Aligning Explainable AI and the Law: The European Perspective

BALINT GYEVNAR and NICK FERGUSON, School of Informatics, University of Edinburgh, UK

The European Union has proposed the Artificial Intelligence Act intending to regulate AI systems, especially those used in high-risk, safety-critical applications such as healthcare. Among the Act's articles are detailed requirements for transparency and explainability. The field of explainable AI (XAI) offers technologies that could address many of these requirements. However, there are significant differences between the solutions offered by XAI and the requirements of the AI Act, for instance, the lack of an explicit definition of transparency. We argue that collaboration is essential between lawyers and XAI researchers to address these differences. To establish common ground, we give an overview of XAI and its legal relevance followed by a reading of the transparency and explainability requirements of the AI Act and the related General Data Protection Regulation (GDPR). We then discuss four main topics where the differences could induce issues. Specifically, the legal status of XAI, the lack of a definition of transparency, issues around conformity assessments, and the use of XAI for dataset-related transparency. We hope that increased clarity will promote interdisciplinary research between the law and XAI and support the creation of a sustainable regulation that fosters responsible innovation.

CCS Concepts: • **Social and professional topics** → **Governmental regulations**; • **Applied computing** → **Law**; • **Computing methodologies** → *Artificial intelligence*.

Additional Key Words and Phrases: explainable AI, AI regulation, AI and the Law, European Union, transparency

ACM Reference Format:

Balint Gyevnar and Nick Ferguson. 2023. Aligning Explainable AI and the Law: The European Perspective. In . ACM, New York, NY, USA, 17 pages. https://doi.org/XXXXXXX.XXXXXXX

1 INTRODUCTION

Artificial intelligence (AI) systems are an increasing presence in twenty-first-century life. AI systems are readily available to everyone from commercial users to non-technical enthusiasts and can outperform humans across a variety of tasks. OpenAI's ChatGPT system is hailed by some as an internet search revolution [45]; Deepmind's AlphaFold has arguably solved the protein folding problem [64]; and generative art from stability.ai's Stable Diffusion offers to generate stunning art with just a few lines of text [56]. Public-facing demonstrations of the capability of AI may lead to the perception that AI is a nascent field, but AI systems are already ubiquitous. Facial recognition, recommender systems, and medical diagnoses are applications which have been utilising AI for many years.

However, as we see AI increasingly deployed in safety-critical applications, unexpected issues invariably arise. Such infringements are well-documented: biased and discriminatory decisions are inherent to data-driven systems [3, 7]; mass surveillance is a threat worsened by AI-based biometric identification [37]; and subversive manipulation of groups is easier than ever [59]. If our rights are impeded by AI systems, how can we recognise a system's effects and contest their decisions, if often not even their developers can understand and explain how they work?

Recognising the urgent need to address concerns around AI, regulators are drawing up proposals to tackle the challenges of its adoption. In the USA, the White House has introduced the "Blueprint for an AI Bill of Rights", proposing

Manuscript submitted to ACM

five main areas of regulation for AI [32]. Brazil's Congress has passed a bill establishing a legal framework for artificial intelligence [5]. The UK has outlined its own pro-innovation approach to AI regulation and is working on a "National AI Strategy" [17]. In the EU, a sweeping new regulation, the Artificial Intelligence Act, is under consideration [13]. The Act places requirements on AI *providers* (i.e., developers of AI systems) relating to transparency and explainability. It takes a *risk-based* approach to defining these requirements, and proposes strong conformity assessment conditions that providers must fulfil. As such, it is the main legal source through which we ground our discussion in this work.

Scientists also acknowledge the need for more transparent and trustworthy AI [30, 35] that respect the principles of law [18] and ethics [27], and are also robust in the real-world [16]. Explanations are regarded by scholars as an essential part of human-AI interactions, and works now focus on achieving human-aligned AI via explanations [15, 40, 58, 62].

However, the legal requirements of AI and the technological methods for satisfying those requirements are not always well aligned. This can have serious consequences. On one hand, how can legal certainty be guaranteed if the law defines AI itself one way while the scientific community agrees on a different definition? On the other hand, how could scientists pursue innovation if legislation is too restrictive due to incorrect or outdated definitions? Given the ludicrous speed of new research, how can we guarantee lasting regulations without a unified conceptual understanding?

Take *transparency* for example: in the AI Act, it is an idea that permeates the entire ecosystem of the AI product. However, in the field of *explainable AI* (XAI), transparency has a concrete meaning, describing the properties of an AI algorithm and the explanations one can generate about its reasoning. While the two definitions are passable on their own, when they interact in the body of AI regulation, this misalignment becomes dangerous. In the legal interpretation, XAI would need to satisfy unreasonable expectations, while in the scientific context, overreaching algorithmic transparency could damage the cybersecurity of the AI system and the intellectual property rights of its provider.

In this paper, we urge that now is the time for scientists and lawyers to consolidate understanding, so that legislation and scientific research are brought into alignment. With this, can we begin to support legal certainty and the longevity of regulations, build public trust, and foster lawful investment and research into AI.

1.1 Scope

From the legal perspective, we focus on the widely discussed proposal of the EU: the 2021 final EU Commission proposal of the AI Act (hereafter *the Act*) [13]. In addition, we discuss the closely related General Data Protection Regulation (GDPR) [48]. From a technological perspective, we focus on explainable AI (XAI). The broad aims of this paper are to understand the transparency and explainability requirements of the Act; to discuss the similarities and differences between the Act's requirements and the properties of explainable AI; and to highlight potential issues that differences between the fields could induce. By focusing on the Act, our argument is grounded in tangible legislation. Moreover, due to the "Brussels Effect" and the partial extraterritorial reach of the Act, we expect many upcoming frameworks to take inspiration from it, similar to the one observed in Data Protection law and the GDPR. This makes our arguments potentially relevant not just within the jurisdiction of the EU, but to a more global audience.

1.2 Important Findings

Our contributions are three-fold. We give an overview of the field of XAI, identifying important XAI-specific concepts, such as transparency and explainability, and discuss their legal relevance. We then give a detailed reading of the transparency and explainability requirements of the Act. Finally, we suggest four topics that we believe need more attention from lawyers and XAI researchers alike to advance the state of the art in AI regulation and trustworthy XAI:

(1) **Clarify the legal status of XAI (Section 4.1)**. XAI methods themselves are frequently based on methods which fit the legal definition of AI. If an AI system makes use of XAI, then should the two systems be considered separately or as one unit for the assessment of conformity or liability?
(2) **Clearly define *transparency* (Section 4.2)**. A functional but implicit meaning of transparency can be derived through regulatory requirements. However, to increase legal certainty, inform the design of XAI systems, and protect the cybersecurity of AI systems from overreaching transparency, notions such as transparency, interpretability, and explainability should be explicitly defined and scoped.
(3) **Focus on XAI and conformity assessment (Section 4.3)**. Conformity assessment is essential to achieve compliance with regulation. However, the Act leaves open the possibility of self-regulation by providers. XAI tools for externally auditing AI systems should be designed to support the enforcement of the regulation.
(4) **Explainable data (Section 4.4)**. The Act defines strong requirements on data quality and governance in terms of, among others, design choices, preprocessing, and bias. XAI research should extend to explaining the effects of changes in these properties on the functioning of the AI system.

1.3 Structure

In Section 2, we overview relevant concepts and approaches in XAI literature and identify their legal relevance. In Section 3, we situate the importance of transparency and explainability in the legal context of the GDPR and the Act, identifying shared concepts and differences between XAI and the law. We follow this in Section 4 with an identification of legal and technological issues, and discuss potential solutions that could further inform the discourse between legal scholars and XAI researchers. In Section 5, we finish by discussing additional interesting legal and scientific challenges.

1.4 Illustrative example

We will use a running example throughout the paper to illustrate the concepts discussed and issues raised. The healthcare industry has long been a test bed for novel AI systems. Technology and applications in this domain have evolved over the years, and include the use of expert systems for diagnosing a variety of conditions [43]; machine learning for drug discovery [38, 65]; neural networks for cancer diagnosis[20]; and even embodied AI in the form of robots for social care[6, 60]. Most software systems, whether involving AI or not, that serve a medical diagnostic function are already regulated [49, 50]. The precise relation between the AI Act and these regulations is likely to lead to complicated legal issues, for example possible double-regulation or regulatory inconsistencies. For this paper, we assume that the AI Act creates new transparency duties for medical devices, which will however be enforced, in most cases, by the existing Notified Bodies that certify medical devices. In this work, we consider a hypothetical medical diagnosis system:

Medical diagnosis example Our system uses X-ray image data, the patient's medical history, and other personal data to predict a diagnosis of a medical problem, such as the existence of a tumour. The system predicts the likelihood of such an event, and provides a localised area on the X-ray in which the tumour is believed to exist.

2 TRANSPARENCY AND ARTIFICIAL INTELLIGENCE

Under the Act, AI providers are required to deliver transparent systems. Parts of the requirements can be satisfied through documentation. However, some AI systems, particularly "black-boxes" such as large language models, go beyond intuitive human understanding, even on questions of high-level design choices. For example, what justifies a particular setting of hyperparameters of neural networks, and how do we explain their effects on the output?

Therefore, the use of explainable AI – often called an "interpretation tool" or "algorithmic scrutiny" in legal contexts, seems unavoidable. We will devote this section to a brief overview of the field of XAI, focusing on relevant terminology and high-level concepts rather than on the workings of any specific method.

2.1 What are Explanations?

Explanations are an essential part of human cognition. To build a cross-disciplinary understanding of what an explanation is, we can start with Miller's review [40]:

> *An explanation is an answer to a* why*-question. It is both a process and a product that relies on cognitive and social processes to attribute causality to observations.*

While some explanations answer *how* and *what*-questions, which require different types of reasoning, in XAI literature, these are considered less complex to answer [40], so we do not consider them here.

Explanations for *why*-questions describe events in terms of their cause-and-effect relationships. These explanations require cognitive reasoning (e.g., experimentation with 'what-if' scenarios and abductive reasoning) to infer a causal chain and select the most relevant causes. Explanations are also social because they need to be communicated. This dual aspect of the explanation process is important to keep in mind, as the perceptions of what is a correct (cognitive process) and clear (social process) explanation will affect the interpretation of legal requirements on explainability.

The word "explanation" has two meanings, referring to either a process–the cognitive and social action of communicating and understanding the cause of an event; or a product–the resulting natural language utterance or graphical representation which contains this information. Regulation will need to properly define the scope of explainability and how it extends to both the process of generating an explanation and to the properties of the generated explanation as a product. Failing to do so would result in too lax or stringent regulations that demand inappropriate explanations.

Explanations are tied to observations. However, what people observe depends greatly on what their stakes are in the AI system. For example, a doctor analysing an X-Ray does not have access to the raw weights of the AI system, unlike the AI provider. Importantly, the generated explanations must consider what observations are available and relevant to the stakeholders. When designing legal requirements and XAI systems for achieving transparency we must, therefore, carefully balance the expectations of different stakeholders.

2.2 Explainable AI

Before discussing XAI, let us give a brief overview of the two prominent AI paradigms. The first AI systems were based on symbolic logic. In this paradigm, world knowledge is represented using mathematical symbols. The first commercially viable AI, known as *expert systems*, were built on this paradigm [57]. The representation of knowledge in this manner lends an inherent interpretability in the form of a causal chain of reasoning.

On the other end of the spectrum, deep learning models that dominate today are not built on such tangible representations of data. These *neural network*-based models consist of many millions of parameters. The values of these parameters are *learned*, requiring vast amounts of data, and serve to mathematically transform the input data to an output prediction or classification. Fundamentally, this family of approaches lacks intrinsic interpretability due to the built-in parameterisation and abstraction. However, such models are highly performant and extremely widespread.

Recognising that a lack of interpretability diminishes the trust in AI systems, methods that explain AI model outputs were developed, forming the field of XAI. Most surveys of the field [8, 24, 41, 63] define XAI as a self-explanatory system that reveals the reasoning behind its outputs. In contrast, we give a practical definition in terms of four attributes to help

regulators pin down what XAI should be. In taking an overview of XAI, we will deconstruct this definition and address the four attributes separately, while connecting concepts from XAI with relevant legal considerations. Explainable AI:

(1) explains the output of an AI system;
(2) using partially or fully automated methods;
(3) to clearly defined stakeholders;
(4) in a relevant and accurate manner.

2.2.1 Explaining systems' outputs. The goal of XAI is to explain the output of AI systems. Yet, different AI systems have different inherent properties that make their decisions more or less suited for explanations. There is a significant amount of debate in XAI regarding the meanings of *interpretability*, *explainability*, and *transparency*, and their relationships. It is important to clarify these terms as they are not interchangeable, and their interpretations lead to different understandings of the responsibilities and capabilities of AI systems. Issues arising from this are discussed in Section 4.2.

On the one hand, Miller [40] equates interpretability with explainability. However, others argue, and we support this view, that interpretability is considered an inherent property of an AI system [23, 24, 41]. According to Molnar [42], an interpretable AI system is inherently human-understandable due to its low complexity. These systems can support human understanding even though they may not provide explanations by design. For example, a simple linear regression function used to predict the future value of the population of a country based on historical data is interpretable; but a multi-layered deep neural network used to identify different animals in images can not.

On the other hand, explainability is the property of any AI system whose output can be explained to support human understanding. It is the ability of an AI system to relevantly communicate the reasoning behind its decision process. Returning to the example from the previous paragraph, a linear regressor on its own is not explainable: showing model weights to a layperson will likely mean nothing to them. However, matched with a suitable XAI technique a linear regressor can intelligibly communicate the relevant weights affecting its decision and thus becomes explainable.

2.2.2 Methods of XAI. The number of approaches has exploded in recent years. To complicate matters, many XAI methods are themselves complex AI systems [8]—at least as far as the definition of an AI system in the Act is concerned. Current methods include, among others, analysis of feature importance [55], saliency maps [52], counterfactual methods [25], recognising textual entailment [39], and knowledge distillation [12].

Researchers often categorise AI systems based on their opacity, contrasting non-interpretable black-box systems such as neural networks against interpretable white-box systems such as decision trees [8]. This property of an AI system has a crucial influence on the design choices of the XAI method: using white-box models, we can more easily guarantee correctness, verifiability, and causality. However, this usually comes at the cost of expressiveness and accuracy [10]. In regulation, one type of system may be preferred over the other, however, careful balancing is needed so that interpretability and accuracy are present at the desired levels.

Aside from the technical methods for generating explanations, Burkart and Huber [8] consider two views on how they are generated: *ante-hoc* explanations are generated from the internal representations of white-box systems, while *post-hoc* explanations are inferred from an output after a decision was made. Thus, ante-hoc explanations correct by design, but can only be generated for white-box systems. Post-hoc explanations may distort the causality underlying the model's decision process and require more effort to generate, but they are applicable to white- and black-box systems.

An XAI system can also be *model-agnostic*, meaning that it can be applied to explain many AI algorithms, or it can be *model-specific*, meaning that it applies to one specific AI algorithm. While model-agnostic XAI offers off-the-shelf explainability for AI systems and, thus, can provide significant savings in resources, it raises issues of liability when the system is not sufficiently tested. For example, if a widely available model-agnostic XAI method is used to erroneously explain the decision of a high-risk AI system and the user acts incorrectly based on this explanation, then to what extent can the developers of the general-purpose XAI method be held liable? Is the due diligence of the AI provider called into question due to their selection of an unsuitable XAI system?

2.2.3 Stakeholders. We must also consider for whom we are creating explanations. Recently, Langer et al. [36] have given a taxonomy of how XAI can address the needs of different stakeholders. They consider a feedback loop between the explanation process and the stakeholders based on the understanding and satisfaction of four main stakeholder groups: developers, users, deployers, and affected parties. Their categorisation of stakeholders also aligns with the definitions of the Act, which defines similar groups.

Additionally, Mohseni et al. [41] suggest three distinct end-users for XAI: *AI novices*, *data experts*, and *AI experts*. In this work, the first category is of utmost importance, as it relates to end-users with a negligible amount of expertise on how the system works. Their concerns may include, among others, bias, privacy, and trust, which are the very issues that regulators are addressing in the Act. Multi-modal explanations, natural language communication, conversational agents, and cognitive modelling are some of the tools that are popular for addressing concerns of AI novices [15, 30, 40, 53]. Much theoretical and practical progress has been made in developing social XAI that addresses these concerns, but additional stakeholder-focused interdisciplinary research is sorely needed.

2.2.4 Accuracy and Relevancy. Finally, the question of evaluating the quality of generated explanations remains. Here, we must specify desired characteristics of performance metrics, which will depend on the various stakeholders. For example, for AI novices, measures of trust and intelligibility will be essential, while for regulators, we might expect objective correctness to be more important. There has been work on creating metrics for explanation performance evaluation [31], however, the design of metrics that clearly show compliance with regulation is a new challenge [61].

Metrics are also essential as different XAI methods applied to the same data can output very different explanations [54]. This means it is difficult to establish trustworthy baselines unless we define and measure clearly what aspects of explanations are important, and for whom. Comparisons to baselines are essential for demonstrating the abilities of any XAI system and regulatory requirements on explainability may well demand such comparisons to show conformity.

Medical diagnosis example Given our understanding of explanations, we consider natural language explanations which we might like our medical diagnosis system to deliver. Explanations are answers to *why*-questions, such as '*Why do you think I don't have a tumour?*' Possible explanations may refer to feature importance: '*The X-ray does not exhibit any of the relevant characteristics, and patients in your age group are at very low risk*', or a contrastive approach: '*Because your X-ray shows characteristics closer to those produced by pneumonia.*'

3 TRANSPARENCY AND THE LAW

We examine the history of how explanations took centre stage in a legal debate around the GDPR, how elements of this debate informed the Act, and the resulting transparency and explainability requirements for AI. Equipped with our summary of XAI, this section reflects on the common themes shared between regulations and XAI.

3.1 A Right to Explanation?

First, we examine the recent legal history of the debate surrounding the *"right to explanation"* in the landmark data protection act of the EU, the General Data Protection Regulation (GDPR) [48]. It is important to devote some space to address this debate because XAI researchers have continually used the "right to explanation" to justify their motivations (e.g., in [1, 24, 31, 63]). However, **XAI practitioners should remember that a "right to explanation" under the GDPR is not a legal norm, and so far no case law exists on its interpretation in this aspect.** While the GDPR is specifically concerned with the processing of personal data, the most questionable practices of AI use precisely this kind of data. Therefore, researchers also need to consider its provisions proactively.

Before we delve deeper into details, we have to note that, unlike the field of XAI, legal terminology differentiates between *ex-ante* and *ex-post* explainability [67] – not to be confused with ante-hoc and post-hoc explanations. Ex-ante explainability concerns explanations about the overall usage, architecture, and purpose of a system, and is created prior to the running of the system. On the other hand, ex-post explanations are created after an automated decision is made, and aims to describe both the input-specific reasoning behind the decision and the overall functionality of the system.

Goodman and Flaxman [22] first suggested that one can derive a *"right to explanation"* from Article 22(3) and Articles 13–15 GDPR, whereby a data subject has the right to "express his or her point of view and to contest the decision" which is "based solely on automated processing", and to obtain "meaningful information about the logic involved" in the processing of personal data. This requirement for an explanation also appears explicitly in the non-binding Recital 71.[1] Furthermore, Winikoff and Sardelić [68] suggested a "right to explanation" could be derived from human rights in specific cases, for example, discrimination due to machine bias, which is a recurring issue of automated profiling.

3.1.1 Issues with the Right to Explanation. However, both the existence and the utility of such a "right to explanation" has been called into question by Wachter et al. [67]. Importantly, they argue that there are both legal and functional issues with such a right, and the intentionally vague phrasing of the GDPR makes the interpretation of Article 22 challenging. For example, it is unclear whether the GDPR requires an ex-ante or an ex-post explanation.

Doshi-Velez et al. [18] give further substance to this challenge, as they note that "explanations are not free" and "generating them takes time and effort". Calling for a universal "right to explanation" would therefore place an unreasonable burden on AI providers. This notion was also acknowledged by the former data protection advisory board of the EU, Article 29 Data Protection Working Party, stating that explanations under the GPDR need not be "complex mathematical explanation[s] about how algorithms or machine-learning work", but rather they should be "clear and comprehensive ways to deliver information to the data subject" [51].

The impracticality of a "right to explanation" prompts Edwards and Veale [19] to warn against the "illusion of a legal right". Instead, they highlight several actionable legal routes to ensure a "right to better decisions", such as the data protection impact assessment. The arguments for the "illusion of a legal right" are further supported by Bayamlıoğlu [2] who emphasise that it is a "right to contest" in Article 22(3) that should drive the discussion around tangible ways to achieve transparency. Relatedly, Jongepier and Keymolen [34] argue that, regardless of the involvement of machines, explanations should be a moral right if "choices are made which significantly affect us but which we do not understand".

We can then say with confidence that relying on the GDPR to serve as the source of regulatory transparency and explainability is not enough. Yet, all who have engaged in this debate acknowledge that explanations are crucial to transparency and explainability. The fact that the GDPR conceivably contains a "right to explanation" is an indication

[1] Recitals are part of the preamble to a treaty that "articulate shared assumptions, goals and explanations concerning the treaty." [28].

of the directions future jurisprudence on automated processing could take. Indeed, recognising the inherent dangers of AI systems used for automated data processing is a core motivation of the AI Act. As we will see later, transparency and explainability become concrete requirements in the Act that are not up to the minutiae of legal interpretations.

3.2 Explainability and the AI Act

Having seen the legal uncertainty around explanations in the GDPR, in this section we discuss how the AI Act clarifies and operationalises transparency for AI providers and the effects of the proposal on the legal requirements for explainability. Since its first appearance, the Act has undergone significant changes in the EU Parliament and Council, so for consistency with previous works on the Act, we refer to the initial proposal by the EU Commission released on 21 April 2021 [13]. However, where appropriate, we will mention relevant amendments to the original proposal.

The AI Act is an ambitious proposal for an EU-wide regulation presenting a sweeping set of rules aimed at harmonising and standardising compliance requirements for AI systems.[2] It takes a risk-based approach to defining the transparency requirements of different AI systems. *High-risk* systems, such as facial recognition and law enforcement systems[3] are subject to greater regulation than low-risk ones.

Article 1(c) declares that the Act lays down "harmonised *transparency rules* for AI systems intended to interact with natural persons [...]" (emphasis added). The concept of transparency is explicitly addressed in the Act, and as we will see, this has wide-ranging implications on the explainability requirements of high-risk AI systems.

In the following sections, we describe the transparency and explainability requirements of the Act building on the work of Sovrano et al. [61]. We review their discussion of explainability requirements, expand their reading with further requirements, and give an updated view that includes recent revisions of the Act. The Act (implicitly) distinguishes between *user-empowering* and *compliance-oriented* transparency.

3.2.1 User-Empowering Transparency. Article 13(1) addresses user-empowering transparency directly, stating that "high-risk AI systems shall be designed and developed in such a way to ensure that their operation is sufficiently transparent". This is to "enable users to *interpret* the system's output sufficiently" (emphasis added), not just to facilitate the correct use of the system. The first concrete user-empowering requirement for explanations appears in Article 13(2) which introduces an ex-ante explainability requirement in the form of instructions for use that are "concise, complete, correct, and clear". That this is an explainability requirement is confirmed in Article 13(3)(b) which requires the instructions for use to contain relevant information as regards "the characteristics, capabilities, and limitations of performance of the high-risk AI system". Recital 47 also makes it clear, that these requirements are essential for the cases where high-risk "AI systems [are] incomprehensible or too complex for natural persons".

In addition, *human oversight* is a core element of the Act, which creates further explainability requirements. Article 14(1) stipulates that "high-risk AI systems shall be designed and developed in such a way [...], that they can be effectively overseen by natural persons". According to Article 14(4), these measures must enable people to "fully understand the capacities and limitations", and to "correctly interpret the high-risk AI's output". Even more importantly, Article 14(4)(c) explicitly addresses, among others, explainable AI which it refers to as "interpretation tools and methods". Therefore, these paragraphs seem to place ex-post explainability requirements on high-risk AI systems. Recitals 38–40 mention the explicit cases of law enforcement, migration, and administration of justice, where such human oversight will definitely need to be ascertained, as biased decisions have particularly far-reaching effects in these applications.

[2]We cannot give a complete overview at the AI Act here. The inclined reader should refer to Veale and Borgesius [66] for a critical reading of the Act.
[3]See Annex II and III of the Act for a full list of systems which are classified as high-risk.

3.2.2 Compliance-Oriented Transparency. The Act also places strong requirements on compliance-oriented transparency. In particular, Articles 9 and 17 establish the requirements for a risk-management and quality-management system. These systems place detailed transparency requirements, achieved through documentation, monitoring, and verification, on the providers of high-risk AI systems to guarantee compliance with the Regulation. Article 11 expands on the *technical documentation* requirements of providers, which need to be drawn up before the high-risk AI system is placed on the market. Referring to Annex IV(2)(b)-(d), Article 11 requires that such documentation includes "the general logic of the AI system and of the algorithms; the key design choices [...]; [and] the main classification choices". These are clear requirements on the ex-ante explainability of high-risk AI.

As Article 29(4) states, "users shall monitor the operation of a high-risk AI system on the basis of the instructions for use". This means that compliance-oriented transparency requirements need to enable users to monitor the operation of the high-risk AI system, forming a crucial interaction of user-empowering and compliance-oriented transparency requirements [61]. Additionally, Article 12 requires *record-keeping*, or logging, of the high-risk AI system's operation. Relatedly, the most recent version of the Act (at the time of writing) from the Czech presidency of the EU Council adds Article 13(3)(f) requiring an ex-post "description mechanism [...] that allows users to properly collect, store, and interpret the logs" [14]. Finally, Recital 58 states that responsibility has to extend to the users to maintain the correct operation of high-risk AI systems. Therefore, an accurate and relevant explanation of the system is essential, otherwise, the user would not be capable of handling the system correctly.

3.2.3 Criticisms of the Act. The transparency and explainability requirements of the Act are not without criticism [66]. Many of the hundreds of amendments to the Commission's proposal highlight a tension between intellectual property (IP) rights and transparency, incorporating several exclusion criteria to the documentation and transparency requirements of the Act, which are in addition to the confidentiality requirements of Article 70 [46, 47]. Interestingly, however, the Czech presidency's version of the Regulation does not incorporate these clauses, highlighting the Council's increased emphasis on transparency rights over IP rights or trade secrets.

An additional criticism is that the Act is overreaching in its requirements for compliance-oriented transparency. A further amendment to Article 11 allows for small and medium enterprises and start-ups, which have very different resources available to them, to fulfil the requirements of Annex IV in equivalent (but possibly less demanding) forms of documentation [14]. Overreaching and infeasible technological requirements are recurring criticisms of the Act and also form a core part of our discussions in the next section. Importantly, both lawyers and scientists should monitor the up-to-date draft of the Act in case they intend to work with the proposal.

4 DIFFERENCES BETWEEN XAI AND THE LAW

We have introduced how XAI addresses algorithmic transparency and explainability, and have also seen that transparency and explainability are deeply embedded into the fabric of the Act. Much can be found in common between the concepts and methods of XAI solutions and the Act, however, there are major differences which could hinder interdisciplinary work and public adoption in these areas. In this section, we examine four main topics to understand the scope and effects of these differences and suggest paths towards the support of an interdisciplinary approach.

4.1 Uncertainty in Legal Definitions

The definitions in both the GDPR and the Act leave room for flexible interpretations. This is desirable to some extent, as definitions which are too strict would hinder progress. Yet, any one particular interpretation could have significant

effects on the utility of the regulations and the applicability of XAI for transparency. Thus, we should lay down more concrete foundations in support of legal certainty.

4.1.1 The definition of AI. The definition of AI itself is a much-debated and amended point [14, 46, 47]. The current definition of the Act, given in Article (3)(1), is considered too broad: AI is defined in terms of a list of technologies. While the EU Commission is empowered to amend this list, it currently includes technologies which are not usually considered AI, including statistical modelling methods. Given the speed of innovation in AI, relying on a list as a definition is a brittle approach as it requires constant monitoring and updating. One frequently cited option is to instead consider the definition of the Organisation for Economic Co-operation and Development (OECD) as a common standard [44], which defines an AI system as a "machine-based system that can, for a given set of human-defined objectives, make predictions, recommendations, or decisions influencing real or virtual environments [...] with varying levels of autonomy."

Regarding transparency considerations, the final definition of AI will have wide-ranging consequences on the legal status and potential utility of XAI systems. Many XAI methods themselves rely on practices that fall under the currently proposed definition of AI. Does then the XAI system qualify as the same or as a different AI system separate from the underlying AI algorithm? This consideration has significant effects on, for example, the determination of liability in the case of *reasonably foreseeable misuse* or a *serious incident*. For example, who is liable when a user fails to operate the AI system according to its intended purposes because an explanation from the accompanying XAI system misled them?

We argue that the XAI system should not be treated separately from the AI system it explains, even if the provider of each system is different. The XAI system will need to be calibrated to work well with the AI system, regardless of whether it is model-agnostic or model-specific. Thus, taking the XAI system out of context by assessing it as a separate entity will inevitably lead to erroneous conclusions. Regulators should clarify the legal status of XAI under the Act, making it clear that it should be assessed as part of the AI system.

4.1.2 Further definitions. We highlight two more issues regarding definitions which are important for XAI. The definition of an *output* of an AI system is critical, as transparency requirements will depend directly on this definition. The GDPR refers to the decisions of automated systems, however, Bayamlıoğlu [2] points out that a definition of a decision is never clarified. Bygrave [9] suggests that a decision could be defined by a list of concepts (e.g., recommendation), and the intention of a user to act upon it. However, the lack of definition leaves requirements open to misinterpretation.

In contrast, the Act refers directly to the output of an AI system, exemplifying it with the terms "content, predictions, recommendations, or decisions". Yet here again, there is no definition of what constitutes an output. For instance, the output could be the internal representations of an AI algorithm – while the user cannot act directly on this output, an XAI system does leverage them to produce an explanation. It is important to clarify what an output is, since applying the same legal requirements to different types of outputs (e.g., internal and external) could become too restrictive.

Furthermore, Articles (1)(29)-(31) of the Act lay down the definitions of *training*, *validation*, and *testing data*. While these definitions are crucial for data-driven AI systems, methods such as planning and reinforcement learning do not rely on the same techniques as supervised learning from where these terms originate [4] It is therefore unclear what requirements should be fulfilled for an XAI system that relies on the input data of such systems to generate its explanations. Thus, a more general definition of data types would be necessary to cover a wider range of algorithms.

4.2 What is Transparency?

4.2.1 Interpretations of transparency. The Act frequently refers to necessary levels of transparency, but is weak on the explicit definition of such a property. While the common interpretation in XAI literature concerns the transparency of the algorithmic process itself, the Act offers a different, much broader interpretation.

In the interpretation of the Act, transparency is an overarching property of the AI system achieved through requirements detailed in Section 3.2. Moreover, Article 52(1) states that users should be made aware that they are interacting with an AI system. For example, if a doctor is advised by such a system, disclosure of the use of that system should be communicated to the patient. In the interpretation of XAI, transparency concerns solely the interpretability of the decision-making process. For white-box models, it would be easy to inspect the rules and knowledge which were used to create a decision. However, transparency is not a virtue of the black-box models which dominate today.

Both interpretations have different consequences on the design of transparent AI systems. In the Act, transparency is required to *an appropriate level*, but no distinction is made on what exactly is appropriate for different applications or tasks. One might imagine that a higher level of transparency is required for high-risk AI systems, but no greater level of detail is given in the Act. In XAI, the interpretation of transparency is often too narrow and technology-focused. It ignores societal and cultural concerns, making XAI less appealing as a solution for transparency.

Additionally, as discussed in section 3.2.1, Article 14(4) places lofty requirements on the understanding that human users are required to have. It is worth questioning whether such aims are even possible: full understanding of capacities and limitations is surely an impossible task for black-box models, while the task of correctly interpreting a system's output is already complicated by our discussion on the term *output*. A more feasible interpretation would at least require understanding of the limitations of the systems with respect to the '*intended purpose*' of the AI system, a term used throughout the Act. For example, is a medical diagnosis system less accurate at predicting diseases for different demographics? This will also allow the user to build an appropriate level of trust in the system.

We consider a hierarchical approach to defining transparency, which enables targeted explanations addressing the right "cognitive holes" of humans. On the lower levels of this hierarchy, for low-risk AI, we could have fully automated explanations from purely numerical to higher-level conceptual explanations. As a threshold is crossed, reaching high-risk systems, we could increasingly introduce human intervention to moderate the output of an XAI system. However, this threshold must be flexible. For example, non-interpretable systems may require more human intervention than white-box systems. Yet, more stringent requirements tend to slow down innovation.

We must be careful about how transparency is defined in the legislation. The definition has consequences for the classes of AI systems which are permitted to be deployed in reality. How to grade different levels of transparency is undoubtedly a difficult task, but the lack of detail given in the Act leaves a lot of room for different interpretations of what is appropriate. This can undermine legal certainty and the harmonisation goals of the Act, or stifle research.

4.2.2 Cybersecurity and transparency. An additional question that is not fully addressed in the Act is the extent to which AI systems become susceptible to cybersecurity threats through increased transparency. Article 14(4) requires human oversight to extend to AI-specific vulnerabilities (e.g., data poisoning, gaming, and adversarial examples). The risks of these attacks occurring would increase with more exposed knowledge of the AI system's workings. Furthermore, XAI systems themselves are not inherently trustworthy, and may provide manipulative or false explanations. Therefore, we urge XAI researchers to consider cybersecurity a high priority when developing their systems. Safety-by-design will be crucial for the wider adoption of XAI, however, present approaches leave this aspect unaddressed.

Medical diagnosis example Transparency, in the Act's interpretation, would ideally allow the patient to contest predictions, and the doctor to justify the system's decisions. However, in another interpretation of transparency explanations may be too vague or too complex to earn the patient's trust. Transparency according to XAI would mean that the doctor receives a list of numeric rankings among features of the data ordering them by importance. However, it is up to the doctor to interpret these numbers and deliver an intelligible explanation.

4.3 Conformity Assessment

Another area where the coverage of the Act may not reach its full potential is the methods by which AI providers achieve conformity to the requirements set out in Chapter 2. Documentation required by Article 11 of the Act is expected to demonstrate conformity with the regulation, and Article 19 requires the providers of high-risk systems to undergo an assessment of that conformity, as outlined in Article 43. Moreover, Article 17 also requires a comprehensive quality management system to be in place to ensure conformity during the entire lifecycle of the AI system.

4.3.1 Self-assessed conformity. Despite the Act's clear requirements for AI providers to document conformity of the system, one clear problem is presented in section 5.2.3 of the Explanatory Memorandum to the Act. It states that "the precise technical solutions to achieve compliance maybe be [...] developed in accordance with general engineering or scientific knowledge at the discretion of the provider of the AI system." This leaves room for AI providers to choose their own methods of demonstrating conformity to the requirements. This self-regulation is potentially problematic as it leaves open the possibility for AI providers to, in the worst case, fraudulently declare conformity of their systems or fail to adequately test the system. Correct technical documentation may still be drawn up as required, but the exact technical solutions for the demonstration of compliance with the regulation may be left to the discretion of the provider.

So far we have considered XAI methods as tools built into the AI system by its provider to improve transparency. In contrast, we may consider the use of XAI in the possibility of, for example, conformity assessment evasion. Here XAI could serve as an external tool for auditing purposes, highlighting potential regulatory violations to authorities.

Medical diagnosis example If the provider's conformity assessment procedures are hidden from view, there are clear opportunities to misrepresent the system's performance. After training, it's conceivable that the system may be tested on a dataset which contains examples from its training data, achieving artificially excellent performance. Alternatively, the test set may be made up of data from the same general demographics and populations which were used to train the system. When deployed in the real world, it will likely achieve substandard performance on demographics which are poorly represented in the training data.

4.3.2 Recertification. Another problem may arise when new models of AI systems are developed and released. Article 43(b) mandates recertification in the event that the system is substantially modified – but what constitutes a substantial modification? The interpretation of this may vary between different AI technologies.

Although in a different industry, one case which demonstrates the danger of misrepresenting changes to ensure conformity of a new version of a product is the introduction of the Boeing 737 MAX aircraft [33]. By playing down the importance of the new system, Boeing hoped to save many millions of dollars, avoiding recertification and the training of staff on the new model. This resulted in over 300 fatalities. Substantial changes to AI systems, for example, a change to a model's architecture to reduce inference time, could qualify as a substantial modification. This would mandate recertification under the Act, but the provider may fraudulently claim sufficient accuracy using the new architecture.

Many XAI methods rely on a specific model design or algorithm to work well. However, in the case of a substantial modification, these methods would likely not work without significant additional engineering. If we hope to use XAI for auditing, then new methods could be developed that address specifically the differences between the original and the modified system, without reliance on a specific model architecture.

4.3.3 Certification for XAI systems. As discussed in Section 4.1, the legal status of XAI systems that explain an AI system is unclear. Would the requirements of Chapter 2 of the Act need to be guaranteed separately for the AI and the XAI system, or could they be covered under the same unified assessment? As argued in Section 4.1.1, the former option would place unreasonable assessment requirements on providers. However, XAI systems introduce further complexity to the AI system, which affects their conformity under the Act. Under a unified assessment scheme, these new complexities could go unnoticed—unintentionally or otherwise, or be construed as the inherent capabilities of the underlying AI system. Current legal requirements should thus clarify how the assessment of XAI tools is carried out, focusing specifically on the improved capabilities of the AI system by virtue of the XAI system.

What is more, the regulation does not address how the effect of explanations on users should be accounted for. One might imagine a system that monitors the kind of explanations that are requested about the output of an AI system. If the same explanations are requested repeatedly by users, then the provider should be compelled to investigate and be held accountable for any oversights. An incorrect explanation is arguably more damaging than no explanation at all, thus, explanations should be subject to stringent quality control and conformity assessment.

This raises a final issue: as we saw, the Act, as worded, cannot on its own provide software developers with enough detail to lead to "actionable" design decisions. This however does not necessarily mean that we argue for a revision of the primary legislation. There are other mechanisms through which a law can gain specificity and clarity: for example through court decisions or official guidance documents. In the case of the GPDR, the guidelines of the former Article 29 Working Party played an important role, as did the guidance of the national Information Commissioner Offices (ICO). Indeed, one role of the ICOs is to work closely with businesses, not just penalising them after the fact, but advising them how to design their compliance mechanisms.

In some cases, including medical devices, the Notified Bodies check the conformity assessment of the developers, in many others, self-certification is sufficient. The combination of a limited role for Notified Bodies and the lack of detail in the Act gives Industrial standards a particularly prominent role. The Act does not mandate that developers adhere to industrial standards. They can interpret the requirements of the Act on their own, and decide how to implement transparency requirements. However, in this case the risk of misreading the law rests with them. If by contrast they adhere to any future standards developed by CEN (European Committee for Standardisation) and CENELEC (European Committee for Electrotechnical Standardisation), they are protected by a "presumption of conformity".

The problem with this approach for our topic are however many-fold. We have argued that a close alignment of technical and legal concepts is needed to give the Act effect. As constituted, however, these standard-setting bodies lack the expertise and remit to consider for instance Human Rights implications of their standards. Standard-setting bodies, because of the influence industry plays in them, are likely to emphasise compliance-oriented transparency over user-empowering transparency. Here the Act proposal may be much less efficient in protecting basic rights than, for example, the envisaged UK regulatory framework for AI. Unlike the "top-down" EU approach that may result in an attempt to define transparency for a whole range of disparate applications, the UK approach is emphasising domain specific regulators. These regulators not only often have relevant legal expertise, and a statutory duty to consider

human rights implications, they are also much better placed to understand the different roles transparency plays in their respective fields.

4.4 Explainable Data

While clarifying the definitions of the Act is mostly in the hands of lawmakers, XAI too could learn much from the Act. In particular, Article 10 and Recital 44 of the Act describe detailed data quality and governance requirements. However, XAI is most often concerned with explaining the decision-making process. Yet, data-related requirements and the subsequent design choices have clear transparency-related implications.

It would be beneficial to introduce automated *data explanation* techniques. These would examine the effects of changing inherent properties of the input dataset on the AI system by purposefully biasing, distorting, or introducing irrelevant samples to the dataset. Problems with the dataset and system itself may be uncovered, and at the same time conformity to Article 10(2) of the Act could be demonstrated.

Another promising direction in data governance is data documentation. Gebru et al. [21] introduce the "datasheet for datasets", which is a careful way to trace the motivation, creation, and qualities of a dataset. Measures like this would improve the transparency of AI systems, and the need for such measures should be emphasised in the Act.

Medical diagnosis example In such a critical setting, it is imperative to understand the data used to train the model. A data explanation should give insights into the demographic information across various dimensions (e.g., age, medical history); and clarify the sources of the data. Using XAI for data explanations, we examine the effect on the prediction of different values of different features, exploring how the model handles different populations. However, this example also illustrates the fine line between data expressiveness and privacy preservation, as increasingly expressive data often entails greater levels of personally identifying information.

5 CONCLUSION

In our call for technological and legal collaboration, we face the legal question of what the ultimate purpose of top-level regulations such as the AI Act is. In the view of Hart et al. [26], laws are instructions for public officials, such as judges, for whom interpreting the law is based on a post-hoc normative analysis. In this context, the current definition of transparency may suffice. However, in more recent jurisprudence, focus on "design-centric" legislation, such as data protection-by-design and privacy-by-design [11, 29], has emerged. Here, technology specific law is required [29] which warrants cross-disciplinary work. Ultimately, our recommendations should serve as guidance to independent supervisory bodies that fill in the gaps between the AI Act and the concrete design practices

However, many issues remain, such as the role of XAI in meeting the two types of transparency requirements, user-empowering and compliance-oriented. Additionally, clarifying the discrepancy between *interpretations* and *explanations*, and whether this has legal consequences. The degree to which the human user is responsible for moderating the power of non-interpretable systems also merits further discussion – should the human have greater power to contend the decision of a system, the less interpretable it is? One should also look to the role of standard-setting bodies in the context of the certification problem.

ACKNOWLEDGEMENTS

We would like to thank Burkhard Schafer for his invaluable comments and help with the legal considerations of the paper. We also thank Shay Cohen, Alan Bundy, Liane Guillou, and Kwabena Nuamah for their comments on drafts.

This work was supported in part by the UKRI Centre for Doctoral Training in Natural Language Processing, funded by the UKRI (grant EP/S022481/1) and the University of Edinburgh, School of Informatics and School of Philosophy, Psychology & Language Sciences.

REFERENCES

[1] Shahin Atakishiyev, Mohammad Salameh, Hengshuai Yao, and Randy Goebel. 2021. Explainable Artificial Intelligence for Autonomous Driving: A Comprehensive Overview and Field Guide for Future Research Directions. *arXiv:2112.11561 [cs]* (Dec. 2021). arXiv:2112.11561 [cs]

[2] Emre Bayamlıoğlu. 2022. The Right to Contest Automated Decisions under the General Data Protection Regulation: Beyond the so-Called "Right to Explanation". *Regulation & Governance* 16, 4 (2022), 1058–1078. https://doi.org/10.1111/rego.12391

[3] Emily M. Bender, Timnit Gebru, Angelina McMillan-Major, and Shmargaret Shmitchell. 2021. On the Dangers of Stochastic Parrots: Can Language Models Be Too Big?. In *Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency (FAccT '21)*. Association for Computing Machinery, New York, NY, USA, 610–623. https://doi.org/10.1145/3442188.3445922

[4] Christopher M Bishop. 2006. *Pattern Recognition and Machine Learning*. Springer, New York, NY, USA.

[5] Eduardo Bismarck. 2021. PL 21/2020. https://www.camara.leg.br/propostas-legislativas/2236340.

[6] J. Broekens, M. Heerink, and H. Rosendal. [n. d.]. Assistive Social Robots in Elderly Care: A Review. 8, 2 ([n. d.]), 94–103. https://doi.org/10.4017/gt.2009.08.02.002.00

[7] Joy Buolamwini and Timnit Gebru. 2018. Gender Shades: Intersectional Accuracy Disparities in Commercial Gender Classification. In *Proceedings of the 1st Conference on Fairness, Accountability and Transparency*. PMLR, 77–91.

[8] Nadia Burkart and Marco F. Huber. 2021. A Survey on the Explainability of Supervised Machine Learning. *Journal of Artificial Intelligence Research* 70 (May 2021), 245–317. https://doi.org/10.1613/jair.1.12228

[9] Lee A Bygrave. 2001. Automated Profiling: Minding the Machine: Article 15 of the EC Data Protection Directive and Automated Profiling. *Computer Law & Security Review* 17, 1 (Jan. 2001), 17–24. https://doi.org/10.1016/S0267-3649(01)00104-2

[10] Diogo V. Carvalho, Eduardo M. Pereira, and Jaime S. Cardoso. 2019. Machine Learning Interpretability: A Survey on Methods and Metrics. *Electronics* 8, 8 (Aug. 2019), 832. https://doi.org/10.3390/electronics8080832

[11] Ann Cavoukian et al. 2009. Privacy by design: The 7 foundational principles. *Information and privacy commissioner of Ontario, Canada* 5 (2009), 12.

[12] Lei Chen, Yu Qiu, Junan Zhao, Jing Xu, and Ao Liu. 2021. CPKD: Concepts-Prober-Guided Knowledge Distillation for Fine-Grained CNN Explanation. In *2021 2nd International Conference on Electronics, Communications and Information Technology (CECIT)*. 421–426. https://doi.org/10.1109/CECIT53797.2021.00081

[13] European Commission. 2021. Proposal for an Artificial Intelligence Act 2021/0106(COD).

[14] Council of the European Union. 2022. General Approach on Comission Proposal 2021/0106(COD) (14954/22).

[15] Richard Dazeley, Peter Vamplew, Cameron Foale, Charlotte Young, Sunil Aryal, and Francisco Cruz. 2021. Levels of Explainable Artificial Intelligence for Human-Aligned Conversational Explanations. *Artificial Intelligence* 299 (Oct. 2021), 103525. https://doi.org/10.1016/j.artint.2021.103525

[16] Thomas G. Dietterich. 2017. Steps Toward Robust Artificial Intelligence. *AI Magazine* 38, 3 (Oct. 2017), 3–24. https://doi.org/10.1609/aimag.v38i3.2756

[17] Nadine Dorries. 2022. *Establishing a Pro-Innovation Approach to Regulating AI.* Technical Report. Department for Digital, Culture, Media and Sport, London.

[18] Finale Doshi-Velez, Mason Kortz, Ryan Budish, Chris Bavitz, Sam Gershman, David O'Brien, Kate Scott, Stuart Schieber, James Waldo, David Weinberger, Adrian Weller, and Alexandra Wood. 2019. Accountability of AI Under the Law: The Role of Explanation. *arXiv:1711.01134 [cs, stat]* (Dec. 2019). arXiv:1711.01134 [cs, stat]

[19] Lilian Edwards and Michael Veale. 2018. Enslaving the Algorithm: From a "Right to an Explanation" to a "Right to Better Decisions"? *IEEE Security & Privacy* 16, 3 (May 2018), 46–54. https://doi.org/10.1109/MSP.2018.2701152

[20] Andre Esteva, Brett Kuprel, Roberto A. Novoa, Justin Ko, Susan M. Swetter, Helen M. Blau, and Sebastian Thrun. [n. d.]. Dermatologist-Level Classification of Skin Cancer with Deep Neural Networks. 542, 7639 ([n. d.]), 115–118. https://doi.org/10.1038/nature21056

[21] Timnit Gebru, Jamie Morgenstern, Briana Vecchione, Jennifer Wortman Vaughan, Hanna Wallach, Hal Daumé III, and Kate Crawford. 2021. Datasheets for Datasets. *Commun. ACM* 64, 12 (Nov. 2021), 86–92. https://doi.org/10.1145/3458723

[22] Bryce Goodman and Seth Flaxman. 2017. European Union Regulations on Algorithmic Decision-Making and a "Right to Explanation". *AI Magazine* 38, 3 (Oct. 2017), 50–57. https://doi.org/10.1609/aimag.v38i3.2741

[23] Patrick Grady. 2022. The EU Should Clarify the Distinction Between Explainability and Interpretability in the AI Act.

[24] Riccardo Guidotti, Anna Monreale, Salvatore Ruggieri, Franco Turini, Fosca Giannotti, and Dino Pedreschi. 2018. A Survey of Methods for Explaining Black Box Models. *Comput. Surveys* 51, 5 (Aug. 2018), 93:1–93:42. https://doi.org/10.1145/3236009

[25] Balint Gyevnar, Massimiliano Tamborski, Cheng Wang, Christopher G. Lucas, Shay B. Cohen, and Stefano V. Albrecht. 2022. A Human-Centric Method for Generating Causal Explanations in Natural Language for Autonomous Vehicle Motion Planning. In *Workshop on Artificial Intelligence for Autonomous Driving*. International Joint Conference on Artificial Intelligence. https://doi.org/10.48550/arXiv.2206.08783 arXiv:2206.08783 [cs]

[26] Herbert Lionel Adolphus Hart, Herbert Lionel Adolphus Hart, Joseph Raz, and Leslie Green. 2012. *The concept of law*. oxford university press.

[27] Christian Herzog. 2022. On the Ethical and Epistemological Utility of Explicable AI in Medicine. *Philosophy & Technology* 35, 2 (May 2022), 50. https://doi.org/10.1007/s13347-022-00546-y

[28] Mireille Hildebrandt. 2020. *Law for Computer Scientists and Other Folk* (first edition ed.). Oxford University Press, Oxford, United Kingdom.

[29] Mireille Hildebrandt and Laura Tielemans. 2013. Data protection by design and technology neutral law. *Computer Law & Security Review* 29, 5 (2013), 509–521.

[30] HLEG-AI. 2019. *Ethics Guidelines for Trustworthy AI*. Publications Office of the European Union, Directorate-General for Communications Networks.

[31] Robert R. Hoffman, Shane T. Mueller, Gary Klein, and Jordan Litman. 2019. Metrics for Explainable AI: Challenges and Prospects. *arXiv:1812.04608 [cs]* (Feb. 2019). arXiv:1812.04608 [cs]

[32] The White House. 2022. Blueprint for an AI Bill of Rights. https://www.whitehouse.gov/ostp/ai-bill-of-rights/.

[33] House Committee on Transportation and Infrastructure. [n. d.]. Final Committee Report on the Design, Development and Certification of Boeing 737 MAX.

[34] Fleur Jongepier and Esther Keymolen. 2022. Explanation and Agency: Exploring the Normative-Epistemic Landscape of the "Right to Explanation". *Ethics and Information Technology* 24, 4 (Nov. 2022), 49. https://doi.org/10.1007/s10676-022-09654-x

[35] Davinder Kaur, Suleyman Uslu, Kaley J. Rittichier, and Arjan Durresi. 2022. Trustworthy Artificial Intelligence: A Review. *Comput. Surveys* 55, 2 (Jan. 2022), 39:1–39:38. https://doi.org/10.1145/3491209

[36] Markus Langer, Daniel Oster, Timo Speith, Holger Hermanns, Lena Kästner, Eva Schmidt, Andreas Sesing, and Kevin Baum. 2021. What Do We Want from Explainable Artificial Intelligence (XAI)? – A Stakeholder Perspective on XAI and a Conceptual Model Guiding Interdisciplinary XAI Research. *Artificial Intelligence* 296 (July 2021), 103473. https://doi.org/10.1016/j.artint.2021.103473

[37] Qiang Liu, Pan Li, Wentao Zhao, Wei Cai, Shui Yu, and Victor C. M. Leung. 2018. A Survey on Security Threats and Defensive Techniques of Machine Learning: A Data Driven View. *IEEE Access* 6 (2018), 12103–12117. https://doi.org/10.1109/ACCESS.2018.2805680

[38] Yu-Chen Lo, Stefano E. Rensi, Wen Torng, and Russ B. Altman. [n. d.]. Machine Learning in Chemoinformatics and Drug Discovery. 23, 8 ([n. d.]), 1538–1546. https://doi.org/10.1016/j.drudis.2018.05.010

[39] Bodhisattwa Prasad Majumder, Oana-Maria Camburu, Thomas Lukasiewicz, and Julian McAuley. 2022. Knowledge-Grounded Self-Rationalization via Extractive and Natural Language Explanations. https://doi.org/10.48550/arXiv.2106.13876 arXiv:2106.13876 [cs]

[40] Tim Miller. 2019. Explanation in Artificial Intelligence: Insights from the Social Sciences. *Artificial Intelligence* 267 (Feb. 2019), 1–38. https://doi.org/10.1016/j.artint.2018.07.007

[41] Sina Mohseni, Niloofar Zarei, and Eric D. Ragan. 2021. A Multidisciplinary Survey and Framework for Design and Evaluation of Explainable AI Systems. *ACM Transactions on Interactive Intelligent Systems* 11, 3-4 (Aug. 2021), 24:1–24:45. https://doi.org/10.1145/3387166

[42] Christoph Molnar. 2022. *Interpretable Machine Learning*.

[43] J. D. Myers, H. E. Pople, and R. A. Miller. [n. d.]. CADUCEUS: A Computerized Diagnostic Consultation System in Internal Medicine. In *Proceedings of the Annual Symposium on Computer Application in Medical Care* (1982). 44–47.

[44] OECD. 2019. Recommendation of the Council on Artificial Intelligence. https://legalinstruments.oecd.org/en/instruments/OECD-LEGAL-0449.

[45] OpenAI. 2022. ChatGPT: Optimizing Language Models for Dialogue. https://openai.com/blog/chatgpt/.

[46] European Parliament. 2022. Draft Opinion of the Committee on Industry, Research and Energy (PA\1250560EN).

[47] European Parliament. 2022. Draft Opinion of the Committee on Legal Affairs (PA\1250671EN).

[48] European Parliament and Council of the European Union. 2016. Regulation (EU) 2016/679 (General Data Protection Regulation).

[49] European Parliament and Council of the European Union. 2017. Regulation (EU) 2017/745 on medical devices.

[50] European Parliament and Council of the European Union. 2017. Regulation (EU) 2017/746 on in vitro diagnostic medical devices.

[51] Article 29 Data Protection Working Party. 2018. Guidelines on Automated Individual Decision-Making and Profiling for the Purposes of Regulation 2016/679 (WP251rev.01).

[52] Vitali Petsiuk, Rajiv Jain, Varun Manjunatha, Vlad I. Morariu, Ashutosh Mehra, Vicente Ordonez, and Kate Saenko. 2021. Black-Box Explanation of Object Detectors via Saliency Maps. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 11443–11452.

[53] Sarvapali D. Ramchurn, Sebastian Stein, and Nicholas R. Jennings. 2021. Trustworthy Human-AI Partnerships. *iScience* 24, 8 (Aug. 2021), 102891. https://doi.org/10.1016/j.isci.2021.102891

[54] Xavier Renard, Thibault Laugel, and Marcin Detyniecki. 2021. Understanding Prediction Discrepancies in Machine Learning Classifiers. https://doi.org/10.48550/arXiv.2104.05467 arXiv:2104.05467 [cs]

[55] Marco Tulio Ribeiro, Sameer Singh, and Carlos Guestrin. 2016. "Why Should I Trust You?": Explaining the Predictions of Any Classifier. *arXiv:1602.04938 [cs, stat]* (Aug. 2016). arXiv:1602.04938 [cs, stat]

[56] Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. 2022. High-Resolution Image Synthesis with Latent Diffusion Models. In *2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. IEEE, 10674–10685. https://doi.org/10.1109/CVPR52688.2022.01042

[57] Stuart J. Russell and Peter Norvig. [n. d.]. *Artificial Intelligence: A Modern Approach* (fourth edition ed.). Pearson.

[58] Sascha Saralajew, Ammar Shaker, Zhao Xu, Kiril Gashteovski, Bhushan Kotnis, Wiem Ben-Rim, Jürgen Quittek, and Carolin Lawrence. 2022. A Human-Centric Assessment Framework for AI. arXiv:2205.12749 [cs]

[59] Evan Selinger and Woodrow Hartzog. 2016. Facebook's Emotional Contagion Study and the Ethical Problem of Co-Opted Identity in Mediated Environments Where Users Lack Control. *Research Ethics* 12, 1 (Jan. 2016), 35–43. https://doi.org/10.1177/1747016115579531

[60] Amanda Sharkey and Noel Sharkey. [n. d.]. Children, the Elderly, and Interactive Robots. 18, 1 ([n. d.]), 32–38. https://doi.org/10.1109/MRA.2010.940151

[61] Francesco Sovrano, Salvatore Sapienza, Monica Palmirani, and Fabio Vitali. 2022. Metrics, Explainability and the European AI Act Proposal. *J* 5, 1 (March 2022), 126–138. https://doi.org/10.3390/j5010010

[62] Francesco Sovrano and Fabio Vitali. 2022. Explanatory Artificial Intelligence (YAI): Human-Centered Explanations of Explainable AI and Complex Data. *Data Mining and Knowledge Discovery* (Oct. 2022). https://doi.org/10.1007/s10618-022-00872-x

[63] Ilia Stepin, Jose M. Alonso, Alejandro Catala, and Martín Pereira-Fariña. 2021. A Survey of Contrastive and Counterfactual Explanation Generation Methods for Explainable Artificial Intelligence. *IEEE Access* 9 (2021), 11974–12001. https://doi.org/10.1109/ACCESS.2021.3051315

[64] Kathryn Tunyasuvunakool, Jonas Adler, Zachary Wu, Tim Green, Michal Zielinski, Augustin Žídek, Alex Bridgland, Andrew Cowie, Clemens Meyer, Agata Laydon, Sameer Velankar, Gerard J. Kleywegt, Alex Bateman, Richard Evans, Alexander Pritzel, Michael Figurnov, Olaf Ronneberger, Russ Bates, Simon A. A. Kohl, Anna Potapenko, Andrew J. Ballard, Bernardino Romera-Paredes, Stanislav Nikolov, Rishub Jain, Ellen Clancy, David Reiman, Stig Petersen, Andrew W. Senior, Koray Kavukcuoglu, Ewan Birney, Pushmeet Kohli, John Jumper, and Demis Hassabis. 2021. Highly Accurate Protein Structure Prediction for the Human Proteome. *Nature* 596, 7873 (Aug. 2021), 590–596. https://doi.org/10.1038/s41586-021-03828-1

[65] Jessica Vamathevan, Dominic Clark, Paul Czodrowski, Ian Dunham, Edgardo Ferran, George Lee, Bin Li, Anant Madabhushi, Parantu Shah, Michaela Spitzer, and Shanrong Zhao. [n. d.]. Applications of Machine Learning in Drug Discovery and Development. 18, 6 ([n. d.]), 463–477. https://doi.org/10.1038/s41573-019-0024-5

[66] Michael Veale and Frederik Zuiderveen Borgesius. 2021. Demystifying the Draft EU Artificial Intelligence Act. *Computer Law Review International* 22, 4 (Aug. 2021), 97–112. https://doi.org/10.9785/cri-2021-220402 arXiv:2107.03721 [cs]

[67] Sandra Wachter, Brent Mittelstadt, and Luciano Floridi. 2017. Why a Right to Explanation of Automated Decision-Making Does Not Exist in the General Data Protection Regulation. *International Data Privacy Law* 7, 2 (June 2017), 76–99. https://doi.org/10.1093/idpl/ipx005

[68] Michael Winikoff and Julija Sardelić. 2021. Artificial Intelligence and the Right to Explanation as a Human Right. *IEEE Internet Computing* 25, 2 (March 2021), 116–120. https://doi.org/10.1109/MIC.2020.3045821

Received 06 February 2023; revised 12 March 2023; accepted 5 June 2023