Post-hoc Explainability in AI-Enabled Clinical Decision Support Systems: Key to Enhancing (Shared) Decision Making or a Trade-Off Between Explainability, Privacy and Accuracy?

Halid Kayhan[1][0000-0002-8913-7809]

[1] KU Leuven Centre for IT & IP Law (CiTiP), Belgium
halid.kayhan@kuleuven.be

Abstract.

Artificial Intelligence (AI)-enabled Clinical Decision Support Systems (CDSS) are reshaping healthcare by enhancing diagnostic accuracy, personalising treatment, and improving clinical efficiency. However, the complexity and opacity of AI models, often functioning as "black boxes", pose significant challenges to transparency and trust, particularly in high-stakes contexts like medicine. Post-hoc explainability, a subset of explainable AI (XAI), has emerged as a promising approach to address these concerns by offering insights into complex AI decision-making processes. Nevertheless, the studies and efforts to utilise various XAI methods, including post-hoc explanations, have shown that these methods come with a cost. While contributing to achieving transparency and trust, they raise various other risks and concerns. In the context of healthcare and, more particularly, of AI-enabled CDSS, these tools can be instrumental in enhancing shared decision-making between physicians and patients by informing and empowering both according to their needs and understanding, these may also lead to risks, including misinterpretation, system exploitation, over-reliance on AI outputs, and privacy risks arising from the additional information on AI model explainability and interpretability to end users. In particular, some recent studies, conducted from a technical perspective, highlight a trade-off between explainability, privacy and utility when it comes to XAI methods. This paper examines the extent to which EU legal frameworks address this trade-off in the context of post-hoc explanations within AI-enabled Clinical Decision Support Systems (CDSS), and assesses whether these provisions offer clear guidance on how to find a balance.

Keywords: Clinical Decision Support Systems, Artificial Intelligence, Explainability, Shared Decision Making, Privacy

1 Introduction

Clinical Decision Support Systems (CDSS) are transforming modern healthcare by enhancing clinical decision-making and improving patient outcomes [1]. Although CDSS

have been used for some decades, recent advancements -especially through the integration of artificial intelligence (AI)- have significantly expanded their capabilities [1], in particular in diagnostic support, personalised treatment recommendations, proactive risk prediction, and streamlined clinical documentation [2]. AI's ability to process vast and diverse datasets allows it to uncover patterns and insights that are not possible for humans to reach [3]. As a result, AI-enabled CDSS are seen as tools that can make healthcare more accurate, efficient, timely, and safer, ultimately improving the quality of care [4].

It is underscored that for establishing trust in AI-enabled CDSS, transparency is among the most prominent concerns due to the opaque algorithms, functions as black boxes; and the explainable AI (XAI), in particular, post-hoc explanations can provide a solution [5]. These methods are argued to offer explainability and interpretability, and ultimately transparency and trust in AI-enabled CDSS, enabling physicians and patients to take better-informed and shared clinical decisions utilising well-functioning, accurate, unbiased, transparent, and understandable AI models [6]. However, it is also noted that such explanations may raise further risks such as misinterpretations, system exploitation, over-reliance on AI, and privacy risks, among others [7].

Privacy risks are of particular importance given that XAI is often deemed as an enabler of regulatory compliance, e.g. the European Union's (EU) General Data Protection Regulation (GDPR) [8]. Nevertheless, there is limited legal literature on whether XAI indeed enables legal compliance, in particular with the GDPR. Furthermore, the recent research conducted from a technical perspective on the privacy risks and privacy-preserving methods in XAI points out a trade-off between explainability, accuracy, and privacy [9].

This paper aims to investigate which legal provisions in the EU legal landscape address the trade-off between explainability, privacy, and accuracy, in the context of post-hoc explanations in AI-enabled CDSS, and whether these provisions provide clarity on how to deal with this trade-off. Due to the space restraints, the technical explanations regarding different XAI methods, privacy risks, and privacy-preserving methods will not be provided, and the explanations will be limited to a high-level legal analysis.

2 Post-hoc Explainability in AI-Enabled Clinical Decision Support Systems

2.1 Background

With the increasing involvement of AI in our lives, transparency, interpretability, and explainability have become more important than ever, as they are regarded as requisite to build trust in AI systems and wide-scale acceptance of those tools. [8]. This has resulted in a growing body of literature on methods to make AI tools understandable by humans, which are known as explainable AI (XAI) methods. [8].

A study by the European Data Protection Supervisor (EDPS), offering an introduction to XAI, defines transparency, interpretability, and explainability as follows, by

noting that these concepts do not have commonly agreed-upon definitions, and some use them interchangeably [7]:

- **Transparency** refers to a specific model being understandable. The strictest sense of this concept requires the model to be understandable in its entirety. However, transparency may be assessed at different levels, from the overall model to individual components, such as parameters, to specific training algorithms. A rather less strict interpretation of transparency suggests that each individual aspect of a model, such as inputs, parameters, and computations, should be intuitively understandable. Transparency in AI is crucial for ensuring accountability, as it will provide stakeholders with the ability to examine and audit algorithmic decision-making, identify potential biases or fairness issues, and confirm that its operations comply with ethical norms and legal frameworks.
- **Interpretability** denotes the extent to which a complex, opaque AI model or a decision by such a model is understandable to humans, especially in how and why an input results in a certain output. This is essential to predict how the system will respond to specific inputs and detect when its output may be incorrect.
- **Explainability,** built on interpretability but also incorporating insight from other domains such as law, ethics and human-computer interaction, focuses on human-understandable, clear, reason-based and coherent justifications on why an AI system generates a particular output. Explainability is particularly important in high-stakes domains such as healthcare, as it is key in understanding the reasoning behind AI decisions and predictions.

While explainability is a building block for trust in AI, it may not be necessary for all AI systems, especially for the ones that are rather simple, transparent and interpretable, or certain systems that are being used only by domain experts who do not need explainability methods to understand the reasoning of those systems. Nonetheless, when non-domain experts need to understand such systems, XAI techniques may become necessary. [7].

Some define XAI as methods to build AI models that allow end-users to understand their reasoning and trust their outputs without compromising strong accuracy. [10]. However, ensuring accuracy is not a universally agreed-upon component across existing XAI definitions, while making the AI tools understandable to humans, particularly on how AI models make a decision, is a common component. [11]. Such different approaches to accuracy in XAI are important given the common assumption of a trade-off between accuracy and explainability of AI technologies. This assumption suggests that the most complex and least explainable models, such as deep-learning models, offer greater performance (and accuracy), while simpler and more interpretable models, such as decision trees, cannot offer such high performance. However, it has been pointed out that this is not true for all types of AI models. [12].

Their performance and accuracy, sometimes even surpassing the efficiency and accuracy of predictions and decisions made by clinicians, create a significant popularity of complex AI technologies, such as deep-learning, in healthcare. [11]. The "black-box" nature of these technologies has also rung the alarm bells due to risks related to a

lack of understanding in their decision-making processes, especially in sensitive domains like healthcare [13]. Nevertheless, there is no consensus on this matter, as some argue explainability is a secondary issue to clinical validation and the (clinically-validated) accuracy should be prioritised over explainability, considering the potential benefits of well-performing tools able to make complex analysis [6]. This would be in line with an argument that it would be unethical to use tools that are not best-performing and most accurate while providing patient care [12]. It is, however, stressed that there is a crucial need for further efforts to find the right balance between the performance of complex models and explainability. [12].

In line with the growing interest in XAI in general, there are limited but increasingly more studies on the use of XAI methods in decision-support tools in various sectors, including healthcare, i.e., clinical decision support systems (CDSS). [8]. Explanations are deemed important in particular for the healthcare domain, as black-box AI may lead to irreparable harm to individuals' health and even loss of lives. [11].

2.2 Explanation Types

Before diving into further details, it is noteworthy that numerous XAI techniques exist that can be divided into different categories according to the following criteria [12]:

- Based on their scope:
  - *Global explanations* on overall model logic and behaviour
  - *Local explanations* on individual decisions by revealing why a specific outcome was generated by a model
- Based on their algorithm compatibility:
  - *Model-agnostic methods* applicable across various AI models
  - *Model-specific methods* tailored to particular models
- Based on the design of AI models to be explained:
  - *Ante-hoc methods* "explainable by design or inherently interpretable" (often called white-box or glass-box methods), including techniques like decision trees, logistic regression, and rule-based systems[1]
  - *Post-hoc methods* aiming to explain decisions from inherently opaque, or "black-box," models such as random forests, support vector machines, and neural networks[2]

AI-enabled CDSS often employ advanced technologies such as machine learning algorithms, Natural Language Processing (NLP), and deep learning models, which have sophisticated decision-making processes as a result of processing vast, various and

[1] Despite their by-design explainability, which results in them being model-specific, ante-hoc methods may be able to provide only a limited interpretability in complex scenarios, such as in high-dimensional contexts or when dealing with intricate interaction terms and deeply nested decision trees [12].

[2] While being mostly model-agnostic, post-hoc methods include methods providing global explanations, such as BETA and GAM, as well as methods providing local explanations, such as Anchors, LIME, and SHAP [12].

often complex data, and, thus, function as black boxes [2]. This requires interpreting their decisions after the decisions are made, which could mainly be done by using post-hoc explanations. [7].

2.3 Benefits of Post-hoc Explainability

Without proper explanations, AI models cannot be correctly evaluated by their developers or deployers and the deficiencies, such as bias, cannot be addressed without their functioning and/or outputs being interpretable. For instance, AI models used in medical diagnosis may produce inaccurate results for certain demographic groups if their training data is biased, and, given that these models often operate as "black boxes," it would be challenging for physicians to understand how decisions are made, making it difficult to detect and address such biases [7].

Turning to AI-enabled CDSS, post-hoc explanations can provide information as per the needs of both physicians and patients, enhancing their understanding and trust in AI-supported medical decisions. [14] If an AI-enabled CDSS offers not only high accuracy but also explainability, which is necessary for building trust in it, it can contribute greatly to physicians' decision-making [8]. Explainability is also key for shared making, which is at the core of patient-centred care, as receiving understandable information on AI-based systems' outputs empowers patients to make informed and autonomous decisions. [6]

2.4 Overview of Risks & Mitigation

Despite its benefits, it is also highlighted that these techniques raise several risks, including misinterpretation due to complex or overly simplified explanations (especially in the case of post-hoc explanations), system exploitation through data exposure, and over-reliance on AI outputs [7, 15, 16]. To mitigate these risks, organisations must tailor explanations to different audiences, safeguard sensitive data and proprietary information, and ensure human oversight remains central to decision-making, especially in high-stakes domains like healthcare. Clear communication about AI limitations and accessible human intervention mechanisms is essential to maintain accountability and protect individual rights. [7]

The human aspect should always be considered, ensuring explanations are relevant and meaningful to people. Individuals perceive and process information differently due to factors like preferences for contrastive explanations, selectiveness, trust in the explanations, and the ability to contextualise them [7]. It is especially important for physicians and patients in the healthcare context, given AI models' potential impact on human lives and integrity and the need for trust in them, while avoiding over-reliance.

2.5 Privacy Risk and Privacy-Preserving Methods

A recent scoping review by Allana et al. [9] examined 57 studies from 2019 to 2024 on the privacy risk and privacy-preserving methods, analysing the privacy-explainability relationship, both key pillars of Trustworthy AI, yet often seen as in conflict. They found that several commonly used XAI techniques raise important privacy risks as they can be exploited in privacy attacks or lead to unintentional privacy leaks. The privacy attacks include membership inference, where attackers determine whether specific data were used during training; model inversion, which involves partial or complete recovery of training or query data from outputs; model extraction, also referred to as "model stealing" where adversaries replicate a model as a first step before other attacks such as membership inference or model inversion; and property inference, where characteristics of the training data, such as overall statistics or aggregated information, may be revealed. On the other hand, unintentional privacy leakages emerge from the training process itself; e.g., overfitting, where models demonstrates better performance on training data which facilitates privacy attacks such as membership inference or attribute inference (a type of model inversion, uncovering specific attributes, often sensitive attributes like gender, age, and race, among others); and memorisation of training data, where the model memorises subsets of training data, at the training process prior to overfitting and may result in data leakage if third-party codes are used to deploy models. Other unintentional leakages can occur directly from the explanations' content, through, for instance, example-based or contrastive explanations, surrogate models, and lack of access control. Correlated fields or proxies in explanations may also result in revealing sensitive fields.

To address these privacy risks, research has been increasingly focusing on exploring privacy-preserving methods, mostly focusing on privacy-preserving machine learning techniques and their applicability to XAI. Techniques such as differential privacy and anonymisation have been explored in depth, while knowledge integration, cryptography, and or perturbation are among the underexplored areas. However, such methods, while preserving privacy, may lead to compromises in accuracy, as well as the quality of explanations. To achieve both privacy and explainability, a combination of federated learning and explainable AI (Fed-XAI) is proposed, as it aims to generate interpretable models while safeguarding local data in distributed settings. While this approach, employing post-hoc explainability methods, is seen as promising, it, in its present state, falls short of ensuring privacy, as generated explanations may inadvertently create entry points for malicious attacks. Moreover, incorporating cryptographic methods into Fed-XAI may introduce additional security challenges. [9].

Allana et al. [9] further stressed the gaps and shortcomings in the research around "privacy-preserving explanation", proposed several characteristics to guide the design of future XAI methods to achieve a triad of explainability, privacy and accuracy, and call for interdisciplinary efforts to develop XAI methods that can ensure both transparency and privacy, given their importance in building trust and ensuring long-term adoption of complex AI tools. Nevertheless, it is underlined that balancing explainability and utility, closely linked to accuracy and performance, while ensuring privacy in XAI stands as a persistent dilemma for developers despite growing interest and efforts in this field; and addressing this is crucial, especially in sensitive domains like healthcare.

Nguyen et al. [17] also emphasise that privacy attacks on explanations may complicate compliance with the GDPR. Unauthorised data leakages may raise challenges to uphold data subject rights, such as rights to access and erasure. While privacy-preserving methods may help avoid such data leakages, they may reduce transparency, and raise challenges for deployers to demonstrate compliance and provide individuals with explanations they can understand, ultimately hindering accountability. Furthermore, finding the right balance between privacy and utility of a system is necessary, as measures that are overly restrictive can undermine the fairness and performance (and accuracy) of AI systems, raising additional legal and ethical concerns.

3 Legal Landscape Surrounding Privacy Risks and Privacy-Preserving Methods

3.1 Overview of Applicable Regulatory Frameworks

There are three main regulations relevant to AI-enabled CDSS in the EU, namely, the General Data Protection Regulation (GDPR), Medical Device Regulation (MDR), and the AI Act.

AI-enabled CDSSs are required to comply with the GDPR at all stages of their design, development, and use in clinical practice, as health data, which is a special category of personal data, is collected and/or processed throughout the lifecycle of these tools.

In addition, these tools are also subject to the EU's Medical Device Regulation, as software tools used for various clinical purposes such as diagnosis, prevention, prediction, prognosis, and treatment of a disease, among others. Under MDR (Rule 11 of Annex VIII on Classification Rules), software used to provide information for diagnostic or therapeutic decision-making is generally classified as Class IIa under EU MDR, unless the decisions it supports could lead to death or irreversible harm (Class III), or serious deterioration of health or surgical intervention (Class IIb). Similarly, software that monitors physiological processes is classified as Class IIa, but if it is used to monitor vital parameters where variations could put a patient in immediate danger, it is classified as Class IIb. Any other medical device software falls under Class I.

AI-enabled CDSS categorised as Class IIa or above are required to undergo a third-party conformity assessment by a notified body. This MDR classification intersects with the EU AI Act, where Article 6 defines certain AI systems as "high-risk" if they form part of products requiring third-party certification under harmonised legislation, such as MDR. As a result, AI-enabled CDSS falling under Class IIa or a higher risk class must comply with both regulations. As a result, most of the AI Act rules, which regulate high-risk AI systems, are also applicable to AI-enabled CDSS tools.

As noted before, there is a limited but growing body of literature on XAI as well as XAI in AI-enabled CDSS, but these studies approach the issue from a rather technical perspective, barely touching upon the compliance matters. The EDPS provides a high-level overview of data protection principles linked to XAI, in particular transparency, fairness, and accountability, but also data minimisation and processing of special categories of data, such as health data. This study notes that XAI techniques, and in particular post-hoc explanations provided for black box models, may offer transparent

insights into AI decisions that aid compliance with these personal data protection principles, but providing explanations does not ensure such compliance [7].

It is also argued, in the literature, that these techniques can be useful in the context of transparency and, in particular, right to explanation, under Article 15 (1) (h) of the GDPR and Article 86 (1) of the AI Act [18].

In order to address the research gap of in-depth analysis of the legal frameworks applicable to post-hoc explanations in AI-enabled CDSS, in particular, the aforementioned dilemma for developers on how to address the balance between explainability, privacy, and accuracy, the following section will map out the provision in the GDPR, AI Act, and MDR that regulate or, may provide guidance on, this trade-off.

3.2 The Trade-off Between Explainability, Privacy, and Accuracy

Recent research, as explained earlier, suggests that there is a trade-off between privacy, explainability and accuracy, which directly impacts the utility of AI-based tools. An overview of the concepts argued to be in tension is as follows:

- **Explainability vs. Accuracy (linked to utility and performance):** It is believed that there is often a conflict between explainability and performance and utility of AI systems. The more complex an AI model is, the more utility and performance it can offer, but the less explainability, and vice versa.
- **Explainability vs. Privacy:** To build transparency and trust in complex AI systems that offer greater performance and accuracy, explanations for AI decisions are provided. However, they present a new surface for privacy attacks, or their content may reveal personal data. On the other hand, privacy-preserving methods may result in a decrease in the quality of explanations.
- **Privacy vs. Accuracy (Utility/Performance):** Various privacy-preserving methods have been proposed to address such privacy risks, but they often come with a cost: decreasing accuracy and diminishing the utility of those tools.

While some argue that the accuracy should be prioritised in medical AI over explainability, as it is of utmost importance to protect patients' lives and health with the best-performing tools, there is no consensus on this matter [12]. Thus, it is necessary to analyse the applicable legal frameworks and determine whether this issue has already been regulated, or whether the lawmaker has made any prioritisation or preference in this trade-off, which may address the developers' dilemma in improving the (privacy-preserving) XAI techniques, or building new ones.

GDPR. Although GDPR does not have any direct, explicit reference to explainability, it contains several provisions that are implicitly relevant to, if not mandating, it. As underlined by the EDPS, XAI can play a vital role in fulfilling key data protection principles such as transparency, fairness, and accountability [7]. The following provisions may offer guidance on how to address the aforementioned trade-off between explainability, privacy and accuracy:

Article 5 lists the principles of personal data processing and requires that personal data be processed transparently and accurately, without prioritising one over another, but counting both as fundamentals of privacy.

Article 15 grants data subjects the right of access to information whether their personal data is being processed and, if so, their personal data as well as certain types of information listed by the article, including, as per Article 15(1)(h), "meaningful information about the logic involved" in the case of (at least, certain types of) automated decision-making regulated under Article 22(1) and (4). Providing meaningful, understandable information to individuals subject to algorithmic decisions requires explainability of such algorithmic decisions.

Article 22(1) enshrines that data subjects have a "right not to be subject to a decision based solely on automated processing, including profiling, which produces legal effects concerning him or her or similarly significantly affects him or her". Medical AI tools can indeed have significant effects on patients and, thus, patients clearly have such a right. It could be discussed whether CDSS can be classified as decision-making tools solely on automated processing. By definition, these tools should support clinical decision making; thus, patients should not be subject to decision-making solely on automated processing. Nevertheless, as mentioned before, the lack of explainability or misleading explainability may result in the over-reliance of healthcare practitioners on such tools, which should be avoided. Gambetti et al. [19] stress that CDSS should, indeed, not replace physicians' decision-making but support it, allowing a more informed patient care, which requires AI decisions to be understandable.

Article 22(3) further stipulates that, in certain cases where Article 22(1) is not applicable, data subjects have "at least the right to obtain human intervention on the part of the controller, to express his or her point of view and to contest the decision". This necessitates a level of explainability that enables both data controllers and data subjects to understand and challenge algorithmic outcomes.

Although recitals are not binding unlike the main text, which is composed of articles, of the regulation, they can be highly instrumental in understanding the intentions of the lawmaker and interpreting the regulation accordingly. In this regard, Recital 71 also highlights that data subjects should have "the right not to be subject to a decision, which may include a measure, evaluating personal aspects relating to him or her which is based solely on automated processing and which produces legal effects concerning him or her or similarly significantly affects him or her" and note that such processing includes, among others, health profiling of an individual. The recital then adds that suitable safeguards should be put in place for such automated processing, including: (i) "specific information to the data subject", of which the content and scope of "specific is not clarified but one could argue that this is a reference to Article 15 (1)(h); and (ii) "the right to obtain human intervention, to express his or her point of view, to obtain an explanation of the decision reached after such assessment and to challenge the decision", which also necessitates a level of explainability of the AI models to the data controller to understand the decision, and to provide an explanation to the data subject about how a decision is made after human intervention. One could argue that such an explanation should provide at least a minimum and understandable information about the algorithmic decision-making reasoning. Taking all these into account, it is clear that

Recital 71, although not bindingly, broadens the scope of Article 22, which grants narrower rights to individuals only in certain cases of automated decision-making.

The second paragraph of Recital 71 lists measures that are essential for a fair and transparent data processing, including ensuring "inaccuracies in personal data are corrected and the risk of errors is minimised", securing personal data, and addressing any unfairness and inequity. As explained previously, AI explainability is closely linked to addressing biases and ensuring fairness, and this recital demonstrates that the lawmaker acknowledges this. The lawmaker also links fairness and transparency with accuracy and data security, which are at the core of the above-explained trade-off between explainability, accuracy, and privacy, given that a significant portion of privacy risks is related to privacy attacks.

Article 25 on data protection by design and by default requires data controllers to put in place appropriate measures from the design stage of a personal data processing activity, during this processing activity, and by default -as the standard practice- to ensure implementation of data protection principles and protection of rights of data subjects. In doing so, data controllers must take into account the state of the art, the particular circumstances related to the processing, and potential risks. Article 25(2) specifically stipulates that, as part of data protection by default obligations, personal data must not be accessible to an indefinite number of people, without the data subjects' intervention. This, as well as the entire article, must be a guiding principle in designing and developing XAI methods and privacy-preserving methods to address to privacy risks involved in XAI.

AI Act. Different from the GDPR, the AI Act has more provisions that are related to explainability, as follows:

Recital 27 has the most explicit reference to AI explainability. This recital refers to seven ethical principles developed by the High-Level Ethics Group for AI (AI HLEG) in their Trustworthy AI Guidelines. As one of these principles, transparency requires AI systems to be developed and used with appropriate traceability and explainability. The recital briefly refers to other ethical principles, including human agency and oversight; technical robustness and safety, which includes accuracy; privacy and data governance; and accountability, which are, as mentioned before, related to XAI or privacy risks arising from XAI.

Turning to the binding main text of the AI Act, Article 13 requires high-risk AI systems to be designed and developed by ensuring sufficient transparency that allows their outputs to be interpretable to their deployers. This article also requires instructions for the use of such high-risk AI systems to be provided, including the following:

- "the characteristics, capabilities and limitations of performance of the high-risk AI system"
- "the human oversight measures referred to in Article 14, including the technical measures put in place to facilitate the interpretation of the outputs of the high-risk AI systems by the deployers"

The former lists several elements, among the following are:

- "where applicable, the technical capabilities and characteristics of the high-risk AI system to provide information that is relevant to explain its output"

- "where applicable, information to enable deployers to interpret the output of the high-risk AI system and use it appropriately"

These demonstrate the importance given by the lawmaker to the explainability of the high-risk AI systems to ensure them being deployed transparently and appropriately, by avoiding their unforeseen effects.

Article 14 stipulates rules on human oversight of high-risk AI systems by requiring those systems to be designed and developed in a way that allows natural persons to effectively oversee them throughout their use. This aims to prevent or minimise the risks, including health risks, for individuals. As the fourth paragraph of the article stipulates, the explainability of high-risk AI systems to deployers is crucial for human oversight. This article is closely related to the GDPR Article 22, but has a more proactive, and, thus, broader approach to ensure human oversight of all high-risk AI systems (not only when a decision is solely based on automated decision making and an individual has already been significantly affected) to prevent or minimise potential risks.

Article 15(1) requires high-risk AI systems to be designed and developed by ensuring an "appropriate level of accuracy, robustness, and cybersecurity". It should be highlighted that there is no benchmark or threshold set by the regulation regarding the appropriate level of accuracy. Instead, the following paragraphs require the EU Commission to encourage the development of such benchmarks as well as measurement methodologies, and the accuracy levels and relevant metrics to be declared in the instructions for use, accompanying the high-risk AI systems.

Article 15(5) builds on 15(1) and stipulates that high-risk AI systems must be resilient against third-party attacks. Technical measures must be tailored to relevant circumstances and risks. Such measures must address AI-specific vulnerabilities, including measures to prevent, detect, and address attacks, confidentiality attacks or model flaws. As explained earlier, the privacy attacks targeting XAI methods also often exploit AI-specific vulnerabilities. Thus, this paragraph is relevant to addressing such privacy attacks and building privacy-preserving XAI methods.

Article 86 enshrines that individuals affected by certain high-risk AI systems, listed in Annex III of the AI Act, must have a right to an explanation about the role of the AI system. As an article resembling GDPR Article 22, this is clearly a requirement for deployers to provide explainability of AI outputs to affected individuals. However, this article has an arguably narrow scope as it does not include those AI systems that are classified as high-risk as a result of being covered by certain EU legislation, including the MDR, and being subject to the requirement of undergoing a third-party conformity assessment. Thus, AI-enabled CDSS are not falling under the scope of this article.

However, Recital 171 is particularly relevant here, as it notes that an individual is significantly affected, such as through an adverse impact on health, by a decision based mainly on the output of a high-risk AI system, should have a right to explanation, in a "clear and meaningful" manner, allowing the individual to use their rights. This recital also reminds Recital 71 and Article 22 of the GDPR, but has a broader scope compared to them, as well as the AI Act Article 86. When compared to the latter, this recital covers all high-risk AI systems, including those that are classified as high-risk as a result of being covered by certain EU legislation, including the MDR, and being subject to the requirement of undergoing a third-party conformity assessment. In other words,

this recital is relevant to AI-enabled CDSS. On the other hand, unlike Recital 71 and Article 22 of the GDPR, the criterion in this recital to have a right to explanation is not being affected by a decision solely based on automated decision-making, but a decision based mainly on the output of a high-risk AI system. As noted before, CDSS are meant to support clinical decision-making but not to overtake it; thus, in many cases, it may not be arguable that a decision involving CDSS will qualify as a decision solely based on automated decision-making falling under the scope of GDPR Article 22 and Recital 71. However, given that CDSS are high-risk AI systems in most, if not all, cases, this recital is likely to be relevant – if the decision is based mainly on the output of CDSS and if the individual faces an adverse impact on their health.

MDR. Turning to the MDR, the sector-specific regulation applicable to the AI-enabled CDSS, there is no explicit reference to explainability or the trade-off between explainability, privacy and accuracy. However, there are a couple of provisions in Annex I on general safety and performance requirements that may offer some useful guidance:

Section 15.1 of Annex I requires diagnostic devices to be "designed and manufactured in such a way as to provide sufficient accuracy, precision and stability for their intended purpose, based on appropriate scientific and technical methods." This emphasises the accuracy of a medical device, such as a CDSS.

According to Section 17.1 of Annex I, medical devices incorporating software, or software as medical devices, including CDSS, must be designed to "ensure repeatability, reliability and performance in line with their intended use", which is not only relevant to performance and utility but can also be seen as referring to high accuracy, given their close interaction. Section 17.2, then, puts forward another requirement of developing and manufacturing such a device according to the state of the art, including information security, which manufacturers, or developers in the AI Act terminology, must follow, including with regard to the need to address privacy attacks on explanations in AI-enabled CDSS.

Section 23.4 (h) of Annex I requires that the information in the instructions for using a device must contain "specifications the user requires to use the device appropriately, e.g. if the device has a measuring function, the degree of accuracy claimed for it". While such specifications may not include the explainability of AI models incorporated into a medical device, this rule suggests that absolute accuracy is not a requirement, and limited accuracy could be acceptable if the users of such devices are informed appropriately.

3.3 Any Legal Clarity For a Right Balance?

The previous section provides an overview of provisions in the GDPR, the AI Act, and the MDR that may offer guidance on how to address the trade-off between explainability, privacy, and accuracy. It has become clear that there are limited references to AI explanations in these regulations, let alone provisions directly related to post-hoc explainability, or the trade-off between privacy, explainability and accuracy. This section analyses whether those mapped out provisions indeed provide any guidance on dealing with this trade-off and finding the right balance.

It should first be highlighted that these three regulations have different scopes, purposes, and approaches. As a result, for instance, MDR has more references to utility, accuracy, and performance; it does not put much emphasis on privacy, while the latter is at the very core of the GDPR. Turning to the AI Act, it is the regulation that has the most references to AI explanations.

Starting with the GDPR, this regulation establishes both accuracy and transparency as the key data protection principles, aiming to protect privacy. While there is no explicit reference to AI explainability, it is linked to the right to access (Article 15) and the right not to be subject to a decision based solely on automated processing (Article 22). However, explainability that is implicitly referred to in these articles has a limited scope, as data subjects will be able to request an explanation on the decision and human intervention only if they are subject to a decision based solely on automated processing, as well as legal or similar significant effects arising from such a decision. Given that using AI-enabled CDSS does not automatically mean that patients will be subject to decisions based solely on automated processing, these articles will not be applicable in most cases. Nevertheless, it is worth highlighting that Recital 71 links fair and transparent processing in such cases of automated processing with taking appropriate measures to ensure accuracy, data security and fairness, among others. On the other hand, Article 25 obliges data controllers to protect privacy both by design and by default, and privacy-preserving methods are crucial in this respect. Taking all these into account, it could be argued that the GDPR prioritises privacy but does not provide any guidance on how to address the aforementioned trade-off.

Turning to the AI Act, it has several provisions on explainability and explanations of AI decisions, particularly in relation to transparency (Article 13) and human oversight (Article 14). These provisions make it clear that explainability is essential to avoid any unintended consequences, including harms, that may arise from the high-risk AI systems. In the case of medical AI tools, such as AI-enabled CDSS, this is of particular importance as unintended consequences may lead to serious harm to health or even death. While those provisions focus on explainability to deployers, there are provisions, namely Article 86 and Recital 171, regarding affected individuals having the right to explanation in certain cases. The differences in the letter of all these provisions indicate that explainability to deployers is seen as more important than explainability to affected individuals. For either case, but especially for explainability to deployers, it should be highlighted that some see post-hoc explanations as unreliable in understanding how an AI model really reaches a decision, and potentially misleading [7].

Article 15 of the AI Act, on the other hand, highlights the importance of accuracy, robustness and cybersecurity. While the "appropriate" level of accuracy is not specified in the article, it requires the level of accuracy to be declared in the instructions for use, implying potential limitations to accuracy. The article, however, has a stricter approach on requiring technical measures against third-party attacks, where the privacy-preserving XAI techniques are meant to be instrumental.

As noted earlier, Recital 27 of the AI Act refers to the Trustworthy AI Guidelines of the AI HLEG and its key principles. Those principles point out the crucial role of privacy, accuracy, and explainability, which are either established as principles themselves or fundamental components of some other principles. The guidelines stress that

there may sometimes be tensions between these principles, which, as “abstract ethical prescriptions”, cannot offer a solution, and such trade-offs and tensions should be approached with “reasoned, evidence-based reflection rather than intuition or random discretion” [20]. Thus, AI HLEG acknowledges that these principles do not always work in harmony, and each tension should be dealt with according to the particularities of each case.

On the other hand, despite not being related to explainability, Article 10(5) of the AI Act may provide useful guidance. This provision, arguably, indicates lawmakers’ prioritisation of fairness over privacy as it allows processing of special categories of data when strictly necessary to ensure bias detection and correction, thus ensuring fairness. However, this is possible only under certain conditions, including the use of state-of-the-art security and privacy-preserving measures, and ensuring access control and avoiding any unauthorised access. One might argue to extend this provision to AI explainability, given the close relationship between fairness and explainability, potentially only as long as such explainability is necessary to address fairness. However, it is not easy to draw the line when explainability is necessary to address fairness or not. It should also be highlighted that even when privacy may be compromised for the sake of fairness, taking security and privacy-preserving measures is necessary. Thus, it can be argued that even if the fairness-over-privacy assumption is extendable to explainability-over-privacy as a result of the fairness-explainability link, taking security and privacy-preserving measures will be necessary.

In the case of AI-enabled CDSS, the MDR also comes into play besides the GDPR and AI Act, as a sector-specific legislation regulating medical devices. This regulation puts emphasis on performance, utility, and accuracy, which are interconnected. However, it also implies, Section 23.4 (h) of Annex I, that accuracy is not an absolute requirement as long as users are informed. This could be the basis of an argument that in the trade-off between explainability, privacy, and accuracy, limitations to accuracy could be acceptable.

Overall, the GDPR, the AI Act, and the MDR do not seem to provide clear guidance on how to address the trade-off between privacy, explainability and accuracy. However, it can be argued that the regulations do not expect complete accuracy, as limitations to it might be acceptable. However, the explainability of AI decisions is crucial for accountability, transparency and human oversight, under the GDPR and, especially, the AI Act. Likewise, privacy and security of personal data are also deemed highly important given the objectives of the GDPR and its articles, such as Article 25 on data protection by design and by default, and the emphasis put on privacy-preserving and other cybersecurity techniques by the AI Act. Taking all this into account, further research on privacy-preserving XAI techniques is deemed necessary, despite the potential diminishing of accuracy, but ultimately achieving the triad of explainability, privacy and accuracy, especially for AI-enabled CDSS. Given that they often utilise complex models with the aim of surpassing human-level accuracy and performance, post-hoc explanations are likely to be used, which must be developed by ensuring reliability.

4 Conclusion

This paper provides an overview of the applicable legal frameworks to the conflict between explainability, privacy, and accuracy related to the privacy risks arising from XAI, with a focus on the post-hoc explainability techniques in AI-enabled CDSS. It concludes that the current legal frameworks do not offer much clarity on how to find a balance between the three concepts, which are all deemed crucial for trust in these tools, as well as their uptake and wide-scale use.

Acknowledgments. The I3LUNG project has received funding from the European Union's Horizon Europe call "HORIZON-2021-HLTH-05-02 - Data-driven decision support tools for better healthcare delivery and policymaking with a focus on cancer", under Grant Agreement number 101057695.

Disclosure of Interests. The author has no competing interests.

References

1. Sutton, R.T., Pincock, D., Baumgart, D.C., Sadowski, D.C., Fedorak, R.N., Kroeker, K.I.: An overview of clinical decision support systems: benefits, risks, and strategies for success. NPJ Digit. Med. 3, 17 (2020). https://doi.org/10.1038/s41746-020-0221-y
2. Elhaddad, M., Hamam, S.: AI-Driven Clinical Decision Support Systems: An Ongoing Pursuit of Potential. Cureus 16(4), e57728 (2024). https://doi.org/10.7759/cureus.57728
3. Magrabi, F. et al.: Artificial Intelligence in Clinical Decision Support: Challenges for Evaluating AI and Practical Implications. Yearb. Med. Inform. 28(1), 128–134 (2019). https://doi.org/10.1055/s-0039-1677903
4. Ouanes, K., Farhah, N.: Effectiveness of Artificial Intelligence (AI) in Clinical Decision Support Systems and Care Delivery. J. Med. Syst. 48(1), 74 (2024). https://doi.org/10.1007/s10916-024-02098-4
5. Panigutti, C. et al..: The role of explainable AI in the context of the AI Act. In: Proceedings of the 2023 ACM Conference on Fairness, Accountability, and Transparency (FAccT '23), pp. 1139–1150. ACM, New York (2023). https://doi.org/10.1145/3593013.3594069
6. Amann, J., Blasimme, A., Vayena, E., Frey, D., Madai, V.I., Precise4Q consortium: Explainability for artificial intelligence in healthcare: a multidisciplinary perspective. BMC Med. Inform. Decis. Mak. 20(1), 310 (2020). https://doi.org/10.1186/s12911-020-01332-6
7. European Data Protection Supervisor (EDPS): TechDispatch #2/2023 – Explainable Artificial Intelligence. EDPS, Brussels (2023). https://www.edps.europa.eu/data-protection/our-work/publications/techdispatch/2023-11-16-techdispatch-22023-explainable-artificial-intelligence_en
8. Kostopoulos, G., Davrazos, G., Kotsiantis, S.: Explainable Artificial Intelligence-Based Decision Support Systems: A Recent Review. Electronics 13(14), 2842 (2024). https://doi.org/10.3390/electronics13142842
9. Allana, S., Kankanhalli, M., Dara, R.: Privacy Risks and Preservation Methods in Explainable Artificial Intelligence: A Scoping Review. arXiv preprint arXiv:2505.02828 (2025). https://doi.org/10.48550/arXiv.2505.02828
10. Gunning, D., Aha, D,.: DARPA’s Explainable Artificial Intelligence Program AI Mag. 40(2), 44–58 (2019). https://doi.org/10.1609%2Faimag.v40i2.2850

11. Yang, G., Ye, Q., Xia, J.: Unbox the black-box for the medical explainable AI via multi-modal and multi-centre data fusion: A mini-review, two showcases and beyond. Inf. Fusion 77, 29–52 (2022). https://doi.org/10.1016/j.inffus.2021.07.016
12. Antoniadi, A.M. et al.: Current challenges and future opportunities for XAI in machine learning-based clinical decision support systems: A systematic review. Appl. Sci. 11(11), 5088 (2021). https://doi.org/10.3390/app11115088
13. Campos, F., Petrychenko, L., Teixeira, L.F., Silva, W.: Latent diffusion models for privacy-preserving medical case-based explanations. In: Zaza, G., Casalino, G., Castellano, G. (eds.) Proc. 1st Workshop on Explainable Artificial Intelligence for the Medical Domain (EXPLIMED 2024) co-located with ECAI 2024. CEUR Workshop Proc. 3831, Santiago de Compostela, Spain, 20 Oct 2024. https://ceur-ws.org/Vol-3831/
14. Gerdes, A.: The role of explainability in AI-supported medical decision-making. Discov. Artif. Intell. 4, 29 (2024). https://doi.org/10.1007/s44163-024-00119-2
15. Slack, D., Hilgard, S., Jia, E., Singh, S., Lakkaraju, H.: Fooling LIME and SHAP: Adversarial attacks on post hoc explanation methods. In: Proc. AAAI/ACM Conf. on AI, Ethics, and Society (AIES 2020), pp. 180–186. ACM, New York (2020). https://doi.org/10.1145/3375627.3375830
16. Lakkaraju, H., Bastani, O.: "How do I fool you?": Manipulating user trust via misleading black box explanations. In: Proc. AAAI/ACM Conf. on AI, Ethics, and Society (AIES 2020), pp. 19–25. ACM, New York (2020). https://doi.org/10.1145/3375627.3375833
17. Nguyen, T.T. et al.: Privacy-preserving explainable AI: A survey. Sci. China Inf. Sci. 68, 111101 (2025). https://doi.org/10.1007/s11432-024-4123-4
18. Juliussen, B.A.: The right to an explanation under the GDPR and the AI Act. In: Ide, I., et al. (eds.) MultiMedia Modeling. MMM 2025. Lecture Notes in Computer Science, vol. 15523, pp. 184–197. Springer, Singapore (2025). https://doi.org/10.1007/978-981-96-2071-5_14
19. Gambetti, A., Han, Q., Shen, H., Soares, C.: A survey on human-centered evaluation of explainable AI methods in clinical decision support systems. arXiv preprint arXiv:2502.09849 (2025). https://doi.org/10.48550/arXiv.2502.09849
20. High-Level Expert Group on AI: Ethics guidelines for trustworthy AI. European Commission, Brussels (2019). https://digital-strategy.ec.europa.eu/en/library/ethics-guidelines-trustworthy-ai