Aligning XAI with EU Regulations for Smart Biomedical Devices

Sovrano, Francesco; Lognoul, Michael; Vilone, Giulia

Published in:
Proceedings of the 27th European Conference on Artificial Intelligence

Publication date:
2024

Document Version
Publisher's PDF, also known as Version of record

ECAI 2024
U. Endriss et al. (Eds.)

doi:10.3233/FAIA240568

Aligning XAI with EU Regulations for Smart Biomedical Devices: A Methodology for Compliance Analysis

Francesco Sovrano[a,*], **Michael Lognoul**[b] **and Giulia Vilone**

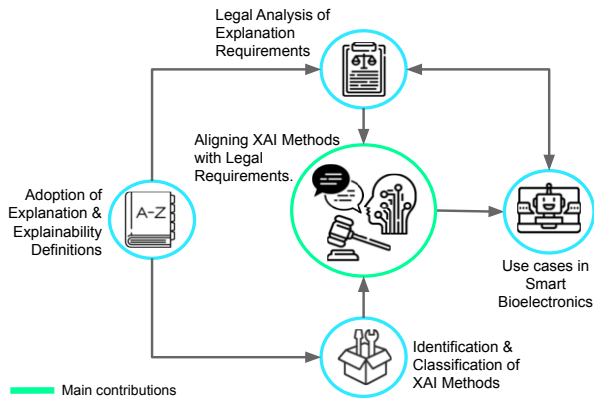
[a]University of Zurich
[b]University of Namur (CRIDS, NADI)
ORCID (Francesco Sovrano): https://orcid.org/0000-0002-6285-1041, ORCID (Michael Lognoul): https://orcid.org/0009-0005-5137-8278, ORCID (Giulia Vilone): https://orcid.org/0000-0002-4401-5664

Abstract. Significant investment and development have gone into integrating Artificial Intelligence (AI) in medical and healthcare applications, leading to advanced control systems in medical technology. However, the opacity of AI systems raises concerns about essential characteristics needed in such sensitive applications, like transparency and trustworthiness. Our study addresses these concerns by investigating a process for selecting the most adequate Explainable AI (XAI) methods to comply with the explanation requirements of key EU regulations in the context of smart bioelectronics for medical devices. The adopted methodology starts with categorising smart devices by their control mechanisms (open-loop, closed-loop, and semi-closed-loop systems) and delving into their technology. Then, we analyse these regulations to define their explainability requirements for the various devices and related goals. Simultaneously, we classify XAI methods by their explanatory objectives. This allows for matching legal explainability requirements with XAI explanatory goals and determining the suitable XAI algorithms for achieving them. Our findings provide a nuanced understanding of which XAI algorithms align better with EU regulations for different types of medical devices. We demonstrate this through practical case studies on different neural implants, from chronic disease management to advanced prosthetics. This study fills a crucial gap in aligning XAI applications in bioelectronics with stringent provisions of EU regulations. It provides a practical framework for developers and researchers, ensuring their AI innovations advance healthcare technology and adhere to legal and ethical standards.

1 Introduction

The 2023 Artificial Intelligence (AI) Index Report by Stanford University reveals that medical and healthcare applications represent one of the largest investment areas in AI (nearly 6 billion USD). Incorporating AI into smart bioelectronics for medical devices represents a tremendous leap in medical technology. This integration has resulted in substantial improvements in patient care, primarily by developing advanced control systems that can adapt in real-time to patient needs, thereby greatly enhancing the effectiveness of treatments and quality of life [22]. A key evolution in medical technology is the shift from open-loop systems, where physicians interpret data to inform decisions, to more sophisticated closed-loop and semi-closed-loop systems, where devices autonomously or semi-autonomously adjust their operations based on continuous monitoring.

Figure 1: Schematic overview of our research methodology for integrating legal requirements and XAI tools (cf. Section 3).



A significant challenge with advanced AI systems is their 'black-box' nature, which makes it hard to understand how they make decisions [5]. This challenge is critical in healthcare, where AI systems must be accurate, transparent, and accountable to enhance trust in their users and enforce responsibility [5]. This is where Explainable Artificial Intelligence (XAI) plays a crucial role, offering tools to make the inner workings of these complex systems more understandable to the diverse stakeholders involved in their operation [48]. Regulatory frameworks, particularly in the EU, implicitly require XAI to ensure AI technologies' transparency, fairness, and accountability, as emphasised in various scholarly works [52].

EU regulations are especially stringent regarding smart bioelectronics for medical devices. These devices must comply with the General Data Protection Regulation (GDPR) [9], the Artificial Intelligence Act (AIA) [11], and the Medical Devices Regulation (MDR) [10], each contributing unique requirements related to explainability. However, navigating the complex regulatory landscape poses significant challenges for developers and researchers. Implementing XAI algorithms in line with EU regulations is a major hurdle, accentuating a disconnect between theory and practice in this field [44, 30]. The motivation for our study stems from this very challenge. We carried out a thorough analysis of various XAI algorithms to determine if they can help satisfy explainability requirements set by the GDPR, the AIA, and the MDR. To this end, we

* Corresponding Author. Email: francesco.sovrano@uzh.ch.

have developed a novel methodology, summarised by Figure 1, to evaluate and understand the role of XAI in adhering to this legal framework. This study advances current understanding by categorising XAI algorithms based on their explanatory goals and matching them to the goals pursued by explainability requirements, guiding developers and researchers in selecting XAI algorithms for bioelectronic devices that better comply with EU regulations.

2 Related Work

Integrating XAI tools into compliance processes to match explanation requirements, as contained in diverse fields of law, is still an open, multi-faceted and multi-disciplinary challenge. One of the significant stumbling blocks discussed by Richmond et al. [44] is harmonising the logic followed by AI algorithms with legal reasoning and legal requirements to provide reasons or explanations.

Previous works delved into the legal and ethical requirements for explainability in Machine Learning (ML) [6], particularly in the context of the GDPR [15], the AIA [52], or other EU regulations related to sensitive fields such as finance [54]. However, our study stands out in its comprehensive and practical approach, as these different studies are not oriented towards offering a methodology to find the right XAI tools. Bibal et al. [6] investigated the increasing legal requirements for AI explainability in private and public decision-making contexts. They emphasised the implementation of these requirements in ML models and advocated for interdisciplinary research in explainability. We went one step further by providing the kind of interdisciplinary research and methodology they suggested.

Similarly to Hashemi [25], which proposes a strategy for choosing the proper XAI method for specific goals, we reviewed the XAI research to provide a synopsis of recent XAI methods and their characteristics that make them suitable candidates for healthcare. In contrast, our research carried forward by explicitly mapping XAI methods to the regulatory requirements of the GDPR, AIA, and MDR in the context of smart bioelectronics for medical devices. Our work delves into the obligations discussed at a general level by Bibal et al. [6] and translates them into operational and practical terms through focused case studies.

Although there are currently no explicit mandates for the use of XAI systems, as noted by Ebers [15] in their analyses of the GDPR, our research echoes Schneeberger et al. [46] in emphasising the crucial role of state-of-the-art XAI for ensuring compliance with various legal texts applicable in the medical sector, for instance, in protection of patient's sensitive data. However, differently from Schneeberger et al. [46], which provides an overview of the EU's legal approach to AI in the medical sector, our study goes beyond the general legislative landscape to perform a detailed analysis of specific XAI methods and their potential use for compliance with these regulations, focusing on applications for smart biomedical devices.

Additionally, our study distinguishes itself from Górski and Ramakrishna [20] by focusing on the medical field, using a broader array of XAI algorithms and systems, and conducting an extensive qualitative analysis of legal requirements, unlike their focus on the accuracy of explainability methods like Grad-CAM, LIME, and SHAP in legal text classifications as assessed by legal professionals.

3 Methodology

This research aims to bridge the gap between the technical capabilities of XAI and the legal requirements set forth by key EU regulations within the domain of smart bioelectronics for medical devices: GDPR, AI Act, and MDR. Our multifaceted methodology combines legal analysis with technical assessment and classification of XAI algorithms to increase regulatory compliance. As shown in Figure 1, our methodology comprises the following steps.

Adoption of Explanation and Explainability Definitions. To map XAI methods with legal explanatory requirements, we needed to select an appropriate definition of *explanation*. We adopted the definition formalised by Sovrano and Vitali [50], which conceptualises explanations as answers to questions that produce understanding. Among the five main definitions in contemporary philosophy, this one, rooted in Ordinary Language Philosophy, is found to align best with the legal interpretation of explanations [49, 52]. According to this definition, an explanation provides sufficient information for an audience's understanding. This differs from other definitions that require an explanation tailored to someone's mental model or showing causal relationships. Indeed, in the legal context, explanations do not necessarily need to be fully personalised [55] and can encompass more than just causality [6, 51].

Legal Analysis of Explanation Requirements. A legal expert (the 2nd author of this paper) thoroughly analysed the GDPR, AI Act, and MDR to pinpoint their explanation requirements and characteristics. Then, following an inductive coding approach [17], the legal expert identified the high-level explanatory goals underlying these requirements, e.g., ensuring that systems' deployers understand risks related to the use of an AI system or can interpret a system's output, guaranteeing that outputs can be reviewed or contested, etc.

Identification and Classification of XAI Methods. Concurrently, two AI experts (the 1st and 3rd author of this paper) conducted in three phases a literature review to compile a comprehensive (but not exhaustive) list of existing XAI methods. Initially, the search query *"XAI survey"* was used on Google Scholar, targeting relevant publications in top journals from 2023. Subsequently, the research scope was broadened to incorporate insights from the XAI survey of Vilone and Longo [53], to ensure a more comprehensive synopsis. Finally, the list of XAI algorithms was integrated with algorithms known to the experts but not mentioned in the surveyed literature. These algorithms were categorised based on their explanation format, input format, and model-agnostic status to discern the types of explanatory questions they could address, such as "what happens if feature X is changed" or "what is the contribution of feature Y to the output". We used a question-driven design process similar to that of Liao et al. [34], in which XAI methods are mapped to explanatory questions based on their characteristics. Differently from Liao et al. [34], our mapping did not involve only interrogative particles (e.g., why, how), but we formulated complete questions (see Table 5) via an inductive coding approach [17], allowing the questions to emerge naturally from the characteristics of the XAI methods.

Aligning XAI Methods with Legal Requirements. By performing a deductive thematic analysis [17], we mapped the XAI questions to the legal explanatory goals enshrined in the GDPR, AIA, and MDR. This was possible because the adopted definition of explanation is framing explanations as answers to questions. Eventually, we could identify congruence where XAI capabilities can be exploited to help meet the stipulated legal explanation requirements. This matching process ensures that the selection of XAI methods is technologically sound and legally robust. To aid developers and researchers in selecting the most appropriate XAI algorithms for different bioelectronic devices, we developed a set of instructions (see Section 8). These instructions and the methodology provide a fundamental framework designed to be flexible and seamlessly incorporate newly emerging XAI algorithms and evolving regulations.

Identification of Case Studies in Smart Bioelectronics. We focused on two use cases in the field of smart bioelectronics determined by the type of control they employed: 1) closed-loop and 2) open- and semi-closed-loop. The distinction is significant as it influences the decision-making processes and the applicable legal frameworks. For example, the GDPR's right to explanation pertains to high-stakes, fully automated algorithmic decision-making in closed-loop systems.

4 Background

This section provides background information on AI-based biomedical technologies, EU regulations, and XAI.

4.1 Smart Bioelectronics and Biomedical Devices

Biomedical devices is an umbrella name that covers a wide variety of tools used to help diagnose, prevent, and treat diseases [31]. *Bioelectronics* refers to a subset of specialised biomedical devices that combine electronic technology, like sensors, with biology and medicine. These devices can interact with biological systems, from whole organs to tiny cellular components, in various ways, such as using light, magnetic, or chemical methods [28]. Based on their decision-making mechanisms, bioelectronics and biomedical devices can be categorised into open-loop, closed-loop, and semi-closed-loop control systems. *Open-loop control systems*, such as Electrocardiograms (ECG), provide only outputs instrumental in the decision-making of healthcare professionals. In contrast, *closed-loop systems* autonomously adjust their operations based on continuous monitoring. For example, artificial pancreas systems for diabetes management autonomously monitor glucose levels and administer insulin [35]. *Semi-closed-loop systems* represent an intermediate approach where the machine instructs a patient to manually perform life-saving actions based on data (e.g., manual insulin injection adjustments based on a Continuous Glucose Monitoring System [35]).

In addition to loop-based categorisation, biomedical devices can be classified based on their potential risks to human health. This classification is influenced by the device's operating mode and characteristics, such as whether it is invasive or non-invasive. The classification is determined with relevant legislation (i.e., MDR) [1].

Neural implants represent a fascinating intersection of AI and neurotechnology and can be split between Brain-Computer Interfaces (BCIs) and Computer-Brain Interfaces (CBIs). Both prosthetic devices establish direct communication between the human brain and external hardware or software. BCIs use decoding algorithms to restore lost functions, while CBIs exploit encoding algorithms to convert external sensory signals to neural stimulation patterns [41]. Depending on their level of autonomy and decision-making mechanisms, neural implants are subject to different explanation requirements (see Table 1). One example of a closed-loop neural implant is the Responsive Neuro Stimulation (RNS) system, which is designed for individuals with epilepsy who do not respond well to medications and are not candidates for epilepsy surgery. Epileptic seizures are caused by abnormal electrical activity in the brain. RNS system records intracranial EEG patterns to timely activate a stimulation designed to mitigate such activity [21]. A Spinal Cord Stimulator (SCS) is instead an example of semi-closed loop neural implant used to alleviate chronic pain. It consists of an implanted device that delivers electrical pulses to the spinal cord to disrupt pain signals before they reach the brain. Unlike closed-loop systems, where all adjustments are fully automated, patients and healthcare professionals often have important control over these devices and their stimulation decisions. They can adjust the stimulation settings within certain limits, such as changing the intensity, frequency, or coverage of the pulses [19].

4.2 EU Regulations Relevant to Smart Biomedical Devices: Scopes and Notions

The *Medical Devices Regulation* [10] governs the placing on the market and use of medical devices in the EU (Art. 1.1) for the diagnosis, prevention, prediction, monitoring, treatment, of diseases, injuries, or disabilities. These devices must primarily operate not through pharmacological, immunological, or metabolic means but can be supported by them (Art. 2.1).

The *General Data Protection Regulation* [9] instead applies to personal data processing *i)* by EU-based data controllers or processors, or *ii)* involving EU residents' data processed by a controller located outside the EU, (Art. 2 and 3). Personal data is information about an identifiable person: the 'data subject' (Art. 4.1). 'Processing' encompasses: collection, organisation, storage, consultation, use, disclosure and erasure (Art. 4.2). A 'data controller' sets personal data processing purposes and means (Art. 4.7), while a 'processor' handles data on behalf of a controller (Art. 4.8).

The *Artificial Intelligence Act* [11], adopted in June 2024, applies to AI systems marketed or used in the EU or whose outputs are employed in the EU, regardless of the provider's or deployer's location (Art. 2). The AI systems covered are software able to infer, from their inputs, how to generate outputs (e.g., predictions, content, recommendations, or decisions, Art. 3.1). A 'provider' develops or commissions AI systems for market placement or service (Art. 3.2), and a 'deployer' employs an AI system, excluding for personal, non-professional use (Art. 3.4). High-risk AI systems include those covered by EU legislation listed in Annex I, like MDR, when requiring third-party conformity assessments and those listed in Annex III, e.g., for remote biometric identification (Art. 6).

4.3 XAI Algorithms

XAI literature features a variety of domain-dependent and context-specific methods that differ in their explanation generation strategies, formats, and applicability to disparate data and learning algorithms [27]. Researchers have developed taxonomies to aid in selecting suitable XAI methods for specific problems. A key contribution to this paper comes from the work of Liao et al. [33], who categorise XAI methods based on the questions they address, and [53], who organised XAI methods by stage, scope, and format. Firstly, the stage category splits explanations between ante-hoc and post-hoc. Ante-hoc methods aim to build inherently explainable models, while post-hoc methods seek to clarify the logic of an already trained model using an external explainer. Secondly, explanations are divided between having a global (explaining the entire model's process) and a local (explaining individual inferences) scope [53]. Thirdly, explanations differ in their format. Some consist of vectors, tensors or matrices of numbers pointing out the most relevant input features. Other explanatory formats are texts, charts and diagrams, rules, or a combination of these formats. As discussed in Section 5 and shown in Table 2, some regulations favour ex-ante explainability [45]. However, much of the research in XAI focuses on post-hoc solutions [53].

5 Explanation Requirements and Legal Explanatory Goals

This section examines the EU regulations outlined in Section 4.2, focusing on their mandates for explanations. Our analysis encompasses

the contents and formats of the explanations required. We also note the interactions and overlaps among rules imposed on devices with varying autonomy. Next, we delineate and classify the legal objectives derived from these requirements. This offers a holistic view of the goals behind the explanation requirements in the EU's regulatory framework for AI and digital health technologies.

Medical Devices Regulation. The MDR (Art. 10.11 and Annex I) mandates that medical devices include user instruction that, aimed at users or patients, must cover: *i)* the device's intended purpose; *ii)* indications, contra-indications, residual risks, and side effects; *iii)* target patient groups; *iv)* performance characteristics; *v)* suitability information for healthcare professionals; *vi)* user requirements for proper device usage; *vii)* any preparatory treatment needs, like calibration; and *viii)* guidance on verifying device installation and operational readiness (Annex I.23.4).

Art. 2.37 of MDR distinguishes between 'patient' and 'user' (healthcare professional or layperson), affecting the complexity of explanations based on the intended audience. For healthcare professionals, detailed instructions are suitable, while simpler information and language are necessary for laypersons or patients. Regardless of the audience, the MDR mandates "easily legible and comprehensible" instructions. The MDR does not specify a format but suggests written explanations, allowing for graphical and numerical forms. The MDR's primary goals are to enhance transparency and safety around medical devices for public health and to empower users and patients to make informed decisions (recital 43).

General Data Protection Regulation. The GDPR regulates decisions made solely through automated means and involving the processing of personal data, with legal or significant effects on individuals (Art. 22.1). It imposes safeguards, including providing data subjects with "meaningful information about the logic involved" and the significance and consequences of processing (Art. 13.2(f), 14.2(g), 15.1(f)). Additionally, Recital 71 mentions the obtaining of "an explanation of the decision reached" to enable individuals to understand and contest decisions (Art. 22, Recital 71) and potentially influence future behaviour to obtain a desired outcome, though this aspect is less emphasised (e.g., see p. 26 [7]). The European Data Protection Board (EDPB) advises data controllers to explain "the rationale behind, or the criteria relied on" for these decisions (p. 25 [7]). The GDPR's key principles include transparency in data processing and empowering data subjects (Art. 5, 12-22; Recitals 29, 58, 60). Therefore, it requires explanations to be adapted to the recipients' background knowledge [47] and to be "concise, transparent, intelligible, and easily accessible, using clear and plain language" (Art. 12.1). The EDPB further clarifies that explanations should enable understanding of the reasons behind decisions without disclosing complex algorithmic details (p. 25) [7]. Both the GDPR and EDPB guidelines do not prescribe a specific format for explanations, allowing for flexibility.

Artificial Intelligence Act. The AIA sets strict transparency standards for high-risk AI systems, demanding them to be "sufficiently transparent to enable deployers to interpret and use them appropriately" (Art. 13.1) and to include detailed instructions for use (Art. 13.2), covering: *i)* system characteristics, capabilities, and limitations of performance, including: its intended purpose, accuracy (with metrics), robustness, and cybersecurity, potential health, safety, or fundamental rights concerns, when available, ability to provide information explaining its output, when appropriate, performance characteristics for target groups, when appropriate, specifications about input data and information on training, validation, or testing datasets, when available, information enabling deployers to interpret the system's output and use it appropriately; *ii)* any planned changes to the system or its performance; *iii)* human oversight measures, including the measures to facilitate system outputs interpretation; *iv)* the system's expected lifetime and maintenance requirements, as well as the resources needed to run it; and *v)* a description of the mechanisms to collect, store and interpret the system's logs.

The AIA also mandates human oversight measures (Art. 14), requiring instructions for deployers and human oversight personnel (Art. 13, 14) with sufficient information for understanding system capacities and limitations, interpreting outputs, making usage decisions, and intervening if necessary. These instructions should be "concise, complete, correct, clear, relevant, accessible, and comprehensible" (Art. 13.2). The AIA also emphasises the necessary competence, training, and authority needed for oversight personnel (Art. 26), suggesting detailed explanations are essential. Explanations may be in various formats, including digital (Art. 13), potentially using text, visuals, or interactive tools (Art. 14). The AIA aims to ensure that deployers can understand and properly use high-risk systems and maintain operator control (Recitals 72, 73), supporting a "high level of protection of health, safety, and fundamental rights" (Recital 1).

Legal Explanatory Goals. The explanation requirements detailed so far apply to any bioelectronic component and biomedical device that enters the scope of GDPR, AIA, and MDR and meets the conditions which trigger their explanation requirements (e.g., fully automated high-stakes decisions based on personal data for the GDPR). These requirements are not mutually exclusive, and a cumulative application of two (or more) requirements may be needed. In particular, the device's autonomy level will influence the number of requirements to be complied with, as illustrated in Table 1.

Based on the analysis of the legal explanatory requirements led in this section and following the methodology described in Section 3, we identified 11 high-level legal explanatory goals to which these requirements pertain. Table 2 presents the identified goals, specifying the relationship with the EU regulations. Importantly, we noted that goals and regulations have a many-to-many relationship, as each goal may be related to one or more regulations. On top of that, building on the notions described in 4.3, we clarify in Table 2 whether each goal requires global or local explanations to be achieved, as well as the stage at which they should be provided: ex-ante, or ex-post.

6 A Categorisation of XAI in Terms of Explanatory Goals

This section presents a categorisation of XAI methods identified in the XAI literature and elaborates on their roles in fulfilling the legal explanatory goals of Section 5. This classification stems from the methodology outlined in Section 3, considering that XAI explanations answer specific questions about AI models and their outputs.

We organised the XAI methods based on explanation format, scope, input type, stage of application, and model specificity. This classification, grounded on established taxonomies (cf. Section 4.3) and Liao et al. [33]'s methodology, aids in pinpointing which explanatory question can be answered by each XAI method and, subsequently, which explanatory goals it addresses. Liao et al. [33] exploited the explanation format to identify which questions can be answered by the XAI methods. For example, counterfactual methods [26] inspect how the output changes when the input instance is modified, generating a 'what-if' scenario that manifests what leads to a desired outcome. Thus, these explanations can answer the question "What minimal changes would need to be made to input to change its prediction?". Instead, similarity-based XAI methods [40] show

Table 1: Applicability of Legal Requirements to Different Device Types.

Device Type	MDR (Art. 10.11 and Annex I.23.4)	AIA (Art. 13-14)	GDPR (Art. 13-15 and 22)
Open-loop	Applicable (if medical device)	Applicable (if high-risk AI system)	Not applicable (no fully-automated decision)
Semi-closed-loop	Applicable (if medical device)	Applicable (if high-risk AI system)	Not applicable (no fully-automated decision)
Closed-loop	Applicable (if medical device)	Applicable (if high-risk AI system)	Applicable (if high-stakes decision)

Table 2: Legal Explanatory Goals, Related Regulations, and their Scope (Global/Local) and Stage (Ex-Post/Ex-Ante).

ID	Legal Explanatory Goal	Related Regulation(s)	Scope	Stage
A	Understand the risks related to the use of the system	MDR, AIA	Global	Ex-ante
B	Understand the conditions under which the intended users should use the system or opt-out	MDR, AIA	Global	Ex-ante
C	Understand the consequences of decisions taken by the system	GDPR	Any	Any
D	Ensure that decisions taken by the system can be reviewed or contested	GDPR, AIA	Local	Ex-post
E	Understand what to do to change a future decision of the system	GDPR	Any	Any
F	Detect and address anomalies, dysfunctions, or unexpected performance	AIA, MDR	Any	Any
G	Understand why a specific decision has been taken	GDPR, AIA	Local	Ex-post
H	Understand how to use the system	MDR, AIA	Global	Ex-ante
I	Understand the general logic of the system	AIA	Global	Ex-ante
J	Understand the accuracy scores and the performance of the system's outputs	AIA, MDR	Global	Ex-ante
K	Interpret the system's output	AIA	Local	Ex-post

how the model behaved with similar inputs, addressing questions like "Which past instances yielded similar predictions to this input?".

Our methodology for aligning XAI questions with legal explanatory goals (Tables 3, 4, and 5), as detailed in Section 3, involves evaluating each XAI question against regulatory requirements. Specifically, we matched legal explanatory goals requiring global explanations to global XAI methods only, while legal goals with a local perspective were linked to local or global XAI methods, as global explanations can sometimes provide local insights. Global feature attribution XAI methods like TreeSHAP [37] are suitable to respond to the question "What are the most important features influencing all the model's predictions?", but not "How does a specific feature influence the prediction for an individual instance?", which has a more local scope and can be addressed with XAI methods such as LIME [42]. Indeed, TreeSHAP's focus is on understanding the slightest changes in the input features that would lead to a different prediction and showing how alternative outcomes could be achieved. Therefore, TreeSHAP best aligns with goals A, B, F, H, and I. In general, global feature attribution XAI methods aid in achieving goals B and F as they allow intended users to make informed decisions, such as detecting anomalies and opting out of using the system if they cannot provide these features. Goal H emphasises understanding system usage beyond following user manuals. It can involve experimenting with the system to determine the correct inputs for the desired outcome, focusing on clarity about the inputs to use. Knowing which features have more impact on the output can speed up the process of finding this sweet spot. Thus, it was associated with global feature attribution and rule-based methods, like Shapley Flow [56].

Instead, local model-agnostic XAI methods, like the counterfactual and contrastive explainers, can identify minimal input changes for different outcomes, aiding in goals D, E, F, G, and K. Yet, it cannot clarify when/how to use the system or the logic underlying its predictions. Not all local XAI methods can answer those five goals. For instance, given that we interpreted goal E as needing specific instructions to alter the system's output, local feature attribution and saliency maps do not contribute to reaching that goal. They simply highlight important features without suggesting how to modify them. Often, these features cannot be eliminated merely by zeroing them out, so goal E was not linked to either approach. These two local XAI methods usually address only goals D, F, G, and K. Also, activation maximisation and layer-wise relevance propagation provide similar explanations to feature attribution methods. The main difference is that they do not provide information to review or contest a decision (D) since they do not explain the contribution of the input's features to the output.

We also found that no XAI tool can explain the consequences of a decision; it only explains the process and reasons behind it. Thus, explaining consequences should be done manually to achieve goal C. Similarly, we found that goal J does not normally pertain to XAI, as it is more about the usual testing and validation steps performed when building an AI system.

For goal A, relevant risks under the considered laws are harms caused when the system works as intended (e.g., discriminatory yet accurate predictions stemming from biased data) or malfunctions. For this reason, we could only associate global feature attribution and rule-extraction XAI algorithms with A.

The stage of an XAI method (ex-ante or ex-post) also influences the question framing. Ex-ante methods help understand the data and its features before a prediction is made or finalised, leading to questions like "What set of rules does the model follow to make all predictions?". Such questions are associated with legal goals A, B, H-J. Instead, ex-post methods drive questions about interpreting these predictions, such as "Why did the model make this specific prediction for this instance?", which are linked to goals D, G and K.

The XAI methods can be segmented into model-specific and model-agnostic, as systematically presented in Tables 3-5. This arrangement facilitates the identification of appropriate XAI methods for case studies in smart bioelectronics for medical devices (see Section 7). Model-specific XAI methods, listed in Table 3 alongside the questions and legal goals we mapped them with, are tailored to specific model types, such as tree-based models or Deep Neural Networks (DNNs). Within this category, we decided to show separately, in Table 4, those methods designed for time series AI models. They are particularly relevant to biomedical devices because they must work with data sequences that vary over time, such as the heart rate in the case of ECGs. Temporal integrated gradients [14], for instance, help identify patterns and segments in time series that are deemed critical by a model to interpret outputs from devices like pacemakers or ECGs. Conversely, model-agnostic XAI algorithms (Table 5) apply to any model type, providing flexibility in their application. For instance, LIME [42], SHAP and KernelSHAP [36] identify the most important features influencing predictions in any AI model.

7 Case Studies: Closed-Loop and Semi-Closed-Loop Control

AI-enhanced neural implants can detect early signs of stroke, improve memory, and help control paralysed limbs to perform fine motor tasks, e.g., holding a glass (cf. Section 4.1).

Table 3: Model-Specific XAI Methods with Explanatory Goals.

Question	XAI Algorithms	Applicable Model Types	Expl. Goal ID(s)	Explanation
What are the most important features influencing all predictions of the model?	Global feature attribution methods: TreeExplainer [37], CAVs [29]	Tree-based models, DNNs	A, B, F, H, I	Identify key features for predictions to understand system's conditions of use, risks, general logic, and detect anomalies.
How do interactions between features affect all predictions of the model?	Global feature attribution methods [37]	Tree-based models, Bayesian networks	A, B, F, H, I	Identify key features for predictions to understand system's conditions of use, risks and general logic and detect anomalies.
What is the model's inner logic?	NNKX [8], ExtracRule [18]	DNNs, SVM	A, B, D-I, K	Transparent models mimicking the behaviour of a black-box.
What set of rules does the model usually follow to make all predictions?	Rule-based algorithms [12, 56]	Decision trees, random forests, linear models, DNNs	A, B, D-I, K	Clarify the rules to understand system usage, risks, logic, decisions, outputs, how to contest or change them, and detect anomalies.
What parts of an input (e.g., an image) influence the model's decision?	Saliency maps [16], LRP [3]	DNNs	D, F, G, K	Identify influential input areas to review or contest decisions, detect anomalies, interpret specific decisions and outputs.
What contributions do individual neurons in a DNN make to the final prediction?	Activation maximisation [32]	DNNs	F, G, K	Analyse neuron contributions for anomaly detection, decision and output interpretation.
How do different layers in a DNN contribute to a prediction?	Layer-wise relevance propagation [32, 29]	DNNs	F, G, K	Examine layer contributions for anomaly detection, decision and output interpretation.
How to interpret a neural net's internal state in terms of human-friendly concepts?	Concept-based methods [29]	DNNs	D, F, G, K	Evaluate key contributions to decisions and outputs, review or challenge them, and identify anomalies.
What are the most similar instances to a given input with respect to a model's prediction?	Self-Organising Maps (SOM) [24]	SVM	D, F, G, K	Compare similar instances to review/contest/interpret decisions/specific outputs and detect anomalies.

Table 4: XAI Methods for Time Series Models (neural networks) with Explanatory Goals.

Question	XAI Algorithms	Expl. Goal ID(s)	Explanation
What points in the time series are most important for the model's decision?	Feature attribution methods [14]	D, F, G, K	Identify key points to review/contest/interpret decisions/specific outputs, and detect anomalies.
What are the key segments in a time series that influence the model's output?	Saliency maps [2]	D, F, G, K	Clarify relevant segments to review/contest/interpret decisions/specific outputs and detect anomalies.
What minimal changes in a time series would alter its predicted outcome?	Counterfactual explanations [26]	D, E, F, G, K	Identify minimal changes leading to review/contest/interpret decisions/specific outputs, detect anomalies, and make changes to future decisions.
What parts of an input (e.g., an image) influence the model's decision?	Visual attribution methods [39]	D, F, G, K	Determine influential input parts to review/contest/interpret decisions/specific outputs and detect anomalies.
How does a specific feature influence the prediction for an individual instance?	Feature importance analysis [38]	D, F, G, K	Assess feature influence on individual predictions to review/contest/interpret decisions/specific outputs and detect anomalies.

Responsive Neuro Stimulations (RNS) are closed-loop systems. They are, in principle, subject to GDPR explanation requirements as they process (sensitive) personal data to make high-stakes, fully automated decisions. Indeed, a stimulation performed at the wrong time, on the wrong area of the brain or with the wrong electrical pulses might have side effects with varying severity [21], including pain, discomfort, sensory disturbances, etc. RNS systems are also subject to the explanation requirements of the MDR and the AIA. According to the MDR rules, they must undergo a third-party conformity assessment, making them high-risk systems as well. Since intracranial EEG patterns are extrapolated from time series, we look at Table 4 for suitable XAI methods. According to our methodology, surrogate models (as also suggested by Rudin [45]) or a mix of counterfactual, rule-based, and global feature attribution XAI methods have high chances of meeting the GDPR (goals D, E, G), the AIA (goals A, B, D, F-I, K) and MDR (goals A, B, F, H) explanation requirements. However, as explained in Section 8, these combinations may not necessarily meet all the legal explanatory goals identified, requiring the integration of other XAI tools or human intervention.

Table 5: Model-Agnostic XAI Methods with Explanatory Goals.

Question	XAI Algorithms	Expl. Goal ID(s)	Explanation
What is the inner logic of the model?	Surrogate models [4]	A, B, D-I, K	Transparent models mimicking the behaviour of a black-box.
How does a specific feature influence the prediction for an individual instance?	LIME [42], SHAP [36]	D, F, G, K	Assess feature impact on individual outputs to review/contest/interpret decisions/specific output and detect anomalies.
What are the most important features influencing all predictions of the model?	Global feature attribution methods like SHAP [36]	A, B, F, H, I	Identify key features for predictions to understand the system's conditions of use, risks and general logic and detect anomalies.
What are the most similar instances to a given input with respect to a model's prediction?	Similarity-based methods [40]	D, F, G, K	Compare similar instances to review/contest/interpret decisions/specific outputs and detect anomalies.
What minimal changes would need to be made to an input to change its prediction?	Counterfactual explanations [23]	D, E, F, G, K	Identify changes for different outcomes pertinent to review/contest/interpret decisions/specific outputs, detect anomalies, and make changes to future decisions.
Why this output instead of another?	Contrastive explanations [13]	D, E, F, G, K	Clarify reasoning behind outputs to review/contest/interpret decisions/specific outputs, detect anomalies, and make changes to future decisions.
What are the conditions or features of the input that, when held fixed, are most responsible for a particular model's prediction or classification?	Anchors [43]	D, F, G, K	Highlights key features to review/contest/interpret decisions/specific outputs and detect anomalies.
How would changing multiple features simultaneously affect the model's prediction for a specific instance?	Counterfactual and interaction detection methods [23]	D, E, F, G, K	Examines combined feature effects relevant to review/contest/interpret decisions/specific outputs, detect anomalies, and make changes to future decisions.
How can we understand the model's decision for a specific instance in the context of its training data?	Contextual analysis methods [40]	D, F, G, K	Provide decision context to review/contest/interpret decisions/specific outputs and detect anomalies.

On the other hand, Spinal Cord Stimulator (SCS) systems, which are semi-closed loop devices, are in principle not subject to GDPR's explanation requirements. However, as RNS, SCS systems are medical devices that can lead to complications like paresthesia, infections, epidural hematoma, nerve injury, paralysis, and even death [19]. Therefore, they will need to comply with the explanation requirements contained in the MDR and the AIA: they are high-risk AI systems, as they have to undergo a third-party conformity assessment under the MDR. Hence, according to our methodology, surrogate models or a mix of counterfactual, rule-based, and global feature attribution XAI methods still have a high chance of meeting the AIA and MDR legal explanatory goals.

8 Instructions for Use & Discussion of Findings

This study introduces a multi-faceted, multi-step, multi-domain methodology for aligning XAI tools with EU regulations, address-

ing a major gap in the field [44]. Matching the legal and the XAI fields presents a few challenges, such as integrating their respective technical languages and reconciling the differing explanatory goals of AI systems and EU regulations [44].

Nevertheless, this methodology provides a practical approach for selecting appropriate XAI tools for new AI use cases in healthcare. It involves a multi-phase process (see Figure 1) tailored to specific applications, ensuring the chosen XAI tools are appropriate to help meet legal explanatory requirements. Readers interested in applying this work should start from Section 5, where requirements' main features are discussed, and the legal explanatory goals are identified, and then move to Section 6, where the XAI algorithms are concretely mapped to the objectives pursued by explanation requirements.

Our methodology is adaptable and can accommodate future developments in AI applications, XAI algorithms, and evolving legal requirements. For instance, interested parties should follow these steps when applying our methodology to new or different explanation requirements. First, the identification of: *i)* the recipients of the explanations and their background knowledge; *ii)* the level of detail required from the explanation; *iii)* the format imposed on the explanation, if any; *iv)* the explanandum (i.e., the pieces of information) required; *v)* the specific objectives pursued by the law; *vi)* the types of questions that the explanation should answer; *vii)* the moment at which they should be provided; *viii)* the scope of the required explanation. Second, based on these major features, taking an inductive approach, interested parties could either relate the requirements studied to one or more of the high-level legal explanatory goals identified in our work or determine other explanatory goals. Finally, by contrasting the identified goals with the question(s) that specific XAI methods or algorithms are meant to answer, interested parties should find matches between both and be able to select appropriate tools to help answer explanation requirements.

Conversely, readers can start by determining their chosen explanation and input format, the model's applicability (model-specific or model-agnostic) and then integrate novel XAI algorithms into our methodology. With this foundation, they can inductively formulate explanatory questions that these algorithms are meant to address, generating complete questions. Subsequently, they can follow a similar process of matching these questions with the relevant legal explanatory goals, ensuring alignment between legal requirements and XAI capabilities.

Finally, each regulation we have examined in our research serves multiple orthogonal objectives (e.g., MDR aligns with the legal explanatory goals A, B, F, H, J). Consequently, even if an XAI tool is designed to address one specific goal (or more), it may not fully encompass all the legal explanatory goals intended by a given regulation. Additionally, individual XAI tools frequently only partially fulfil the objectives associated with a particular goal. Therefore, practitioners and researchers need a case-by-case assessment when applying our work in real-world scenarios. This assessment should determine the extent to which the selected XAI methods are required to implement a sufficiently diverse set of tools to ensure compliance with legal requirements while avoiding unnecessary redundancies.

9 Threats to Validity

Extrinsic Threats. Extrinsic threats include the potential for new interpretations of the Regulations discussed (e.g., through EU case-law), which may alter the applicability of our findings. Additionally, while our prescribed compliance methods assist in obtaining the necessary information, the effectiveness of conveying this information to individuals with different expertise and background knowledge is yet to be determined. Furthermore, while our study concentrates on biomedical devices and their significant requirements, it is crucial to acknowledge that other EU or national laws might impose additional explanation requirements in specific contexts. As a result, some devices may encounter extra constraints, potentially necessitating a broader range of XAI tools than those discussed in this paper.

Another extrinsic limitation arises from the inherent complexities in explainability. Most existing XAI methods, such as surrogate models, SHAP, and LIME, often rely on imperfect heuristics and usually operate effectively under specific conditions, lacking theoretical guarantees. For example, surrogate models are entirely transparent but usually perform less effectively than their corresponding black-box models. This discrepancy can lead to explanations that do not accurately represent the underlying logic of the model. Instead, SHAP-based algorithms necessitate independent input features, a condition not always met in real-world applications. Other algorithms, like LIME, also have specific requirements for their correct implementation. The incorrect use of an XAI algorithm can result in misleading explanations that do not accurately address the identified legal objectives. Hence, simply employing an XAI algorithm does not guarantee adherence to the regulations discussed in this study.

Finally, our entire methodology is based on the definition of explanation from Ordinary Language Philosophy, as outlined in Section 3. Considering alternative definitions could, therefore, introduce external threats to validity and require a different methodology.

Intrinsic Threats. There are possible alternative interpretations of the law's explanatory goals and high-level objectives, which our study may not fully encompass. The limited choice of case studies is another intrinsic issue, as it does not capture the complete range of nuances within the field, potentially affecting the generalisability of our results. Lastly, the list of XAI algorithms considered in this study is not exhaustive. However, as discussed in Section 8, our approach allows for the inclusion of new XAI algorithms and legal explanatory goals, which helps to mitigate this concern.

10 Conclusion

This paper analysed many XAI methods and their compliance with key EU regulations for smart biomedical devices. Significant contributions include a novel methodology for combining legal analysis, technical assessment, and a detailed categorisation of XAI methods to analyse their legal alignment. This constitutes a practical framework for selecting suitable XAI methods that help meet the explainability requirements of the GDPR, AIA, and MDR. The findings highlight the importance of XAI in meeting such demands for legal explainability. Future research should extend the case studies to various bioelectronic and biomedical devices, analysing stakeholders' perceptions of the explanations generated by XAI.

Acknowledgements

F. Sovrano acknowledges the support of the Swiss National Science Foundation for the SNF Project 200021_197227. M. Lognoul acknowledges the support of the European Union H2020 research and innovation program, Grant Agreement No. 958339 (DENiM).

References

[1] S. Almpani, Y. Kiouvrekis, P. Stefaneas, and P. Frangos. Computational argumentation for medical device regulatory classification. *International Journal on Artificial Intelligence Tools*, 2022.

[2] B. Aydemir, L. Hoffstetter, et al. Tempsal-uncovering temporal information for deep saliency prediction. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2023.

[3] S. Bach, A. Binder, G. Montavon, F. Klauschen, K.-R. Müller, and W. Samek. On pixel-wise explanations for non-linear classifier decisions by layer-wise relevance propagation. *PloS one*, 2015.

[4] O. Bastani, C. Kim, and H. Bastani. Interpretability via model extraction. In *Fairness, Accountability, and Transparency in Machine Learning Workshop*, 2017.

[5] J. Bernal and C. Mazo. Transparency of artificial intelligence in healthcare: insights from professionals in computing and healthcare worldwide. *Applied Sciences*, 2022.

[6] A. Bibal, M. Lognoul, A. De Streel, and B. Frénay. Legal requirements on explainability in machine learning. *Artificial Intelligence and Law*, 2021.

[7] E. D. P. Board. Guidelines on automated individual decision-making and profiling for the purposes of regulation 2016/679 (wp251rev.01), 2018. URL https://ec.europa.eu/newsroom/article29/items/612053/en.

[8] A. Bondarenko, L. Aleksejeva, V. Jumutc, and A. Borisov. Classification tree extraction from trained artificial neural networks. *Procedia Computer Science*, 2017.

[9] E. Commission. General data protection regulation, 2016. URL https://eur-lex.europa.eu/eli/reg/2016/679/oj.

[10] E. Commission. Regulation (eu) 2017/745 of the european parliament and of the council of 5 april 2017 on medical devices, 2017. URL https://eur-lex.europa.eu/legal-content/DE/TXT/?uri=CELEX:32017R0745.

[11] E. Commission. Artificial intelligence act, 2024. URL https://eur-lex.europa.eu/legal-content/EN/TXT/HTML/?uri=OJ:L_202401689.

[12] S. Dash, O. Gunluk, and D. Wei. Boolean decision rules via column generation. *Advances in neural information processing systems*, 2018.

[13] A. Dhurandhar, P.-Y. Chen, R. Luss, C.-C. Tu, P. Ting, K. Shanmugam, and P. Das. Explanations based on the missing: Towards contrastive explanations with pertinent negatives. In *Advances in Neural Information Processing Systems 31 (NIPS)*, 2018.

[14] J. Duell, M. Seisenberger, et al. A formal introduction to batch-integrated gradients for temporal explanations. In *2023 IEEE 35th International Conference on Tools with Artificial Intelligence (ICTAI)*, 2023.

[15] M. Ebers. Regulating explainable ai in the european union. an overview of the current legal framework (s). *Nordic Yearbook of Law and Informatics 2020: Law in the Era of Artificial Intelligence*, 2020.

[16] N. Feldhus, L. Hennig, M. D. Nasert, C. Ebert, R. Schwarzenberg, and S. Möller. Constructing natural language explanations via saliency map verbalization. *arXiv preprint arXiv:2210.07222*, 2022.

[17] J. Fereday and E. Muir-Cochrane. Demonstrating rigor using thematic analysis: A hybrid approach of inductive and deductive coding and theme development. *International journal of qualitative methods*, 2006.

[18] G. Fung, S. Sandilya, and R. B. Rao. Rule extraction from linear support vector machines. In *11th SIGKDD international conference on Knowledge discovery in data mining*, 2005.

[19] K. Garcia, J. K. Wray, and S. Kumar. Spinal cord stimulation. 2020.

[20] Ł. Górski and S. Ramakrishna. Explainable artificial intelligence, lawyer's perspective. In *Proceedings of the eighteenth international conference on artificial intelligence and law*, 2021.

[21] F. V. Gouveia, N. M. Warsi, et al. Neurostimulation treatments for epilepsy: Deep brain stimulation, responsive neurostimulation and vagus nerve stimulation. *Neurotherapeutics*, 2024.

[22] C. Guger, N. F. Ince, M. Korostenskaja, and B. Z. Allison. Brain-computer interface research: A state-of-the-art summary 11. In *Brain-Computer Interface Research: A State-of-the-Art Summary 11*. 2024.

[23] R. Guidotti. Counterfactual explanations and how to find them: literature review and benchmarking. *Data Mining and Knowledge Discovery*, 2022.

[24] L. Hamel. Visualization of support vector machines with unsupervised learning. In *Symposium on Computational Intelligence and Bioinformatics and Computational Biology*, 2006.

[25] M. Hashemi. Who wants what and how: a mapping function for explainable artificial intelligence. *CoRR*, 2023.

[26] M. He, B. An, J. Wang, and H. Wen. Counterfactual explanations for sequential recommendation with temporal dependencies. In *International Conference on Web Information Systems Engineering*, 2023.

[27] M. R. Islam, M. U. Ahmed, S. Barua, and S. Begum. A systematic review of explainable artificial intelligence in terms of different application domains and tasks. *Applied Sciences*, 2022.

[28] M. Jafari, G. Marquez, H. Dechiraju, M. Gomez, and M. Rolandi. Merging machine learning and bioelectronics for closed-loop control of biological systems and homeostasis. *Cell Reports Physical Science*, 2023.

[29] B. Kim, M. Wattenberg, J. Gilmer, et al. Interpretability beyond feature attribution: Quantitative testing with concept activation vectors (tcav). In *International Conference on Machine Learning*, 2018.

[30] T. Kirat, O. Tambou, V. Do, and A. Tsoukiàs. Fairness and explainability in automatic decision-making systems. a challenge for computer science and law, 2022.

[31] R. H. Lam and W. Chen. Biomedical devices. *Materials, Design, and Manufacturing. Springer, Reading, Massachusetts,*, 2019.

[32] J. Li, C. Zhang, J. T. Zhou, H. Fu, S. Xia, and Q. Hu. Deep-lift: Deep label-specific feature learning for image annotation. *IEEE transactions on Cybernetics*, 2021.

[33] Q. V. Liao, D. Gruen, and S. Miller. Questioning the ai: informing design practices for explainable ai user experiences. In *Proceedings of the 2020 CHI conference on human factors in computing systems*, 2020.

[34] Q. V. Liao, M. Pribić, et al. Question-driven design process for explainable ai user experiences. *arXiv preprint arXiv:2104.03483*, 2021.

[35] Y.-J. Lin, F.-L. Mi, et al. Strategies for improving diabetic therapy via alternative administration routes that involve stimuli-responsive insulin-delivering systems. *Advanced Drug Delivery Reviews*, 2019.

[36] S. Lundberg and S. Lee. A unified approach to interpreting model predictions. In *Advances in Neural Information Processing Systems*, 2017.

[37] S. M. Lundberg, G. Erion, H. Chen, A. DeGrave, et al. From local explanations to global understanding with explainable ai for trees. *Nature machine intelligence*, 2020.

[38] H. Meng, C. Wagner, and I. Triguero. An initial step towards stable explanations for multivariate time series classifiers with lime. In *2023 IEEE International Conference on Fuzzy Systems (FUZZ)*, 2023.

[39] P. S. Parvatharaju, R. Doddaiah, T. Hartvigsen, and E. A. Rundensteiner. Learning saliency maps to explain deep time series classifiers. In *Proceedings of the 30th ACM International Conference on Information & Knowledge Management*, 2021.

[40] A. Poché, L. Hervier, and M.-C. Bakkay. Natural example-based explainability: a survey. 2023.

[41] R. P. Rao. Brain co-processors: using ai to restore and augment brain function. In *Handbook of neuroengineering*. 2023.

[42] M. T. Ribeiro, S. Singh, and C. Guestrin. Model-agnostic interpretability of machine learning. In *22nd IGKDD International Conference on Knowledge Discovery and Data Mining*, 2016.

[43] M. T. Ribeiro, S. Singh, et al. Anchors: High-precision model-agnostic explanations. In *32nd Conference on Artificial Intelligence*, 2018.

[44] K. M. Richmond, S. M. Muddamsetty, T. Gammeltoft-Hansen, H. P. Olsen, and T. B. Moeslund. Explainable ai and law: An evidential survey. *Digital Society*, 2024.

[45] C. Rudin. Stop explaining black box machine learning models for high stakes decisions and use interpretable models instead. *Nature Machine Intelligence*, 2019.

[46] D. Schneeberger, K. Stöger, and A. Holzinger. The european legal framework for medical ai. In *International Cross-Domain Conference for Machine Learning and Knowledge Extraction*, 2020.

[47] A. D. Selbst and J. Powles. Meaningful information and the right to explanation. *International Data Privacy Law*, 12 2017.

[48] F. Sovrano and F. Vitali. Explanatory artificial intelligence (yai): human-centered explanations of explainable ai and complex data. *Data Mining and Knowledge Discovery*, 2022.

[49] F. Sovrano and F. Vitali. Perlocution vs illocution: How different interpretations of the act of explaining impact on the evaluation of explanations and xai. In *World Conference on Explainable Artificial Intelligence*. Springer, 2023.

[50] F. Sovrano and F. Vitali. An objective metric for explainable ai: how and why to estimate the degree of explainability. *Knowledge-Based Systems*, 2023.

[51] F. Sovrano, F. Vitali, and M. Palmirani. Modelling gdpr-compliant explanations for trustworthy ai. In *Electronic Government and the Information Systems Perspective*, 2020.

[52] F. Sovrano, S. Sapienza, M. Palmirani, and F. Vitali. A survey on methods and metrics for the assessment of explainability under the proposed ai act. In *Legal Knowledge and Information Systems*. IOS Press, 2021.

[53] G. Vilone and L. Longo. Classification of explainable artificial intelligence methods through their output formats. *Machine Learning and Knowledge Extraction*, 2021.

[54] G. Vilone, F. Sovrano, and M. Lognoul. On the explainability of financial robo-advice systems. In *World Conference on Explainable Artificial Intelligence*, pages 219–242. Springer, 2024.

[55] S. Wachter, B. Mittelstadt, and L. Floridi. Why a right to explanation of automated decision-making does not exist in the general data protection regulation. *International Data Privacy Law*, 2017.

[56] J. Wang, J. Wiens, and S. Lundberg. Shapley flow: A graph-based approach to interpreting model predictions. In *International Conference on Artificial Intelligence and Statistics*, 2021.