decisions"? *IEEE Security & Privacy*, 16(3): 46–54. Conference Name: IEEE Security & Privacy.

European Commission. 2021. Proposal for a Regulation of the European Parliament and of the Council laying down harmonised rules on artificial intelligence (Artificial Intelligence Act) and amending certain Union legislative acts. https://artificialintelligenceact.eu/wp-content/uploads/2022/05/AIA-COM-Proposal-21-April-21.pdf. 2021/0106 (COD), SEC(2021) 167 final, SWD(2021) 84 final, SWD(2021) 85 final.

European Commission. 2024a. AI Pact. https://digital-strategy.ec.europa.eu/en/policies/ai-pact. Accessed: 2024-05-14.

European Commission. 2024b. Commission Decision Establishing the European AI Office. https://digital-strategy.ec.europa.eu/en/library/commission-decision-establishing-european-ai-office. Accessed: 2024-05-14.

European Parliament. 2023a. Artificial Intelligence Act Amendments adopted by the European Parliament on 14 June 2023 on the proposal for a regulation of the European Parliament and of the Council on laying down harmonised rules on artificial intelligence (Artificial Intelligence Act) and amending certain Union legislative acts. Texts Adopted P9_TA(2023)0236. COM(2021)0206 - C9-0146/2021 - 2021/0106(COD). Ordinary legislative procedure: first reading. The matter was referred back for interin-

stitutional negotiations to the committee responsible, pursuant to Rule 59(4), fourth subparagraph (A9-0188/2023).

European Parliament. 2023b. Artificial Intelligence Act: deal on comprehensive rules for trustworthy AI. Press Release. IMCO LIBE.

European Parliament. 2024a. Artificial Intelligence Act: MEPs adopt landmark law. Press Release. PLENARY SESSION IMCO LIBE.

European Parliament. 2024b. Corrigendum to the position of the European Parliament adopted at first reading on 13 March 2024 with a view to the adoption of Regulation (EU) 2024/... of the European Parliament and of the Council laying down harmonised rules on artificial intelligence and amending Regulations (EC) No 300/2008, (EU) No 167/2013, (EU) No 168/2013, (EU) 2018/858, (EU) 2018/1139 and (EU) 2019/2144 and Directives 2014/90/EU, (EU) 2016/797 and (EU) 2020/1828 (Artificial Intelligence Act). Plenary sitting cor01. P9_TA(2024)0138, COM(2021)0206 - C9-0146/2021 - 2021/0106(COD).

European Parliament and Council of the European Union. 2016. Regulation (EU) 2016/679 of the European Parliament and of the Council of 27 April 2016 on the protection of natural persons with regard to the processing of personal data and on the free movement of such data, and repealing Directive 95/46/EC (General Data Protection Regulation). Official Journal of the European Union L 119, 4.5.2016, p. 1-88. Text with EEA relevance. In force, with current consolidated version as of 4/5/2016.

European Parliament and Council of the European Union. 2024. Regulation (EU) 2024/1689 of the European Parliament and of the Council of 13 June 2024 laying down harmonised rules on artificial intelligence and amending Regulations (EC) No 300/2008, (EU) No 167/2013, (EU) No 168/2013, (EU) 2018/858, (EU) 2018/1139 and (EU) 2019/2144 and Directives 2014/90/EU, (EU) 2016/797 and (EU) 2020/1828 (Artificial Intelligence Act).

European Parliament Committee on Legal Affairs. 2022. Opinion of the Committee on Legal Affairs for the Committee on the Internal Market and Consumer Protection and the Committee on Civil Liberties, Justice and Home Affairs on the proposal for a regulation of the European Parliament and of the Council laying down harmonised rules on artificial intelligence (Artificial Intelligence Act) and amending certain Union Legislative Acts. https://artificialintelligenceact.eu/wp-content/uploads/2022/09/AIA-JURI-Rule-57-Opinion-Adopted-12-September.pdf. Rapporteur for opinion: Axel Voss. COM(2021)0206 - C9-0146/2021 - 2021/0106(COD).

Goodman, B.; and Flaxman, S. 2017. European Union Regulations on Algorithmic Decision-Making and a "Right to Explanation". *AI Magazine*, 38(3): 50–57. Number: 3.

Grochowski, M.; Jabłonowska, A.; Lagioia, F.; and Sartor, G. 2021. Algorithmic Transparency and Explainability for EU Consumer Protection: Unwrapping the Regulatory Premises. *Critical Analysis of Law*, 8(1): 43–63. Number: 1.

Gryz, J.; and Rojszczak, M. 2021. Black box algorithms and the rights of individuals: No easy solution to the "explainability" problem. *Internet Policy Review*, 10(2): 1–24.

Hacker, P.; Krestel, R.; Grundmann, S.; and Naumann, F. 2020. Explainable AI under contract and tort law: legal incentives and technical challenges. *Artif Intell Law*, 28(4): 415–439.

Helberger, N.; and Samuelson, P. 2024. The Digital Services Act as a Global Transparency Regime. *VerfBlog*.

Heymans, F.; Gils, T.; and Ooms, W. 2024. From Policy to Practice: Prototyping The EU AI Act's Transparency Requirements. *Available at SSRN 4714345*.

Jongepier, F.; and Keymolen, E. 2022. Explanation and Agency: exploring the normative-epistemic landscape of the "Right to Explanation". *Ethics Inf. Technol.*, 24(4): 49.

Kaminski, M. E. 2019. "The Right to Explanation, Explained". *Berkeley Technology Law Journal*, 34(1).

Kim, T. W.; and Routledge, B. R. 2022. Why a right to an explanation of algorithmic decision-making should exist: A trust-based approach. *Business Ethics Quarterly*, 32(1): 75–102.

Laux, J.; Wachter, S.; and Mittelstadt, B. 2023. Three Pathways for Standardisation and Ethical Disclosure by Default under the European Union Artificial Intelligence Act. *SSRN Electronic Journal*.

Laux, J.; Wachter, S.; and Mittelstadt, B. 2024. Trustworthy artificial intelligence and the European Union AI act: On the conflation of trustworthiness and acceptability of risk. *Regulation & Governance*, 18(1): 3–32.

Malgieri, G. 2019. Automated decision-making in the EU Member States: The right to explanation and other "suitable safeguards" in the national legislations. *Computer Law & Security Review*, 35(5): 105327.

Malgieri, G.; and Comandé, G. 2017. Why a Right to Legibility of Automated Decision-Making Exists in the General Data Protection Regulation. *International Data Privacy Law*, 7(4): 243–265.

Malgieri, G.; and Pasquale, F. 2024. Licensing high-risk artificial intelligence: Toward ex ante justification for a disruptive technology. *Computer Law & Security Review*, 52: 105899.

Munch, L. A.; Bjerring, J. C.; and Mainz, J. T. 2024. Algorithmic decision-making: The right to explanation and the significance of stakes. *Big Data & Society*, 11(1): 20539517231222872.

Mylly, U.-M. 2023. Transparent AI? Navigating Between Rules on Trade Secrets and Access to Information. *IIC - International Review of Intellectual Property and Competition Law*, 54(7): 1013–1043.

Nannini, L. 2024. Habemus a Right to an Explanation: so What? - A Framework on Transparency-Explainability Functionality and Tensions in the EU AI Act. Available at SSRN.

Nannini, L.; Alonso-Moral, J. M.; Catala, A.; Lama, M.; and Barro, S. 2024. Operationalizing Explainable AI in the EU Regulatory Ecosystem. *IEEE Intelligent Systems*, 1–13.

Nannini, L.; Balayn, A.; and Smith, A. L. 2023. Explainability in AI Policies: A Critical Review of Communications, Reports, Regulations, and Standards in the EU, US, and UK. In *Proceedings of the 2023 ACM Conference on Fairness, Accountability, and Transparency, FAccT 2023, Chicago, IL, USA, June 12-15, 2023*, 1198–1212. ACM.

Novelli, C.; Hacker, P.; Morley, J.; Trondal, J.; and Floridi, L. 2024. A Robust Governance for the AI Act: AI Office, AI Board, Scientific Panel, and National Authorities. *Available at SSRN*.

Olsen, H. P.; Hildebrandt, T. T.; Wiesener, C.; Larsen, M. S.; and Flügge, A. W. A. 2024. The Right to Transparency in Public Governance: Freedom of Information and the Use of Artificial Intelligence by Public Agencies. *Digit. Gov.: Res. Pract.*, 5(1).

Panigutti, C.; Hamon, R.; Hupont, I.; Llorca, D. F.; Yela, D. F.; Junklewitz, H.; Scalzo, S.; Mazzini, G.; Sánchez, I.; Garrido, J. S.; and Gómez, E. 2023. The role of explainable AI in the context of the AI Act. In *FAccT*, 1139–1150. ACM.

Rudin, C. 2019. Stop explaining black box machine learning models for high stakes decisions and use interpretable models instead. *Nat. Mach. Intell.*, 1(5): 206–215.

Selbst, A. D.; and Powles, J. 2017. Meaningful information and the right to explanation. *International Data Privacy Law*, 7(4): 233–242.

Sovrano, F.; Sapienza, S.; Palmirani, M.; and Vitali, F. 2022. Metrics, Explainability and the European AI Act Proposal. *J*, 5(1): 126–138.

Sovrano, F.; Vitali, F.; and Palmirani, M. 2020. Modelling GDPR-Compliant Explanations for Trustworthy AI. In *Electronic Government and the Information Systems Perspective: 9th International Conference, EGOVIS 2020, Bratislava, Slovakia, September 14–17, 2020, Proceedings*, 219–233. Berlin, Heidelberg: Springer-Verlag. ISBN 978-3-030-58956-1.

Sovrano, F.; Vitali, F.; and Palmirani, M. 2021. Making Things Explainable vs Explaining: Requirements and Challenges under the GDPR. In Rodríguez-Doncel, V.; Palmirani, M.; Araszkiewicz, M.; Casanovas, P.; Pagallo, U.; and Sartor, G., eds., *AI Approaches to the Complexity of Legal Systems XI-XII*, 169–182. Cham: Springer International Publishing. ISBN 978-3-030-89811-3.

Söderlund, K.; Engström, E.; Haresamudram, K.; Larsson, S.; and Strimling, P. 2024. Regulating high-reach AI: On transparency directions in the Digital Services Act. *Internet Policy Review*, 13(1).

Vredenburgh, K. 2021. The Right to Explanation. *Journal of Political Philosophy*, 30(2): 209–229.

Wachter, S.; Mittelstadt, B.; and Floridi, L. 2017. Why a Right to Explanation of Automated Decision-Making Does Not Exist in the General Data Protection Regulation. *International Data Privacy Law*, 7(2): 76–99.

Worker Info Exchange. 2023. Historic digital rights win for WIE and the ADCU over Uber and Ola at Amsterdam Court of Appeal. https://www.workerinfoexchange.org/post/historic-digital-rights-win-for-wie-and-the-adcu-over-uber-and-ola-at-amsterdam-court-of-appeal.