Bridging the transparency gap

ECAI 2023
K. Gal et al. (Eds.)

doi:10.3233/FAIA230367

Bridging the Transparency Gap: What Can Explainable AI Learn from the AI Act?

Balint Gyevnar[a;*], Nick Ferguson[a] and Burkhard Schafer[b]

[a]School of Informatics, University of Edinburgh
[b]Edinburgh Law School, University of Edinburgh

Abstract. The European Union has proposed the Artificial Intelligence Act which introduces detailed requirements of transparency for AI systems. Many of these requirements can be addressed by the field of explainable AI (XAI), however, there is a fundamental difference between XAI and the Act regarding what transparency is. The Act views transparency as a means that supports wider values, such as accountability, human rights, and sustainable innovation. In contrast, XAI views transparency narrowly as an end in itself, focusing on explaining complex algorithmic properties without considering the socio-technical context. We call this difference the "transparency gap". Failing to address the transparency gap, XAI risks leaving a range of transparency issues unaddressed. To begin to bridge this gap, we overview and clarify the terminology of how XAI and European regulation – the Act and the related General Data Protection Regulation (GDPR) – view basic definitions of transparency. By comparing the disparate views of XAI and regulation, we arrive at four axes where practical work could bridge the transparency gap: defining the scope of transparency, clarifying the legal status of XAI, addressing issues with conformity assessment, and building explainability for datasets.

1 Introduction

Artificial Intelligence (AI) systems are an increasing presence in twenty-first-century life and are now available to most, from commercial users to non-technical enthusiasts. While public-facing demonstrations of the capability of AI may lead to the perception that AI is an emerging field, the use of AI systems is already widespread. Facial recognition, recommender systems, and medical diagnoses are applications which have been utilising AI for many years. However, as we see black-box AI increasingly deployed in safety-critical applications, issues relating to their deployment in society have arisen.

Recognising the urgent need to address such concerns, regulators have been drawing up proposals to tackle the challenges of the "AI-era". In the USA, the White House has introduced the "Blueprint for an AI Bill of Rights", proposing five main areas of regulation for AI [26]. The UK has outlined its own pro-innovation approach to AI regulation and is working on a "National AI Strategy" [14]. In the EU, an ambitious regulation called the *AI Act* is under consideration [11]. The Act places requirements on AI providers (developers of AI systems) relating to transparency and explainability. It takes a proportional risk-based approach to defining these requirements and proposes comprehensive conformity assessment conditions.

* Corresponding Author. Email: balint.gyevnar@ed.ac.uk

Scientists also acknowledge the need for transparent and trustworthy AI [24, 30, 63] that respect the law, ethics, and social considerations, but which are also robust in the real-world. Researchers regard explanations as essential to transparency in human-AI interaction and explainable AI (XAI) now focuses to a large extent on achieving human-aligned AI via social explanations [39, 58, 62].

However, the definitions, scope, and purpose of transparency in regulations are not in agreement with how technological approaches understand them [44], let alone how they translate into more trustworthy systems [34]. XAI views transparency merely as an algorithmic property that offers practical solutions but through a limited, technology-focused scope [35, 40]. By contrast, the law treats transparency as a quality of complex socio-technical interactions between the AI and its users, developers, owners, and wider society. We call this mismatch of views the **"transparency gap"**. Transparency in the view of the law is not a goal in itself, but a *means* that is needed to promote a range of very different values. As a consequence regulation in the EU differentiates among various forms of technical, enabling, and protective transparency [20], which may be amenable only in varying degrees to computational solutions. XAI by contrast discusses concepts of algorithmic transparency – e.g., black box or interpretable systems – as formal properties of a computer system in isolation [60]. In this view, achieving transparency is an *end* in itself, necessary by virtue of the complexity of AI systems. Thus, the transparency gap is also the difference in viewing transparency as a means or an end.

The transparency gap is also one of the reasons why many of the legal requirements are being criticised by writers from the computer science community as being ineffective, overreaching, or technically infeasible [32, 68]. The mistake here is to assume that the broad and ambitious transparency requirements that the Act lays out are engineering instructions, and can be addressed solely through appropriate design decisions. A much better understanding however is to ask which design decisions facilitate, and which ones hinder, the type of transparency aspiration that the Act aims for, without equating algorithmic explainability with legal transparency outright.

This type of mismatch is not a new phenomenon. System theory teaches us that law, like all social systems, is cognitively open but normatively closed [65]. This means, for instance, that the concept of "causation" – in science understood as a purely descriptive term used for the explanation of observations – is "normatively constrained" [21] so that not every causation in the sense of science is also causation in law, turning scientific explanations into legal justifications. This is sometimes experienced as a misunderstanding: if only legislators understood technology better, many scientists might feel, they could

legislate more appropriately, and for scientists that means in a way so that the legal norms directly translate into system requirements. It is certainly true that some technology regulation is deeply misguided due to insufficient technological competency by the legislator (e.g., [13]). However, much more common is a mutual and indeed necessary or inevitable misunderstanding. The law cannot but distort the technical conceptions that originate in science discourse if it wants to stay true to its own logic [66]. We argue instead that good XAI should not try to usurp the role of the law, or "solve" the legal problem of transparency, but can nonetheless anticipate how the law will (mis)understand its concepts, and in this way find new approaches to assist the legal system in achieving its objective, and bridge the transparency gap.

To this end, we first clarify and compare the terminology of how XAI and European regulation – the GDPR and the Act – view basic definitions of transparency (Sections 2 and 3), establishing the two disparate perspectives. Based on this discussion, we then identify four axes (Section 4) along which cross-disciplinary work should be placed to begin to bridge the transparency gap:

1. **Define transparency (Section 4.1)**. To increase legal certainty and to inform the design of XAI systems basic notions such as transparency, interpretability, and explainability should be clearly defined and scoped.
2. **Clarify the legal status of XAI (Section 4.2)**. XAI methods are often based on methods which fit the legal definition of AI system in the Act. An AI system and an XAI tool should be considered as one unit, but the Act may treat the systems separately, conflating the transparency requirements and their consequences.
3. **Conformity assessments (Section 4.3)**. The Act lacks guidance on the process of certifying algorithmic transparency, which raises the question of how XAI systems can be certified for transparency while leaving open the possibility of self-regulation by providers.
4. **Explainable data (Section 4.4)**. The Act defines strong requirements for data quality and governance. XAI has so far neglected data transparency but should extend to explaining the effects of inherent properties of datasets on the functioning of AI systems.

2 Transparency and Explainable AI

Before discussing XAI, it is important to understand from where current concerns about opaque AI systems originate. The first AI systems were based on symbolic logic, where knowledge about the world 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, which aligns with human cognition making them intelligible to people [38, 39].

In contrast, deep learning models that dominate today are not built on such tangible representations of data. These models, usually based on *neural networks*, 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. While these models are highly performant, public opinion has shifted towards expressing fundamental concerns about their social, economic, ethical, and political effects [31, 59] often attributed to a loss of autonomy and a socio-technical blindness exacerbated by unclear scientific public discourse [27].

Recognising that a lack of transparency is diminishing trust in AI systems, methods that explain AI models were developed, forming the field of XAI. Most surveys of the field [6, 18, 41, 42, 64] define the methods of XAI as self-explanatory systems that reveal the reasoning behind the outputs of AI models. In addition, we give a practical lower-level definition in terms of four attributes to help regulators pin down what XAI is. An explainable AI system should:

1. explain 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.

2.1 Explaining systems' outputs

Different AI systems have different inherent properties that make their outputs more or less suited for explanations. There is a significant amount of debate in XAI regarding the meanings of interpretability, explainability, justification, 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 in a legal context are further discussed in Section 4.1.

Miller [39] equates *interpretability* with explainability. However, others argue, and we support this view, that while interpretability and explainability are connected, they are distinct properties. According to Molnar [43], what makes some AI systems interpretable, and with that inherently human-understandable, is their low complexity. A reasonably skilled human user can understand the output of such a system, and how it was derived from the input, even in the absence of an explanation generated by the AI. 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: a reasonably skilled user understands how the data informs the outcome.

In contrast, *explainability* is the property of any AI system whose output comes with an automatically generated output that has the syntactic form of an explanation. It is, in other words, the ability of an AI system to relevantly communicate the reasoning behind its decision-making process. 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 which affected its decision and thus becomes explainable.

A *justification* gives a teleological rather than a mechanistic explanation. In Miller's words, 'it explains why a decision is good, but does not necessarily aim to give an explanation of the actual decision making process' [39, p8]. A self-driving car might justify stopping for a pedestrian as the "lawfully correct and safe action" without explaining the mechanisms that transformed the input into that decision.

Finally, if we focus solely on the algorithmic properties of an AI system, then *transparency* becomes the same concept as interpretability [2, 41]. However, as shown in Section 4.1, this interpretation is too restrictive because it clashes with a broader vision of transparency, such as the top-down view advocated by the Act.

2.2 Methods of XAI

Current XAI methods include, among others, analysis of feature importance [54], saliency maps [51], counterfactual methods [19], recognising textual entailment [36], and knowledge distillation [10]. Many XAI methods are themselves AI systems (e.g., natural language inference [7]) by the definition of an AI system in the Act which raises further legal questions, discussed further in Section 4.2.

Researchers often categorise AI systems based on interpretability [6, 56], contrasting non-interpretable black box systems such as neural networks against interpretable white box systems such as decision trees. Interpretability has a crucial influence on the design choices of XAI methods: using white-box models, we can more easily guarantee verifiability and causality. However, this usually comes at the cost of expressiveness and accuracy [8]. Both in regulation and technology, careful balancing is needed so that interpretability and accuracy are present at the desired levels.

Explainability also plays a key role in determining the kind of transparency XAI systems can offer. *Ante-hoc* explanations are generated directly from the internal representations and processes of white box systems, while *post-hoc* explanations are inferred from an output after a decision was made. Thus, ante-hoc explanations are truthful to the decision process by design. Post-hoc explanations may distort the causality underlying the model's decision process and require more effort to generate, but they apply to both white and black box systems.

Finally, 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 certified. Is the due diligence of the AI provider called into question due to their selection of an unsuitable XAI system even when the underlying AI system functions properly? This raises issues around the certification process, which we tackle in Section 4.3.

2.3 Stakeholders

We must also consider how transparency can be achieved for different stakeholders of AI. To this end, Langer et al. [33] have given a taxonomy of XAI from the perspective of stakeholders. They consider a feedback loop between the explanation process and the stakeholders based on four groups: developers, users, deployers, and affected parties. Their categorisation aligns with the definitions of the Act, which considers similar stakeholder groups. For example, Article 3 of the Act gives legal definitions of the provider, user, importer, etc.

Additionally, Mohseni et al. [42] suggest three distinct end-users for XAI: *AI novices*, *data experts*, and *AI experts*. 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 include, among others, bias, privacy, and trust, which are issues that regulators are addressing in the Act, and which increased transparency is supposed to alleviate. 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 [52]. 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.

2.4 Accuracy and Relevancy

Finally, the question of evaluating the quality of generated explanations remains, especially its effects on the perceived transparency of the AI system. Here, we must specify the 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 providers, we might expect objective correctness to be more important. There has been work on creating metrics for explanation performance evaluation [25], however, the design of metrics that clearly show compliance with regulation is a new challenge [62].

Metrics are also essential as different XAI methods applied to the same data can produce very different explanations [53]. 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.

3 Transparency and the Law

As we saw, XAI delivers algorithmic transparency, but its approaches are focused on technical aspects, despite recent calls for a stakeholder-directed approach focusing on trustworthiness.

This leads us back to the transparency gap. In the legal context, transparency is most often seen as a means to achieve broader goals, most importantly here, algorithmic accountability [22, 29], yet *not* necessarily trustworthiness. While a trustworthy system gives the user good reasons to accept the output of the system as correct, an accountable system allows them to allocate blame appropriately if the outcome turns out to be incorrect. In considering accountability, laws often fulfil a dual function. They try to prevent harm from occurring, but they also allocate responsibility if harm does incur nonetheless. While both objectives require transparency, they may require different conceptions of transparency to fulfil these objectives. In this context, the transparency gap is not just the inter-disciplinary means-end friction we saw in the introduction, but also an *intra*-disciplinary friction as XAI figures out to what end it is building transparent systems.

To bridge the transparency gap, we need to understand the transparency requirements in the AI Act, and here it helps to put them in a historical context. A "right to explanation" for automated decisions was first trialled in the landmark data protection act of the EU, the General Data Protection Regulation (GDPR) [49]. By tracing the genealogy of the concept of explainability in the AI Act to its predecessor in the GDPR, and identifying both continuities and differences in the legislative language, we can get a better sense of the scope and limits of this provision and how they relate to the transparency gap.

3.1 The Right to an Explanation

Goodman and Flaxman [17] 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]

Kaminski [29] supports this view by arguing that the 'GDPR establishes multiple layers of transparency' in which 'there is a clear relationship between the individual rights the GDPR establishes—contestation, correction, and erasure—and the kind of individualized transparency it requires.' Malgieri and Comandé [37] further refined the "right to explanation" by combining the rights to transparency and comprehensibility to distinguish between different levels of information. Much of these arguments are substantiated by the guidelines of the former data protection advisory board of the EU, Article 29 Data Protection Working Party (WP29) [50], which include a discussion of a "right to be informed" and notice mechanisms for automated

[1] Recitals are part of the preamble to a treaty that 'articulate shared assumptions, goals and explanations concerning the treaty' [23, p86].

decision-making. Additionally, Casey et al. [9] cite the EU data protection authorities to argue that algorithmic auditing and "data protection by design" methodologies codified by the GDPR are really what substantiate a "right to explanation". Furthermore, Winikoff and Sardelić [70] 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.

However, both the existence and the utility of such a "right to explanation" have been called into question [69]. Importantly, it is argued 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. This was acknowledged by WP29 [50], which stated that explanations need not be 'complex mathematical explanation[s] about how algorithms or machine-learning work', instead they should be 'clear and comprehensive ways to deliver information to the data subject'.

Edwards and Veale [15] also warn against the "illusion of a legal right". Given the inherent technological difficulties in generating helpful explanations, the danger is that the Act has the unintended consequence of permitting low-quality automated decisions as long as the affected party gets some form of explanation, even if they won't be able to in practice use this explanation to challenge the outcome. Instead, Edwards and Veale suggest several actionable routes to ensure a "right to better decisions", e.g., via data protection impact assessments. The arguments for the "illusion of a legal right" are further supported by Bayamlıoğlu [4] who emphasise that it is a "right to contest" in Article 22(3) that should drive the discussion around tangible ways to achieve transparency. In this view, transparency is no longer an end in itself, but a means to achieve effective contestation, and should be evaluated by the contribution they make to this aim. As we will see, this is a view that the Act takes to a great extent.

XAI researchers have continually used the "right to explanation" to justify their design choices (e.g., in [64, 3, 25, 18]), but they should remember that existence and scope of a new algorithm-centric "right to explanation" under the GDPR remain contested and so far no case law exists on its interpretation in this aspect. However, recently the Court of Justice of the European Union (CJEU) has advanced the debate in the beginnings of a landmark case which might favour the interpretation that a "right to explanation" exists [67].

Even if such an interpretation is upheld, the transparency gap remains, as XAI would still regard the law as a justification for its own motivations. A better way to bridge the transparency gap may come from Jongepier and Keymolen [28]. Regardless of the involvement of machines, there is often a legal and/or moral right to an explanation if "choices are made which significantly affect us but which we do not understand". If such a general right exists for a given context, regardless of whether the decision maker was a machine or a human, then we can argue that the replacement of the human decision-maker by an AI must not undermine the right to an explanation, the automated process "inherits" it from the manual, human decision maker.

3.2 Explainability and the AI Act

In the GDPR, it is the right to contestation that is the end to achieve, in part via the means of transparency, opening the transparency gap. However, the uncertainty around interpretations renders this argument weak, because the motivations for XAI could be readily adapted to bridge the gap. By contrast, the AI Act [11] clarifies, broadens, and operationalises transparency requirements and their effects on the legal requirements for explainability.[2] Article 1(c) declares that the Act lays down 'harmonised *transparency rules* for AI systems intended to interact with natural persons [. . .]' (emphasis added). 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. It takes a proportional risk-based approach to defining the transparency requirements, where *high-risk* systems, such as facial recognition and law enforcement systems, are subject to greater regulation than low-risk ones.[3]

In the following sections, we describe the transparency and explainability requirements of the Act building on the work of Sovrano et al. [62]. 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. To start, the Act 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 – created prior to running the system – 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 are essential for when 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. To codify this view, Article 52(1) states that users should be made aware that they are interacting with an AI system. Furthermore, 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 – created after a decision was made – requirements on high-risk AI systems. Recitals 38–40 mention the cases of law enforcement, migration, and administration of justice, where human oversight needs to be ascertained, as biased decisions have particularly far-reaching effects in these applications.

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

[2] The Act has undergone significant changes in the EU Parliament and Council. For consistency with prior work on the Act, we use the final proposal by the EU Commission released on 21 April 2021. However, where appropriate, we will mention relevant amendments to the original proposal.

[3] See Annex II and III of the Act for a full list of systems classified as high-risk.

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 [62]. Additionally, Article 12 requires *record-keeping*, or logging, of the high-risk AI system's operation. Relatedly, the version of the Act 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' [12]. 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.

Finally, we note that the Act has also attracted a lot of criticism [68]. A detailed discussion of these criticisms is out of scope, but we mention two recurring issues that will merit further academic attention. First, many of the proposed amendments to the Act would incorporate exclusion criteria to the documentation and transparency requirements due to a tension between intellectual property rights and transparency [46, 47]. The Act is also criticised for posing overreaching compliance-oriented transparency requirements. To address this, amendments to Article 11 would allow for start-ups, small, and medium enterprises to fulfil the requirements in equivalent but less demanding forms [12]. However, this might enable conformity avoidance due to self-regulation as discussed in Section 4.3.

4 Bridging the Transparency Gap

As we saw, there are major differences in how technology and the law understand transparency and this leads to the transparency gap. XAI uses a specialist vocabulary that addresses the distinct but narrow challenge of algorithmic transparency as an end in itself to be achieved. Regulations, by contrast, consider a wider view of transparency that views it as one of many means through which wider values are supported. Without a mutual conceptual understanding of what transparency is, expert discussions (e.g., XAI literature) would lack sufficient breadth to address all concerns of society, while courts and regulators might lack the ability to assess algorithmic transparency. After all, it is up to these legal bodies, informed by expert discussions, to determine appropriate interpretations to the normative demands set by high-level laws, such as the Act, that are then given specificity in their interactions with reality. In the following, we identify and discuss four critical axes – informed by our previous discussions – along which the transparency gap may be addressed.

4.1 Scope of Transparency

In the interpretation of the Act, transparency is an overarching property of the AI system achieved through the requirements detailed in Section 3.2. Yet as we saw, the transparency gap means that XAI focuses solely on the algorithmic properties of the decision-making process. Both interpretations have different consequences on the actual design of transparent AI.

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 different levels of transparency would be required for different high-risk AI systems, but no greater level of detail is given in the Act. A proportional approach to transparency could be emphasised by focusing on stakeholders similarly to how XAI might address stakeholders (cf., Section 2.3). Nevertheless, a more feasible interpretation would at least require an understanding of the limitations of the systems concerning its "*intended purpose*", a term used throughout the Act. This will also allow the user to build an appropriate level of trust in the system, rather than over- or under-relying on it. 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 possible: a 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 complicated by what is legally meant by an *output*, discussed in Section 4.2.

In contrast, our discussion of terminology in Section 2.1 showed that the interpretation of transparency-related concepts in XAI is often too specific and technology-focused. It often ignores societal and cultural concerns, making XAI less appealing as a solution for transparency, especially because transparency should be understood not as an end in itself, but as a means to achieve a range of important, but also heterogeneous and potentially conflicting, social goods. While social and human-centred XAI [39] have long been trying to bridge the transparency gap, the uptake of these methods has been slowed down due to brittle conceptual frameworks [40], quickly changing external requirements [1], and the difficulty of subjective evaluation [55]. Besides addressing these issues, the Act will in many ways affect how the XAI systems of the future are designed, and the field should look towards the Act to find inspiration for conceptual and requirements-related clarity, not least to bridge the transparency gap.

We also suggest a hierarchical approach to XAI, which enables targeted explanations addressing the right "cognitive holes" of humans based on risk levels. On the lower levels of this hierarchy, for low-risk AI, we would have fully automated explanations from purely numerical to higher-level conceptual explanations. As a threshold is crossed, reaching high-risk systems, we would increasingly introduce human interaction to guide the output generation of XAI systems, e.g., via dialogue systems. This restores human agency by allowing people to intervene or contest decisions in extraordinary circumstances.

4.2 Uncertainty in Legal Definitions

Legal definitions in the Act leave room for flexible interpretations which have significant effects on the interpretation of the regulations and the utility of XAI for transparency. This is expected on some level given the high-level nature of the Act, but XAI and the transparency gap significantly complicates the applicability of these definitions. We illustrate this by focusing on the crucial concepts of *AI system* and its *output*, comparing how the Act and XAI approach them.

The Act's definition of an AI system is a much-debated and amended point [12, 46, 47]. Originally, Article (3)(1) defined AI systems in terms of a list of technologies, but this was later revised in the European Parliament [48] to reflect the definition by the Organisation for Economic Co-operation and Development (OECD) [45]: "a machine-based system that can, for a given set of human-defined ob-

jectives, make predictions, recommendations, or decisions influencing real or virtual environments [...] with varying levels of autonomy."

The final definition of AI has wide-ranging consequences on how transparency is achieved. Many XAI tools rely on practices that fall under the currently proposed definition of AI. Does then the XAI tool qualify as the same or as a different AI system separate from the underlying AI algorithm which they explain? We argue that the XAI tools should *not* be treated separately from the AI system they explain. This is a natural view to take as we argue for bridging the transparency gap, and in this sense, we argue against the general use of model-agnostic methods. XAI tools will need to be calibrated to work well with not just the AI system but its entire ecosystem. Taking the XAI system out of context by assessing it as a separate entity will inevitably lead to erroneous conclusions. Regulators need to clarify the legal status of XAI under the Act, making it clear that XAI systems should be assessed as part of the overall AI system, and XAI should focus on developing task-specific tools that retain domain knowledge.

In addition, the definition of an *output* of an AI system is also critical, as transparency requirements depend in part on what needs to be explained. In defining AI systems, the Act exemplifies outputs with the terms 'content, predictions, recommendations, or decisions'. However, there is no explicit definition of output. This approach is insufficient as XAI tools differentiate between internal (e.g., feature weights) and external outputs (e.g., a classification) of the AI system. While the user cannot act directly on internal outputs, XAI tools – especially ante-hoc systems seen in Section 2.2 – leverage them to produce an explanation. To bridge the transparency gap, it is important for both regulators and XAI to define the outputs they work with since applying the same requirements to different types of outputs could result in conflicting technologies and regulations.

4.3 Conformity Assessment

Another area where addressing the transparency gap is important is the methods by which AI providers are required to achieve conformity to the Act's transparency requirements. 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.

However, XAI tools introduce further complexities to the AI system which affect their conformity under the Act, often without considering the wider effects due to the transparency gap. We suggest that regulators and XAI should work towards a unified assessment scheme, where the AI system and XAI tool are considered as one unit so that the new complexities introduced by XAI would not go unnoticed or be construed as the inherent capabilities of the underlying AI system. 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, while scientists should measure the socio-technical effects of XAI tools via the involvement of human participants from various stakeholder groups.

Interestingly, the Act does not address how the actual effects of explanations on users should be accounted for. An incorrect explanation is arguably more damaging than no explanation at all, thus, explanations should be subject to stringent quality control and conformity assessment too. In bridging the transparency gap, we can use explainability fact sheets that provide a comprehensive checklist for the assessment of the correctness of XAI methods [61].

The Act, as worded, cannot provide software developers with enough detail to lead to actionable design decisions due to its high-level nature. In some cases, e.g., medical devices, *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. Though the Act does not mandate that developers adhere to industrial standards, this may be a ready way to narrow the transparency gap. While developers could interpret the requirements of the Act on their own, and decide how to implement transparency requirements, 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".

As constituted, however, these standard-setting bodies lack the expertise and remit to consider for instance Human Rights implications of their standards. We discussed at our outset how transparency speaks to a whole range of important human rights that a badly designed AI system may impact: from the right to bodily integrity to the right to non-discriminatory treatment to the right to privacy. We have also seen how multifaceted the concept of transparency is.

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's proposal may be much less efficient in protecting basic rights than, for example, the envisaged UK regulatory framework for AI [14]. 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, but they are also better placed to understand the different roles transparency plays in their respective fields.

Another problem arises 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? For our context, in particular, do changes to the XAI tool count? A possible scenario here is a system that delivers generally correct results but has an XAI component that is increasingly not state-of-the-art. If this component is updated, does the entire system need recertification? After all, *ex hypothesi* the substantial results have not changed, and the AI system still conforms to industry-standards. If the answer is yes, then this might be a deterrent to upgrading the XAI component of a certified system.

It is crucial for solving the transparency gap, that more of academia is involved in the work of not just standard-setting bodies but other regulatory bodies, to enable the work of these institutions and to understand the wider values that are at stake in misreading transparency.

4.4 Explainable Data

So far most of our discussion in XAI has focused on the algorithmic aspect of transparency. However, data-related requirements and the subsequent design choices have clear transparency-related implications and these considerations are strongly regulated in the Act. In particular, Article 10 and Recital 44 of the Act lays down comprehensive requirements for *training*, *validation*, and *testing data* used with AI systems, addressing, either directly or indirectly, many of the Human Rights implications we eluded to in the previous section.

However, the Act misses a crucial technological aspect in Article 10 by constraining its scope to training, validation, and testing

data. Datasets are indeed crucial for data-driven AI systems, however, methods such as planning and reinforcement learning do not rely on the same techniques as supervised learning from where these terms originate [5]. It is unclear what requirements should be fulfilled for XAI systems that rely on the input data of such systems to generate their explanations. More comprehensive coverage of data types is necessary to cover a wider range of algorithms.

Further indicative of the transparency gap is that XAI has a long way to go in addressing data-oriented concerns. Automated *data explanation* techniques could examine the effects of inherent properties of the input dataset on the AI system, for example by purposefully biasing, distorting, or introducing irrelevant samples to the dataset. Such methods could identify issues with data entanglement as well – a result of mixing multiple sources of data. Data explanations may uncover problems with the dataset and the system itself and, at the same time, conformity to Article 10(2) of the Act could be demonstrated.

Another promising direction in practical data governance is data documentation. Gebru et al. [16] 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 AI transparency in a way that is compatible with the Act, bridging the transparency gap.

5 Conclusion

The transparency gap presents the very real risk that the differing views of law and computer science on transparency will prevent the Act and XAI from achieving a beneficial impact. This could happen in a number of ways: one is identifying the wider legal concept of transparency with the much narrower computer science concept, allowing the technological discourse to replace the socio-legal one. This would leave a range of wider transparency harms unaddressed.

The other danger is that legislators overestimate the capabilities of XAI and as a result overload the regulation with unachievable, highly detailed design prescriptions that will often opt to deliver benefits for users and affected third parties, not necessarily because they fail in generating transparency, but because the legal framework does not then give the affected parties the tools to act on what they learned.

In the first scenario, we lower our expectations of the law too much, in the second we raise our expectations of computer science too high. Instead of computer science colonising law, or vice versa, we suggested four approaches that respect the internal logic of the two systems and ask XAI not to "solve" the problem of legal transparency, but to understand how, given the internal logic of the law, it can develop new tools and approaches to facilitate transparency while staying true to its foundations at the same time.

Acknowledgements

This work was supported in part by the UKRI Centre for Doctoral Training in Natural Language Processing (grant EP/S022481/1) and the UKRI Trustworthy Autonomous Systems Node in Governance and Regulation (grant EP/V026607/1).

References

[1] Sajid Ali, Tamer Abuhmed, Shaker El-Sappagh, Khan Muhammad, Jose M. Alonso-Moral, Roberto Confalonieri, Riccardo Guidotti, Javier Del Ser, Natalia Díaz-Rodríguez, and Francisco Herrera, 'Explainable Artificial Intelligence (XAI): What we know and what is left to attain Trustworthy Artificial Intelligence', *Information Fusion*, 101805, (April 2023).

[2] Plamen P. Angelov, Eduardo A. Soares, Richard Jiang, Nicholas I. Arnold, and Peter M. Atkinson, 'Explainable artificial intelligence: An analytical review', *WIREs Data Mining and Knowledge Discovery*, **11**(5), e1424, (2021).

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

[4] Emre Bayamlıoğlu, 'The right to contest automated decisions under the General Data Protection Regulation: Beyond the so-called "right to explanation"', *Regulation & Governance*, **16**(4), 1058–1078, (2022).

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

[6] Nadia Burkart and Marco F. Huber, 'A Survey on the Explainability of Supervised Machine Learning', *Journal of Artificial Intelligence Research*, **70**, 245–317, (May 2021).

[7] Oana-Maria Camburu, Tim Rocktäschel, Thomas Lukasiewicz, and Phil Blunsom, 'E-SNLI: Natural Language Inference with Natural Language Explanations', in *Advances in Neural Information Processing Systems*, volume 31. Curran Associates, Inc., (2018).

[8] Diogo V. Carvalho, Eduardo M. Pereira, and Jaime S. Cardoso, 'Machine Learning Interpretability: A Survey on Methods and Metrics', *Electronics*, **8**(8), 832, (August 2019).

[9] Bryan Casey, Ashkon Farhangi, and Roland Vogl. Rethinking Explainable Machines: The GDPR's 'Right to Explanation' Debate and the Rise of Algorithmic Audits in Enterprise, February 2018.

[10] Lei Chen, Yu Qiu, Junan Zhao, Jing Xu, and Ao Liu, 'CPKD: Concepts-Prober-Guided Knowledge Distillation for Fine-Grained CNN Explanation', in *2021 2nd International Conference on Electronics, Communications and Information Technology (CECIT)*, pp. 421–426, (December 2021).

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

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

[13] Tiffany Curtiss, 'Computer Fraud and Abuse Act Enforcement: Cruel, Unusual, and Due for Reform', *Washington Law Review*, **91**(4), 1813, (December 2016).

[14] Nadine Dorries, 'Establishing a pro-innovation approach to regulating AI', Technical report, Department for Digital, Culture, Media and Sport, London, (July 2022).

[15] Lilian Edwards and Michael Veale, 'Enslaving the Algorithm: From a "Right to an Explanation" to a "Right to Better Decisions"?', *IEEE Security & Privacy*, **16**(3), 46–54, (May 2018).

[16] Timnit Gebru, Jamie Morgenstern, Briana Vecchione, Jennifer Wortman Vaughan, Hanna Wallach, Hal Daumé III, and Kate Crawford, 'Datasheets for datasets', *Communications of the ACM*, **64**(12), 86–92, (November 2021).

[17] Bryce Goodman and Seth Flaxman, 'European Union Regulations on Algorithmic Decision-Making and a "Right to Explanation"', *AI Magazine*, **38**(3), 50–57, (October 2017).

[18] Riccardo Guidotti, Anna Monreale, Salvatore Ruggieri, Franco Turini, Fosca Giannotti, and Dino Pedreschi, 'A Survey of Methods for Explaining Black Box Models', *ACM Computing Surveys*, **51**(5), 93:1–93:42, (August 2018).

[19] Balint Gyevnar, Cheng Wang, Christopher G. Lucas, Shay B. Cohen, and Stefano V. Albrecht, 'Causal Explanations for Stochastic Sequential Multi-Agent Decision-Making', in *Explainable and Transparent AI and Multi-Agent Systems*, London, (May 2023).

[20] Philipp Hacker and Jan-Hendrik Passoth, 'Varieties of AI Explanations Under the Law. From the GDPR to the AIA, and Beyond', in *xxAI - Beyond Explainable AI: International Workshop, Held in Conjunction with ICML 2020, July 18, 2020, Vienna, Austria, Revised and Extended Papers*, eds., Andreas Holzinger, Randy Goebel, Ruth Fong, Taesup Moon, Klaus-Robert Müller, and Wojciech Samek, Lecture Notes in Computer Science, 343–373, Springer International Publishing, Cham, (2022).

[21] Herbert Lionel Adolphus Hart and Tony Honoré, *Causation in the Law*, Oxford University Press UK, 1959.

[22] Mireille Hildebrandt, 'The Dawn of a Critical Transparency Right for the Profiling Era', in *Digital Enlightenment Yearbook 2012*, 41–56, IOS Press, (2012).

[23] Mireille Hildebrandt, *Law for Computer Scientists and Other Folk*, Oxford University Press, Oxford, United Kingdom, first edition edn., 2020.

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

[25] Robert R. Hoffman, Shane T. Mueller, Gary Klein, and Jordan Litman, 'Metrics for Explainable AI: Challenges and Prospects', *arXiv:1812.04608 [cs]*, (February 2019).

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

[27] Deborah G. Johnson and Mario Verdicchio, 'Reframing AI Discourse', *Minds and Machines*, **27**(4), 575–590, (December 2017).

[28] Fleur Jongepier and Esther Keymolen, 'Explanation and Agency: Exploring the normative-epistemic landscape of the "Right to Explanation"', *Ethics and Information Technology*, **24**(4), 49, (November 2022).

[29] Margot E. Kaminski. The Right to Explanation, Explained, June 2018.

[30] Davinder Kaur, Suleyman Uslu, Kaley J. Rittichier, and Arjan Durresi, 'Trustworthy Artificial Intelligence: A Review', *ACM Computing Surveys*, **55**(2), 39:1–39:38, (January 2022).

[31] Patrick Gage Kelley, Yongwei Yang, Courtney Heldreth, Christopher Moessner, Aaron Sedley, Andreas Kramm, David T. Newman, and Allison Woodruff, 'Exciting, Useful, Worrying, Futuristic: Public Perception of Artificial Intelligence in 8 Countries', in *Proceedings of the 2021 AAAI/ACM Conference on AI, Ethics, and Society*, AIES '21, pp. 627–637. Association for Computing Machinery, (July 2021).

[32] LAION.ai. A Call to Protect Open-Source AI in Europe. https://laion.ai/notes/letter-to-the-eu-parliament, April 2023.

[33] Markus Langer, Daniel Oster, Timo Speith, Holger Hermanns, Lena Kästner, Eva Schmidt, Andreas Sesing, and Kevin Baum, '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**, 103473, (July 2021).

[34] Johann Laux, Sandra Wachter, and Brent Mittelstadt, 'Trustworthy artificial intelligence and the European Union AI act: On the conflation of trustworthiness and acceptability of risk', *Regulation & Governance*, (June 2023).

[35] Q. Vera Liao, Yunfeng Zhang, Ronny Luss, Finale Doshi-Velez, and Amit Dhurandhar. Connecting Algorithmic Research and Usage Contexts: A Perspective of Contextualized Evaluation for Explainable AI, June 2022.

[36] Bodhisattwa Prasad Majumder, Oana-Maria Camburu, Thomas Lukasiewicz, and Julian McAuley. Knowledge-Grounded Self-Rationalization via Extractive and Natural Language Explanations, September 2022.

[37] Gianclaudio Malgieri and Giovanni Comandé, 'Why a Right to Legibility of Automated Decision-Making Exists in the General Data Protection Regulation', *International Data Privacy Law*, **7**(4), 243–265, (November 2017).

[38] Bertram F. Malle, 'How People Explain Behavior: A New Theoretical Framework', *Personality and Social Psychology Review*, **3**(1), 23–48, (February 1999).

[39] Tim Miller, 'Explanation in artificial intelligence: Insights from the social sciences', *Artificial Intelligence*, **267**, 1–38, (February 2019).

[40] Tim Miller, 'Explainable AI is Dead, Long Live Explainable AI! Hypothesis-driven Decision Support using Evaluative AI', in *Proceedings of the 2023 ACM Conference on Fairness, Accountability, and Transparency*, FAccT '23, pp. 333–342, New York, NY, USA, (June 2023). Association for Computing Machinery.

[41] Brent Mittelstadt, Chris Russell, and Sandra Wachter, 'Explaining Explanations in AI', in *Proceedings of the Conference on Fairness, Accountability, and Transparency*, pp. 279–288, (January 2019).

[42] Sina Mohseni, Niloofar Zarei, and Eric D. Ragan, 'A Multidisciplinary Survey and Framework for Design and Evaluation of Explainable AI Systems', *ACM Transactions on Interactive Intelligent Systems*, **11**(3-4), 24:1–24:45, (August 2021).

[43] Christoph Molnar, *Interpretable Machine Learning*, February 2023.

[44] Luca Nannini, Agathe Balayn, and Adam Leon Smith. Explainability in AI Policies: A Critical Review of Communications, Reports, Regulations, and Standards in the EU, US, and UK, April 2023.

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

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

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

[48] European Parliament. DRAFT Compromise Amendments on the Draft Report Proposal for a regulation of the European Parliament and of the Council on harmonised rules on Artificial Intelligence (Artificial Intelligence Act) and amending certain Union Legislative Acts (COM(2021)0206 – C9 0146/2021 – 2021/0106(COD)), May 2023.

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

[50] Article 29 Data Protection Working Party. Guidelines on Automated individual decision-making and Profiling for the purposes of Regulation 2016/679 (WP251rev.01), February 2018.

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

[52] Sarvapali D. Ramchurn, Sebastian Stein, and Nicholas R. Jennings, 'Trustworthy human-AI partnerships', *iScience*, **24**(8), 102891, (August 2021).

[53] Xavier Renard, Thibault Laugel, and Marcin Detyniecki. Understanding Prediction Discrepancies in Machine Learning Classifiers, April 2021.

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

[55] Yao Rong, Tobias Leemann, Thai-trang Nguyen, Lisa Fiedler, Tina Seidel, Gjergji Kasneci, and Enkelejda Kasneci. Towards Human-centered Explainable AI: User Studies for Model Explanations, October 2022.

[56] Cynthia Rudin, Chaofan Chen, Zhi Chen, Haiyang Huang, Lesia Semenova, and Chudi Zhong, 'Interpretable machine learning: Fundamental principles and 10 grand challenges', *Statistics Surveys*, **16**(none), 1–85, (January 2022).

[57] Stuart Russel and Peter Norvig, *Artificial Intelligence: A Modern Approach*, Pearson, fourth edn., October 2022.

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

[59] Jakob Schoeffer, Niklas Kuehl, and Yvette Machowski, '"There Is Not Enough Information": On the Effects of Explanations on Perceptions of Informational Fairness and Trustworthiness in Automated Decision-Making', in *2022 ACM Conference on Fairness, Accountability, and Transparency*, pp. 1616–1628, (June 2022).

[60] Gesina Schwalbe and Bettina Finzel. A Comprehensive Taxonomy for Explainable Artificial Intelligence: A Systematic Survey of Surveys on Methods and Concepts, May 2022.

[61] Kacper Sokol and Peter Flach, 'Explainability Fact Sheets: A Framework for Systematic Assessment of Explainable Approaches', in *Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency*, pp. 56–67, (January 2020).

[62] Francesco Sovrano, Salvatore Sapienza, Monica Palmirani, and Fabio Vitali, 'Metrics, Explainability and the European AI Act Proposal', *J*, **5**(1), 126–138, (March 2022).

[63] Stanford. AI Index Report 2023 – Artificial Intelligence Index. https://aiindex.stanford.edu/report/, 2023.

[64] Ilia Stepin, Jose M. Alonso, Alejandro Catala, and Martín Pereira-Fariña, 'A Survey of Contrastive and Counterfactual Explanation Generation Methods for Explainable Artificial Intelligence', *IEEE Access*, **9**, 11974–12001, (2021).

[65] Gunther Teubner, *Law as an Autopoietic System*, Oxford/Cambridge, Blackwell Publishers, 1993.

[66] Gunther Teubner, 'Breaking Frames: The Global Interplay of Legal and Social Systems', *The American Journal of Comparative Law*, **45**(1), 149–169, (January 1997).

[67] The Court of Justice of The Eurpean Union. Request for a preliminary ruling from the Verwaltungsgericht Wiesbaden (Germany) – OQ v Land Hesse (Case C-634/21).

[68] Michael Veale and Frederik Zuiderveen Borgesius, 'Demystifying the Draft EU Artificial Intelligence Act', *Computer Law Review International*, **22**(4), 97–112, (August 2021).

[69] Sandra Wachter, Brent Mittelstadt, and Luciano Floridi, 'Why a Right to Explanation of Automated Decision-Making Does Not Exist in the General Data Protection Regulation', *International Data Privacy Law*, **7**(2), 76–99, (June 2017).

[70] Michael Winikoff and Julija Sardelić, 'Artificial Intelligence and the Right to Explanation as a Human Right', *IEEE Internet Computing*, **25**(2), 116–120, (March 2021).