Varieties of AI Explanations under the Law. From the GDPR to the AIA, and beyond

Philipp Hacker[1] and Jan-Hendrik Passoth[2]

[1] Chair for Law and Ethics of the Digital Society, European New School of Digital Studies, European University Viadrina, Große Scharnstraße 59, 15230 Frankfurt / Oder, Germany
hacker@europa-uni.de
https://www.europeannewschool.eu/law-ethics.html
[2] Chair of Sociology of Technology, European New School of Digital Studies, European University Viadrina, Große Scharnstraße 59, 15230 Frankfurt / Oder, Germany
passoth@europa-uni.de
https://www.europeannewschool.eu/sociology-of-technology.html

Abstract. The quest to explain the output of artificial intelligence systems has clearly moved from a mere technical to a highly legally and politically relevant endeavor. In this paper, we provide an overview of legal obligations to explain AI and evaluate current policy proposals. In this, we distinguish between different functional varieties of AI explanations – such as multiple forms of enabling, technical and protective transparency – and show how different legal areas engage with and mandate such different types of explanations to varying degrees. Starting with the rights-enabling framework of the GDPR, we proceed to uncover technical and protective forms of explanations owed under contract, tort and banking law. Moreover, we discuss what the recent EU proposal for an Artificial Intelligence Act means for explainable AI, and review the proposal's strengths and limitations in this respect. Finally, from a policy perspective, we advocate for moving beyond mere explainability towards a more encompassing framework for trustworthy and responsible AI that includes actionable explanations, values-in-design and co-design methodologies, interactions with algorithmic fairness, and quality benchmarking.

Keywords: artificial intelligence · explainability · regulation

1 Introduction

Sunlight is the best disinfectant, as the saying goes. Therefore, it does not come as a surprise that transparency constitutes a key societal desideratum vis-à-vis complex, modern IT systems in general [67] and artificial intelligence (AI) in particular [18, 74]. As in the case of very similar demands concerning other forms of opaque or, at least from an outsider perspective, inscrutable decision making processes of bureaucratic systems, transparency is seen as a means of

making decisions more understandable, more contestable, or at least more rational. More specifically, explainability of AI systems generally denotes the degree to which an observer may understand the causes of the system's output [64, 15]. Various technical implementations of explainability have been suggested, from truth maintenance systems for causal reasoning in the case of symbolic reasoning systems that were developed mainly from the 1970s to the 1990s to layerwise relevance propagation methods for neural networks today. Importantly, observers, and with them the adequate explanations for a specific context, may vary [3, p. 85].

In recent years, the quest for transparent and explainable AI has not only spurred a vast array of research efforts in machine learning [82, 3, and the chapters in this volume for an overview], but it has also emerged at the heart of many ethics and responsible design proposals [43, 66, 68, 45] and has nurtured a vivid debate on the promises and limitations of advanced machine learning models for various high-stakes scenarios [37, 12, 88].

1.1 Functional Varieties of AI Explanations

Importantly, from a normative perspective, different arguments can be advanced to justify the need for transparency in AI systems [3]. For example, given its relation to human autonomy and dignity, one may advance a 'deontological' conception viewing transparency as an aim in itself [104, 17, 92]. Moreover, research suggests that explanations may satisfy the curiosity of counterparties, their desire for learning or control, or fulfill basic communicative standards of dialogue and exchange [64, 59, 62]. From a legal perspective, however, it is submitted that three major functional justifications for demands of AI explainability may be distinguished: enabling, technical, and protective varieties. All of them subscribe to an 'instrumentalist' approach conceiving of transparency as a means to achieve technically or normatively desirable ends.

First, explainability of AI is seen as a prerequisite for empowering those affected by its decisions or charged with reviewing them ('enabling transparency'). On the one hand, explanations are deemed crucial to afford due process to the affected individuals [23] and to enable them to effectively exercise their subjective rights vis-à-vis the (operators of the) AI system [89] ('rights-enabling transparency'). Similarly, other parties such as NGOs, collective redress organizations or supervisory authorities may use explanations to initiate legal reviews, e.g. by inspecting AI systems for unlawful behavior such as manipulation or discrimination [37, p. 55]('review-enabling transparency'). On the other hand, information about the functioning of AI systems may facilitate informed choice of the affected persons about whether and how to engage with the models or the offers they accompany and condition. Such 'decision-enabling transparency' seeks to support effective market choice, for example by switching contracting partners [14, p. 156].

Second, with respect to technical functionality, explainability may help fine-tune the performance (e.g., accuracy) of the system in real-world scenarios and evaluate its generalizability to unseen data [79, 47, 57, 3]. In this vein, it also

acts as a catalyst for informed decision making, though not of the affected persons, but rather of the technical operator or an expert auditor of the system. That approach may hence be termed 'technical transparency', its explanations being geared toward a technically sophisticated audience. Beyond model improvements, a key aim here is to generate operational and institutional trust in the AI system [37, p. 54], both in the organization operating the AI system and beyond in the case of third-party reviews and audits.

Third, technical improvements translate into legal relevance to the extent that they contribute to reducing normatively significant risks. Hence, technically superior performance may lead to improved safety (e.g., AI in robots; medical AI), reduced misallocation of resources (e.g., planning and logistics tools), or better control of systemic risks (e.g., financial risk modelling). This third variety could be dubbed 'protective transparency', as it seeks to harness explanations to guard against legally relevant risks.

These different types of legally relevant, functional varieties of AI explanations are not mutually exclusive. For example, technical explanations may, to the extent available, also be used by collective redress organizations or supervisory authorities in a review-enabling way. Nonetheless, the distinctions arguably provide helpful analytical starting points. As we shall see, legal provisions compelling transparency are responsive to these different strands of justification to varying degrees. It should not be overlooked, however, that an excess of sunlight can be detrimental as well, as skeptics note: explainability requirements may not only impose significant and sometimes perhaps prohibitive burdens on the use of some of the most powerful AI systems, but also offer affected persons the option to strategically "game the system" and accrue undeserved advantages [9]. This puts differentiated forms of accountability front and center: to whom - users, affected persons, professional audit experts, legitimized rights protection organizations, public authorities - should an AI system be transparent? Such limitations need to be considered by the regulatory framework as well.

1.2 Technical Varieties of AI Explanations

From a technical perspective, in turn, it seems uncontroversial that statements about AI and explainability, as well as the potential trade-off with accuracy, must be made in a context- and model-specific way [57, 81][3, p. 100]. While some types of ML models, such as linear or logistic regressions or small decision trees [47, 57, 22], lend themselves rather naturally to global explanations about the feature weights for the entire model (often called ex ante interpretability), such globally valid statements are much harder to obtain for other model types, particularly random forests or deep neural networks [90, 79, 57]. In recent years, such complex model types have been the subject of intense technical research to provide for, at the minimum, local explanations of specific decisions ex post, often by way of sensitivity analysis [79, 60, 31]. One specific variety of local explanations seeks to provide counterfactuals, i.e., suggestions for minimal changes of the input data to achieve a more desired output [97, 64]. Counterfactuals are

a variety of contrastive explanations, which seek to convey reasons for the concrete output ('fact') in relation to another, possible output ('foil') and which have recently gained significant momentum [77, 65]. Other methods have sought to combine large numbers of local explanations to approximate a global explanatory model of the AI system by way of overall feature relevance [55, 16], while other scholars have sought to fiercely defend the benefits of designing models that are interpretable ex ante rather than explainable ex post [81].

1.3 Roadmap of the Paper

Arguably, much of this research has been driven, at least implicitly, by the assumption that explainable AI systems would be ethically desirable and perhaps even legally required [47]. Hence, this paper seeks to provide an overview of explainability obligations flowing from the law proper, while engaging with the functional and technical distinctions just introduced. The contemporary legal debate has its roots in an interpretive battle over specific norms of the GDPR [96, 89], but has recently expanded beyond the precincts of data protection law to other legal fields, such as contract and tort law [42, 84]. As this paper will show, another important yet often overlooked area which might engender incentives to provide explanations for AI models is banking law [54]. Finally, the question of transparency has recently been taken up very prominently by the regulatory proposals at the EU level, particularly in the Commission proposal for an Artificial Intelligence Act (AIA). It should be noted that controversies and consultations about how to meaningfully regulate AI systems are still ongoing processes and that the questions of what kind of explainability obligations follow already from existing regulations and which obligations should - in the future - become part of AI policy are still very much in flux. This begs the question of the extent to which these diverging provisions and calls for explainability properly take into account the usability of that information for the recipients, in other words: the actionability of explainable AI (XXAI), which is also at the core of this volume.

Against this background, the paper will use the running example of credit scoring to investigate whether positive law mandates, or at least sets incentives for, the provision of actionable explanations in the use of AI tools, particularly in settings involving private actors (Section 2); to what extent the proposals for AI regulation at the EU level will change these findings (Section 3); and how regulation and practice could go beyond such provisions to ensure actionable explanations and trustworthy AI (Section 4). In all of these sections, the findings will be linked to the different (instrumentalist) functions of transparency, which are taken up to varying degrees by the different provisions and proposals. Figure 1 below provides a quick overview of the relations between functions and several existing legal acts surveyed in this paper; Figure 2 (in Section 3) connects these functions to the provisions of the planned AIA.

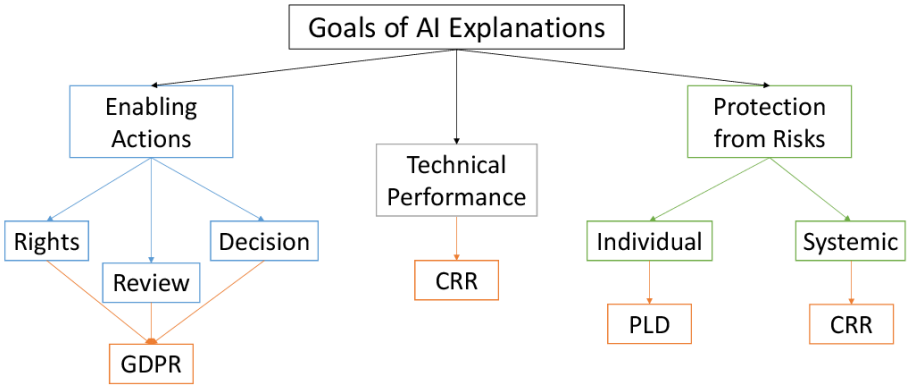

Figure 1: Overview of the functions of different EU law instruments concerning AI explanations; abbreviations: GDPR: General Data Protection Regulation; CRR: Capital Requirements Regulation; PLD: Product Liability Directive

2 Explainable AI under Current Law

The quest for explainable AI interacts with existing law in a number of ways. The scope of this paper will be EU law, and for the greatest part the law governing exchange between private parties more particularly (for public law, see, e.g.[14, 2.2]). Most importantly, and bridging the public-privates divide, the GDPR contains certain rules, however limited and vague, which might be understood as an obligation to provide explanations of the functioning of AI models (Section 2.1.). Beyond data protection law, however, contract and tort law (Section 2.2) and banking law (Section 2.3) also provide significant incentives for the use of explainable AI (XAI).

2.1 The GDPR: Rights-Enabling Transparency

In the GDPR, whether a subjective right to an explanation of AI decisions exists or not has been the object of a long-standing scholarly debate which, until this day, has not been finally settled [96, 36, 89, 61]. To appreciate the different perspectives, let us consider the example of AI-based credit scoring. Increasingly, startups use alternative data sets and machine learning to compute credit scores, which in turn form the basis of lending decisions (see, e.g., [54, 34]). If a particular person receives a specific credit score, the question arises if, under the GDPR, the candidate may claim access to the feature values used to make the prediction, to the weights of the specific features in his or her case (local explanation), or even to the weights of the features in the model more generally (global explanation). For example, the person might want to know what concrete age and income

values were used to predict the score, to what extent age or income contributed to the prediction in the specific case, and how the model generally weights these features.

So far, there is no guidance by the Court of Justice of the European Union (CJEU) on precisely this question. However, exactly this case was decided by the German Federal Court for Private Law (BGH) in 2014 (BGH, Case VI ZR 156/13 = MMR 2014, 489). The ruling came down not under the GDPR, but its predecessor (the 1995 Data Protection Directive) and relevant German data protection law. In substance, however, the BGH noted that the individual information interest of the plaintiff needed to be balanced against the legitimate interests of the German credit scoring agency (Schufa) to keep its trade secrets, such as the precise score formula for credit scoring, hidden from the view of the public, lest competitors free ride on its know-how. In weighing these opposing interests, the BGH concluded that the plaintiff did have a right to access its personal data processed for obtaining the credit score (the feature values), but not to obtain information on the score formula itself, comparison groups, or abstract methods of calculation. Hence, the plaintiff was barred from receiving either a local or a global explanation of its credit score.

2.1.1 Safeguards for Automated Decision Making How would such a case be decided under the GDPR, particularly if an AI-based scoring system was used? There are two main normative anchors in the GDPR that could be used to obtain an explanation of the score, and hence more generally of the output of an AI system. First, Article 22 GDPR regulates the use of automated decision making in individual cases. That provision, however, is subject to several significant limitations. Not only does its wording suggest that it applies only to purely automated decisions, taken independently of even negligible human interventions (a limitation that could potentially be overcome by a more expansive interpretation of the provision, see [96]); more importantly, the safeguards it installs in Article 22(3) GDPR for cases of automated decision making list 'the right to obtain human intervention on the part of the controller, to express his or her point of view and to contest the decision' – but not the right to an explanation. Rather, such a right is only mentioned in Recital 71 GDPR, which provides additional interpretive guidance for Article 22(3) GDPR. Since, however, only the Articles of the regulation, not the recitals, constitute binding law, many scholars are rightly skeptical whether the CJEU would deduce a right to an explanation (of whatever kind) directly from Article 22(3) GDPR [96, 84].

2.1.2 Meaningful information about the logic involved A second, much more promising route is offered by different provisions obliging the data controller (i.e., the operator of the AI system) to provide the data subject not only with information on the personal data processed (the feature values), but also, at least in cases of automated decision making, with 'meaningful information about the logic involved' (Art. 13(2)(f), Art. 14(2)(g), Art. 15(1)(h) GDPR).

A rights-enabling conception of meaningful information Since the publication of the GDPR, scholars have intensely debated what these provisions mean for AI systems (see, e.g. for overviews [20, 49]. For instance, in our running example, we may more concretely ask whether a duty to disclose local or global weights of specific features exists in the case of credit scoring. Some scholars stress the reference to the concept of 'logic', which to them suggests that only the general architecture of the system must be divulged, but not more specific information on features and weights [73, para. 31c][103]. A more convincing interpretation, in our view, would take the purpose of the mentioned provisions into account. Hence, from a teleological perspective, the right to meaningful information needs to be read in conjunction with the individual rights the GDPR confers in Art. 16 et seqq. [89]. Such a rights-enabling instrumentalist approach implies that information will only be meaningful, to the data subject, if it facilitates the exercise of these rights, for example the right to erasure, correction, restriction of processing or, perhaps most importantly, the contestation of the decision pursuant to Article 22(3) GDPR. An overarching view of the disclosure provisions forcing meaningful information and the safeguards in Article 22(3) GDPR therefore suggests that, already under current data protection law, the information provided must be actionable to fulfill its enabling function. Importantly, this directly relates to the quest of XXAI research seeking to provide explanations that enable recipients to meaningfully reflect upon and intervene in AI-powered decision-making systems.

Hence, in our view, more concrete explanations may have to be provided if information about the individual features and corresponding weights are necessary to formulate substantive challenges to the algorithmic scores under the GDPR's correction, erasure or contestation rights. Nevertheless, as Article 15(4) GDPR and more generally Article 16 of the Charter of Fundamental Rights of the EU (freedom to conduct the business) suggest, the information interests of the data subject must still be balanced against the secrecy interests of the controller, and their interest in protecting the integrity of scores against strategic gaming. In this reading, a duty to provide actionable yet proportionate information follows from Art. 13(2)(f), Art. 14(2)(g) and Art. 15(1)(h) GDPR, read in conjunction with the other individual rights of the data subject.

Application to credit scores In the case of AI-based credit scores, such a regime may be applied as follows. In our view, meaningful information will generally imply a duty to provide local explanations of individual cases, i.e., the disclosure of at least the most important features that contributed to the specific credit score of the applicant. This seems to be in line with the (non-binding) interpretation of European privacy regulators (Article 29 Data Protection Working Party, 2018, at 25-26). Such information is highly useful for individuals when exercising the mentioned rights and particularly for contesting the decision: if, for example, it turns out that the most important features do not seem to be related in any plausible way to creditworthiness or happen to be closely correlated with attributes protected under non-discrimination law, the data subject will be in a much better position to contest the decision in a substantiated way. Further-

more, if only local information is provided, trade secrets are implicated to a much lesser extent than if the entire score formula was disclosed; and possibilities to 'game the system' are significantly reduced. Finally, such local explanations can increasingly be provided even for complex models, such as deep neural networks, without loss of accuracy [79, 31].

On the other hand, meaningful information will generally not demand the disclosure of global explanations, i.e., of weights referring to the entire model. While this might be useful for individual complainants to detect, for example, whether their case represents an outlier (i.e., features were weighted differently in the individual case than generally in the model), the marginal benefit of a global explanation vis-à-vis a local explanation seems outweighed by the much more significant impact on trade secrets and incentives to innovation if weights for an entire model need to be disclosed. Importantly, such a duty to provide global explanations would also significantly hamper the use of more complex models, such as deep neural networks (cf. [14, p. 162]. While such technical limitations do not generally speak against certain interpretations of the law (see, e.g., BVerfG NJW 1979, 359, para. 109 – Kalkar), they seem relevant here because such models may, in a number of cases, perform better in the task of credit scoring than simpler but globally explainable models. If this premise holds, another provision of EU law becomes relevant. More accurate models allow to fulfill the requirements of *responsible lending* to a better extent (see Section 2.3 for details): if models more correctly predict creditworthiness, loans will be handed out more often only to persons who are indeed likely to repay the loan. Since this is a core requirement of the post-financial crisis framework of EU credit law, it should be taken into account in the interpretation of the GDPR in cases of credit scoring as well (see, for such overarching interpretations of different areas of EU law, CJEU, Case C-109/17, Bankia, para. 49; [38]).

Ultimately, for local and global explanations alike, a compromise between information interests and trade secrets might require the disclosure of weights not in a highly granular, but in a 'noisy' fashion (e.g., providing relevance intervals instead of specific percentage numbers) [6, para. 54]. Less mathematically trained persons often disregard or have trouble cognitively processing probability information in explanations [64] so that the effective information loss for recipients would likely be limited. Noisy weights, or simple ordinal feature ranking by importance, would arguably convey a measure enabling meaningful evaluation and critique while safeguarding more precise information relevant for the competitive advantage of the developer of the AI system, and hence for incentives to innovation. Such less granular information could be provided whenever the confidentiality of the information is not guaranteed; if the information is treated confidentially, for example in the framework of a specific procedure in a review or audit, more precise information might be provided without raising concerns about unfair competition. The last word on these matters will, of course, have the CJEU. It seems not unlikely, though, that the Court would be open to an interpretation guaranteeing actionable yet proportionate information. This would correspond to a welcome reading of the provisions of the GDPR with a view to

due process and the exercise of subjective rights by data subjects (rights-enabling transparency).

2.2 Contract and Tort Law: Technical and Protective Transparency

In data protection law, as the preceding section has shown, much will depend on the exact interpretation of the vague provisions of the GDPR, and on the extent to which these provisions can be applied even if humans interact with AI systems in more integrated forms of decision making. These limitations should lead us to consider incentives for actionable AI explanations in other fields of the law, such as contract and tort law. This involves particularly product liability (Section 2.2.1), and general negligence standards under contract and tort law (Section 2.2.2). Clearly, under freedom of contract, parties may generally contract for specific explanations that the provider of an AI system may have to enable. In the absence of such explicit contractual clauses, however, the question arises to what extent contract and tort law still compel actionable explanations. As we shall see, in these areas, the enabling instrumentalist variety of transparency (due process, exercise of rights) is to a great extent replaced by a more technical and protective instrumentalist approach focusing on trade-offs with accuracy and safety.

2.2.1 Product liability In product liability law, the first persevering problem is the extent to which it applies to non-tangible goods such as software. Article 2 of the EU Product Liability Directive (PLD), passed in 1985, defines a product as any movable, as well as electricity. While an AI system embedded in a physical component, such as a robot, clearly qualifies as a product under Article 2, this is highly contested for a standalone system such as, potentially, a credit scoring application (see [99, 84]). In the end, at least for professionally manufactured software, one will have to concede that it exhibits defect risks similar to traditional products and entails similar difficulties for plaintiffs in proving them, which speaks strongly in favor of applying the PLD, at least by analogy, to such software independently of any embeddedness in a movable component [29, p. 43]. A proposal by the EU Commission on that question, and on liability for AI more generally, is expected for 2022.

Design defects As it currently stands, the PLD addresses producers by providing those harmed by defective products with a claim against them (Art. 1 PLD). There are different types of defects a product may exhibit, the most important in the context of AI being a design defect. With respect to the topic of this paper, one may therefore ask if the lack of an explanation might qualify as a design defect of an AI system. This chiefly depends on the interpretation of the concept of a design defect.

In EU law, two rivaling interpretations exist: the consumer expectations test and the risk-utility test. Article 6 PLD at first glance seems to enshrine the former variety by holding that a 'product is defective when it does not provide

the safety which a person is entitled to expect'. The general problem with this formulation is that it is all but impossible to objectively quantify legitimate consumer expectations [99]. For example, would the operator of an AI system, the affected person, or the public in general be entitled to expect explanations, and if so, which ones?

Product safety law is often understood to provide minimum standards in this respect [100, para. 33]; however, exact obligations on explainability of AI are lacking so far in this area, too (but see Annex I, Point 1.7.4.2.(e) of the Machinery Directive 2006/42 and Section 3). Precisely because of these uncertainties, many scholars prefer the risk-utility test which has a long-standing tradition in US product liability law (see § 402A Restatement (Second) of Torts). Importantly, it is increasingly used in EU law as well [86][99, n. 48] and was endorsed by the BGH in its 2009 Airbag decision[1]. Under this interpretation, a design defect is present if the cost of a workable alternative design, in terms of development and potential reduced utility, is smaller than the gain in safety through this alternative design. Hence, the actually used product and the workable alternative product must be compared considering their respective utilities and their risks [94, p. p. 246].

With respect to XAI, it must hence be asked if an interpretable tool would have provided additional safety through the explanation, and if that marginal benefit is not outweighed by additional costs. Such an analysis, arguably, aligns with a technical and protective instrumentalist conception of transparency, as a means to achieve safety gains. Importantly, therefore, the analysis turns not only on the monetary costs of adding explanations to otherwise opaque AI systems, but it must also consider whether risks are really reduced by the provision of an explanation.

The application of the risk-utility test to explainability obligations has, to our knowledge, not been thoroughly discussed in the literature yet (for more general discussions, see [87, p. 1341, 1375][42]. Clearly, XAI may be *helpful*, in evidentiary terms, for producers in showing that there was no design defect involved in an accident [19, p. 624][105, p. 217]; but is XAI *compulsory* under the test? The distinguishing characteristic of applying a risk-utility test to explainable AI seems to be that the alternative (introducing explainability) does not necessarily reduce risk overall: while explanations plausibly lower the risk of misapplication of the AI system, they might come at the expense of accuracy. Therefore, in our view, the following two cases must be distinguished:

1. The explainable model exhibits the same accuracy as the original, non-explainable model (e.g., ex post local explanation of a DNN). In that case, only the expected gain in safety, from including explanations, must be weighed against potential costs of including explanations, such as longer run time, development costs, license fees etc. Importantly, as the BGH specified in its Airbag ruling, the alternative model need not only be factually ready for use, but its use must also be normatively reasonable and appropriate for

[1] BGH, 16.6.2009, VI ZR 107/08, BGHZ 181, 253 para 18.

the producer[2]. This implies that, arguably, trade secrets must be considered in the analysis, as well. Therefore, it seems sensible to assume that, as in data protection law, a locally (but not a globally) explainable model must be chosen, unless the explainable add-on is unreasonably expensive. Notably, the more actionable explanations are in the sense of delivering clear cues for operators, or affected persons, to minimize safety risks, the stronger the argument that such explanations indeed must be provided to prevent a design defect.

2. Matters are considerably more complicated if including explanations lowers the accuracy of the model (e.g., switching to a less powerful model type): in this case, it must first be assessed whether explanations enhance safety overall, by weighing potential harm from lower accuracy against potential prevention of harm from an increase in transparency. If risk is increased, the alternative can be discarded. If, however, it can be reasonably expected that the explanations entail a risk reduction, this reduction must be weighed against any additional costs the inclusion of explainability features might entail, as in the former case (risk-utility test). Again, trade secrets and incentives for innovation must be accounted for, generally implying local rather than global explanations (if any).

Importantly, in both cases, product liability law broadens the scope of explanations vis-à-vis data protection law. While the GDPR focuses on the data subject as the recipient of explanations, product liability more broadly considers any explanations that may provide a safety benefit, targeting therefore particularly the operators of the AI systems who determine if, how and when a system is put to use. Hence, under product liability law producers have to consider to what extent explanations may help operators safely use the AI product.

Product monitoring obligations Finally, under EU law, producers are not subject to product monitoring obligations once the product has been put onto the market. However, product liability law of some Member States does contain such monitoring obligations (e.g., Germany[3]). The producers, in this setting, have to keep an eye on the product to become aware of emerging safety risks, which is particularly important with respect to AI systems whose behavior might change after being put onto the market (e.g., via online learning). Arguably, explanations help fulfill this monitoring obligation. This, however, chiefly concerns explanations provided to the producer itself. If these are not shared with the wider public, trade secrets may be guarded; therefore, one might argue that even global explanations may be required. However, again, this would depend on the trade-off with the utility of the product as producers cannot be forced to put less utile products on the market unless the gain in safety, via local or global explanations, exceeds the potentially diminished utility.

[2] BGH, 16.6.2009, VI ZR 107/08, BGHZ 181, 253 para 18
[3] BGH, 17.3.1981, VI ZR 286/78 – Benomyl.

Results In sum, product liability law targets the producer as the responsible entity, but primarily focuses on explanations provided to the party controlling the safety risks of the AI system in the concrete application context, typically the operator. To the extent that national law contains product monitoring obligations, however, explanations to the producer may have to be provided as well. In all cases, the risk reduction facilitated by the explanations must be weighed against the potentially reduced utility of the AI system. In this, product liability law aligns itself with technical and protective transparency. It generates pressure to offer AI systems with actionable explanations by targeting the supply side of the market (producers).

2.2.2 General negligence standards Beyond product liability, general contract and tort law define duties of care that operators of devices, such as AI systems, need to fulfill in concrete deployment scenarios. Hence, it reaches the demand side of the market. While contract law covers cases in which the operator has a valid (pre-)contractual agreement with the harmed person (e.g., a physician with a patient; the bank with a credit applicant), tort law steps in if such an agreement is missing (e.g., autonomous lawnmower and injured pedestrian). However, the duties of care that relate to the necessary activities for preventing harm to the bodily integrity and the assets of other persons are largely equivalent under contract and tort law (see, e.g., [5, para 115]. In our context, this raises the question: do such duties of care require AI to be explainable, even if any specific contractual obligations to this end are lacking?

From Error Reversal to Risk-Adequate Choice Clearly, if the operator notices that the AI system is bound to make or has made an error, she has to overrule the AI decision to avoid liability [84, 42, 33]. Explanations geared toward the operator will often help her notice such errors and make pertaining corrections [80, p. 23][31]. For example, explanations could suggest that the system, in the concrete application, weighted features in an unreasonable manner and might fail to make a valid prediction [79, 71]. What is unclear, however, is whether the duty of care more generally demands explanations as a necessary precondition for using AI systems.

While much will depend on the concrete case, at least generally, the duty of care under both contract and tort law comprises monitoring obligations for operators of potentially harmful devices. The idea is that those who operate and hence (at least partially) control the devices in a concrete case must make reasonable efforts to control the risks the devices pose to third parties (cf. [101, para. 459]). The scope of that obligation is similar to the one in product liability, but not directed toward the producer, but rather the operator of the system: they must do whatever is factually possible and normatively reasonable and appropriate to prevent harm by monitoring the system. Hence, to the extent possible the operator arguably has to choose, at the moment of procurement, an AI system that facilitates risk control. Again, this reinforces technical and protective transparency in the name of safety gains. If an AI system providing actionable

explanations is available, such devices must therefore be chosen by the operator over non-explainable systems under the same conditions as in product liability law (i.e., if the explanation leads to an overall risk reduction justifying additional costs). For example, the operator need not choose an explainable system if the price difference to a non-explainable system constitutes an unreasonable burden. Note, however, that the operator, if distinct from the producer, cannot claim that trade secrets speak against an explainable version.

Alternative Design Obligations? Nonetheless, we would argue that the operator is not under an obligation to redesign the AI system, i.e., to actively install or use explanation techniques not provided by the producer, unless this is economically and technically feasible with efforts proportionate to the expected risk reduction. Rather, the safety obligations of the operator will typically influence the initial procurement of the AI system on the market. For example, if there are several AI-based credit scoring systems available the operator would have to choose the system with the best risk utility trade-off, taking into account explainability on both sides of the equation (potential reduction in utility and potential reduction of risk). Therefore, general contract and tort law sets incentives to use explainable AI systems similar to product liability, but with a focus on actions by, and explanations for, the operator of the AI system.

Results The contractual and tort-law duty of care therefore does not, other than in product liability, primarily focus on a potential alternative design of the system, but on prudently choosing between different existing AI systems on the market. Interpreted in this way, general contract and tort law generate market pressure toward the offer of explainable systems by targeting the demand side of the market (operators). Like product liability, however, they cater to technical and protective transparency.

2.3 Banking Law: More Technical and Protective Transparency

Finally, banking law provides for detailed regulation governing the development and application of risk scoring models. It therefore represents an underresearched, but in fact highly relevant area of algorithmic regulation, particularly in the case of credit scoring (see, e.g., [54]). Conceptually, it is intriguing because the quality requirements inherent in banking law fuse technical transparency with yet another legal and economic aim: the control of systemic risk in the banking sector.

2.3.1 Quality Assurance for Credit Models Significant regulatory experience exists in this realm because econometric and statistical models have long since been used to predict risk in the banking sector, such as creditworthiness of credit applicants [25]. In the wake of the financial crisis following the collapse of the subprime lending market, the EU legislator has enacted encompassing regulation addressing systemic risks stemming from the banking sector. Since

inadequate risk models have been argued to have contributed significantly to the scope and the spread of the financial crisis [4, p. 243-245], this area has been at the forefront of the development of internal compliance and quality regimes – which are now considered for AI regulation as well.

In general terms, credit institutions regulated under banking law are required to establish robust risk monitoring and management systems (Art. 74 of Directive 2013/36). More specifically, a number of articles in the Capital Requirements Regulation 575/2013 (CRR) set out constraints for the quality assurance of banking scoring models. Perhaps most importantly, Article 185 CRR compels banks to validate the score quality ('accuracy and consistency') of models for internal rating and risk assessment, via a continuous monitoring of the functioning of these models. Art. 174 CRR, in addition, specifies that: statistical models and 'other mechanical methods' for risk assessments must have good predictive power (lit. a); input data must be vetted for accuracy, completeness, appropriateness and representativeness (lit. b, c); models must be regularly validated (lit. d) and combined with human oversight (lit. e) (see [58, para. 1]; cf. [26, para. 249]; [21, paras. 68, 256]; for similar requirement for medical products, see [84]).

These provisions foreshadow many of the requirements the AIA proposed by the EU Commission now seeks to install more broadly for the regulation of AI. However, to the extent that AI-based credit scoring is used by banks, these provisions – other than the AIA – already apply to the respective models. While the responsible lending obligation contained in Article 8 of the Consumer Credit Directive 2008/48 only spells out generic duties to conduct creditworthiness assessments before lending decisions, Articles 174 and 185 CRR have complemented this obligation with a specific quality assurance regime. Ultimately, more accurate risk prediction is supposed to not only spare lenders and borrowers the transaction costs of default events, but also and perhaps even more importantly to rein in systemic risk in the banking sector by mitigating exposure. This, in turn, aims at reducing the probability of severe financial crises.

2.3.2 Consequences for XAI What does this entail for explainable AI in the banking sector? While accuracy (and model performance more generally) may be verified on the test data set in supervised learning settings without explanations relating to the relevant features for a prediction, explainability will, as mentioned, often be a crucial element for validating the generalizability of models beyond the test set (Art. 174(d) CRR), and for enabling human review (Art. 174(e) CRR). In its interpretive guidelines for supervision and model approval, the European Banking Authority (EBA) therefore stipulates that banks must 'understand the underlying models used', particularly in the case of technology-enabled credit assessment tools [26, para. 53c]. More specifically, it cautions that consideration should be given to developing interpretable models, if necessary for appropriate use of the model [26, para. 53d].

Hence, the explainability of AI systems becomes a real compliance tool in the realm of banking law, an idea we shall return to in the discussion of the AIA. In banking law, explainability is intimately connected to the control of systemic

risk via informed decision making of the individual actors. One might even argue that both local and global explainability are required under this perspective: local explainability helps determine accuracy in individual real-world cases for which no ground truth is available, and global explanations contribute to the verification of the consistency of the scoring tool across various domains and scenarios. As these explanations are generated internally and only shared with supervisory authorities, trade secrets do not stand in the way.

The key limitation of these provisions is that they apply only to banks in the sense of banking law (operating under a banking license), but not to other institutions not directly subject to banking regulation, such as mere credit rating agencies [7]. Nevertheless, the compliance and quality assurance provisions of banking law seem to have served as a blue print for current AI regulation proposals such as the EU Artificial Intelligence Act (esp. Art. 9, 14, 15 and 17), to which we now turn.

3 Regulatory Proposals at the EU Level: the AIA

The AIA, proposed by the EU Commission in April 2021, is set to become a cornerstone of AI regulation not only in the EU, but potentially with repercussions on a global level. Most notably, it subscribes to a risk-based approach and therefore categorically differentiates between several risk categories for AI. Figure 2 offers a snapshot of the connections between the functions of transparency and various Articles of the AIA.

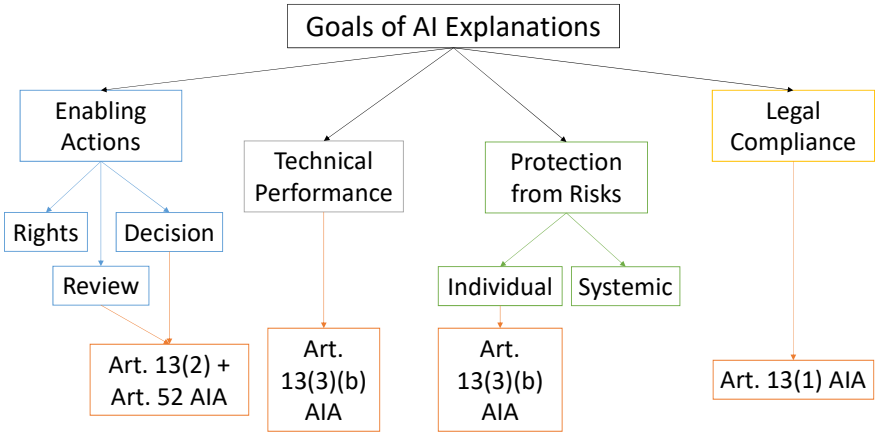

Figure 2: Overview of the functions of different Articles of the AIA transparency provisions

3.1 AI with Limited Risk: Decision-Enabling Transparency (Art. 52 AIA)?

For specific AI applications with limited risk, Article 52 AIA spells out transparency provisions in an enabling but highly constrained spirit (see also [95, 38]). Thus, the providers of AI systems interacting with humans, of emotion recognition systems, biometric categorization systems and of certain AI systems meant to manipulate images, audio recordings or videos (e.g., deep fakes) need to disclose the fact that an AI system is operating and, in the last case, that content was manipulated. Transparency, in this sense, does not relate to the inner workings of the respective AI systems, but merely to their factual use and effects.

The aim of these rules arguably is also of an enabling nature, but primarily with respect to informed choice, or rather informed avoidance (decision-enabling transparency), not the exercise of rights. Whether these rules will have any meaningful informational and behavioral effect on affected persons, however, must at least be doubted. A host of studies document rational as well as boundedly rational ignorance of standard disclosures in digital environments [13, 72, 1]. But regardless of the individual benefit, the more or less complete information about the use of low-risk AI systems alone is indirectly helpful in providing overviews and insights to civil society initiatives or journalistic projects, for example. Moreover, in the specific case of highly controversial AI applications such as emotion recognition or remote biometric identification, compulsory disclosure might, via coverage by media and watchdogs, engender negative reputational effects for the providers, which may lead some of them to reconsider the use of such systems in the first place.

3.2 AI with High Risk: Encompassing Transparency (Art. 13 AIA)?

The regulatory environment envisioned by the AIA is strikingly different for high-risk AI applications. Such applications are supposed to be defined via a regularly updated Annex to the AIA and, according to the current proposal, comprise a wide variety of deployment scenarios, from remote biometric identification to employment and credit scoring contexts, and from the management of critical infrastructure to migration and law enforcement (see Annex III AIA). In this regard, the question of the process of updating the AIA Annex is still open in terms of participation and public consultation. The requirements for low-risk AI systems to at least document the use and effects of the selected technologies, however, leads us to expect case-related disputes about whether an AI application should be classified as high risk, in which stakeholder representatives, civil and human rights protection initiatives, and manufacturers and users of technologies will wrestle with each other. This public struggle can also be seen as a rights-enabling transparency measure.

3.2.1 Compliance-Oriented Transparency For such high-risk applications, Article 13 AIA spells out a novel transparency regime that might be inter-

preted as seeking to fuse, to varying degrees, the several instrumentalist approaches identified in this paper, while notably foregrounding another goal of transparency: legal compliance.

Hence, Article 13(1) AIA mandates that high-risk AI systems be 'sufficiently transparent to enable users to interpret the system's output and use it appropriately'. In this, an 'appropriate type and degree of transparency' must be ensured. The provision therefore acknowledges the fundamentally different varieties of explanations that could be provided for AI systems, such as local, global or counterfactual explanations; or more or less granular information on feature weights. The exact scope and depth of the required transparency is further elaborated upon in Article 13(3) AIA and will need to be determined in a context-specific manner. Nothing in the wording of Article 13, however, suggests that global explanations, which may be problematic for complex AI systems, must be provided on a standard basis. However, explanations must be faithful to the model in the sense that they need to be an, at least approximately, correct reconstruction of the internal decision making parameters: explanation and explanandum need to match [57]. For example, local ex post explanations would have to verifiably and, within constraints, accurately measure feature relevance (or other aspects) *of the used model*.

Notably, with respect to the general goal of transparency, the additional explanatory language in Article 13(1) AIA introduces a specific and arguably novel variety of transparency instrumentalism geared toward effective and compliant application of AI systems in concrete settings. In fact, Article 13(1) AIA defines a particular and narrow objective for appropriate transparency under the AIA: facilitating the fulfillment of the obligations providers and users have under the very AIA (Chapter 3 = Art. 16-29). Most notably, any reference to rights of users or affected persons is lacking; rather, Article 29 AIA specifies that users may only deploy the AI system within the range of intended purposes specified by the provider and disclosed under Article 13(2) AIA. Hence, transparency under the AIA seems primarily directed toward compliance with the AIA itself, and not towards the exercise of rights affected persons might have. In this sense, the AIA establishes a novel, self-referential, compliance-oriented type of transparency instrumentalism.

3.2.2 Restricted Forms of Enabling and Protective Transparency For specific applications, the recitals, however, go beyond this restrained compliance conception and hold that, for example in the context of law enforcement, transparency must facilitate the exercise of fundamental rights, such as the right to an effective remedy or a fair trial (Recital 38 AIA). This points to a more encompassing rights-enabling approach, receptive of demands for contestability, which stands in notable tension, however, with the narrower, compliance-oriented wording of Article 13(1) AIA. To a certain extent, however, the information provided under Article 13 AIA will facilitate audits by supervisory authorities, collective redress organizations or NGOs ('review-enabling transparency').

Furthermore, the list of specific items that need to be disclosed under Article 13(3) AIA connects to technical and protective instrumentalist conceptions of transparency (see also [41]). Hence, Article 15 AIA mandates appropriate levels of accuracy, as well as robustness and cybersecurity, for high-risk AI systems. According to Article 13(3)(b)(ii) AIA, the respective metrics and values need to be disclosed. In this, the AIA follows the reviewed provisions of banking law in installing a quality assurance regime for AI models whose main results need to be disclosed. As mentioned, this also facilitates legal review: if the disclosed performance metrics suggest a violation of the requirements of Article 15 AIA, the supervisory authority may exercise its investigative and corrective powers.The institutional layout of this oversight and supervisory regime however is still not fully defined: The sectoral differentiation of AI applications in the AIA's risk definitions on the one hand suggest an equally sectoral organization of supervisory authorities; the technical and procedural expertise needed for such oversight procedures on the other hand calls for a less distributed supervisory regime.

Similarly, Article 10 AIA installs a governance regime for AI training data, whose main parameters, to the extent relevant for the intended purpose, also need to be divulged (Art. 13(3)(b)(v) AIA). Any other functionally relevant limitations and predetermined changes must be additionally informed about (Art. 13(3)(b)(iii),(iv), (c) and (e)). Finally, disclosure also extends to human oversight mechanisms required under Article 14 AIA – like the governance of training data another transplant from the reviewed provisions on models in banking law. Such disclosures, arguably, cater to protective transparency as they seek to guard against use of the AI system beyond its intended purpose, its validated performance or in disrespect of other risk-minimizing measures.

Hence, transparency under Article 13 is intimately linked to the requirements of human oversight specified in Article 14 AIA. That provision establishes another important level of protective transparency: high-risk AI applications need to be equipped with interface tools enabling effective oversight by human persons to minimize risks to health, safety and fundamental rights. Again, as discussed in the contract/tort and banking law sections, local explanations particularly facilitate monitoring and the detection of inappropriate use or anomalies engendering such risks (cf. Art. 14(4)(a) AIA). While it remains a challenge to implement effective human oversight in AI systems making live decisions (e.g., in autonomous vehicles), the requirement reinforces the focus of the AIA on transparency vis-à-vis professional operators, not affected persons.

3.3 Limitations

The transparency provisions in the AIA in several ways represent steps in the right direction. For example, they apply, other than the GDPR rules reviewed, irrespective of whether decision making is automated or not and of whether personal data is processed or not. Furthermore, the inclusion of a quality assurance regime should be welcomed and even be (at least partially) expanded to non-high-risk applications, as disclosure of pertinent performance metrics may be of substantial signaling value for experts and the market. Importantly, the

rules of the future AIA (and of the proposed Machinery Regulation) will likely at least generally constitute minimum thresholds for the avoidance of design defects in product liability law (see Section 2.2.1), enabling decentralized private enforcement next to the public enforcement foreseen in the AIA. Nonetheless, the transparency provisions of the AIA are subject to significant limitations.

First and foremost, self-referential compliance and protective transparency seems to detract from meaningful rights-enabling transparency for affected persons. Notably, the transparency provisions of Article 13 AIA are geared exclusively toward the users of the system, with the latter being defined in Article 3(4) AIA as anyone using the system with the exception of consumers. While this restriction has the beneficial effect of sparing consumers obligations and liability under the AIA (cf. [102]), for example under Article 29 AIA, it has the perhaps unintended and certainly significant effect of excluding non-professional users from the range of addressees of explanations and disclosure[27, 91]. Therefore, the enabling variety of transparency, invoked in lofty words in Recital 38 AIA, is missing from the Articles of the AIA and will in practice be largely relegated to other, already existing legal acts – such as the transparency provisions of the GDPR reviewed above. In this sense, the AIA does not make any significant contribution to extending or sharpening the content of the requirement to provide 'meaningful information' to data subjects under the GDPR. In this context, information facilitating a review in terms of potential bias with respect to protected groups is missing, too.

Second, this focus on professional users and presumed experts continues in the long list of items to be disclosed under Article 13(3) AIA. While performance metrics, specifications about training data and other disclosures do provide relevant information to sophisticated users to determine whether the AI system might present a good fit to the desired application, such information will only rarely be understandable and actionable for users without at least a minimal training in ML development or practice. In this sense, transparency under the AIA might be described as transparency 'by experts for experts', likely leading to information overload for non-experts. The only exception in this sense is the very reduced, potentially decision-enabling transparency obligation under Article 52 AIA.

Third, despite the centrality of transparency for trustworthy AI in the communications of the EU Commission (see, e.g., European Commission, 2020), the AIA contains little incentive to actually disclose information about the inner workings of an AI system to the extent that they are relevant and actionable for affected persons. Most of the disclosure obligations refer either to the mere fact that an AI system of a specific type is used (Art. 52 AIA) or to descriptions of technical features and metrics (Art. 13(3) AIA). Returning briefly to the example of credit scoring, the only provision potentially impacting the question of whether local or even global explanations of the scores (feature weights) are compulsory is the first sentence of Article 13(1) AIA. According to it, users (i.e., professionals at the bank or credit scoring agency) must be able to interpret the system's output. The immediate reference, in the following sentence, to the

obligations of users under Article 29 AIA, however, detracts from a reading that would engage Article 13 AIA to provide incentives for clear and actionable explanations beyond what is already contained in Articles 13-15 GDPR. The only interpretation potentially suggesting local, or even global, explanations is the connection to Article 29(4) AIA. Under this provision, users have to monitor the system to decide whether use according to the instructions may nonetheless lead to significant risks. One could argue that local explanations could be conducive to and perhaps even necessary for this undertaking to the extent that they enable professional users to determine if the main features used for the prediction were at least plausibly related to the target, or likely rather an artifact of the restrictions of training, e.g., of overfitting on training data (cf. [79]). Note, however, that for credit institutions regulated under banking law, the specific provisions of banking law take precedence over Article 29(4) and (5) AIA.

Fourth, while AI systems used by banks will undergo a conformity assessment as part of the supervisory review and evaluation process already in place for banking models (Art. 43(2)(2) AIA), the providers of the vast majority of high-risk AI systems will be able to self-certify the fulfilment of the criteria listed in the AIA, including the transparency provisions in Art. 13 (see Art. 43(2)(1) AIA). The preponderance of such self-assessment may result from an endeavor to exonerate regulatory agencies and to limit the regulatory burden for providers, but it clearly reduces enforcement pressure and invites sub-optimal compliance with the already vague and limited transparency provisions (cf. also [95, 91].

In sum, the AIA provides for a plethora of information relevant for sophisticated users, in line with technical transparency, but will disappoint those that had hoped for more guidance on and incentives for meaningful explanations enabling affected persons to review and contest the output of AI systems.

4 Beyond Explainability

As the legal overview has shown, different areas of law embody different conceptions of AI explainability. Perhaps most importantly, however, if explanations are viewed as a social act enabling a dialogical exchange and laying the basis for goal-oriented actions of the respective recipients, it will often not be sufficient to just provide them with laundry lists of features, weights or model architectures. There is a certain risk that the current drive toward explainable AI, particularly if increasingly legally mandated, generates information that does not justify the transaction costs it engenders. Hence, computer science and the law have to go beyond mere explainability toward interactions that enable meaningful agency of the respective recipients [103], individually, but even more so by strengthening the ability of stakeholder organizations or civil and human rights organizations. This includes a push for actionable explanations, but also for connections to algorithmic fairness, to quality benchmarking and to co-design strategies in an attempt to construct responsible, trustworthy AI [3, 45].

4.1 Actionable Explanations

The first desideratum, therefore, is for explanations to convey actionable information, as was stressed throughout the article. Otherwise, for compliance reasons and particularly under the provisions of the AIA, explanations might be provided that few actors actually cognitively process and act upon. This implies a shift from a focus on the technical feasibility of explanations toward, with at least equal importance, the recipient-oriented design of the respective explanations.

4.1.1 Cognitive optimization Generally, to be actionable, explanations must be designed such that information overload is avoided, keeping recipients with different processing capabilities in mind. This is a lesson that can be learned from decades of experience with the disclosure paradigm in US and EU consumer law: most information is flatly ignored by consumers [8, 72]. To stand a chance of being cognitively processed, the design of explanations must thus be recipient-oriented. In this, a rich literature on enhancing the effectiveness of privacy policies and standard information in consumer and capital markets law can be exploited [10, 64]. Information, in this sense, must be cognitively optimized for the respective recipients, and the law, or at least the implementing guidelines, should include rules to this effect.

To work, explanations likely must be salient and simple [93] and include visualizations [48]. Empirical studies indeed show that addressees prefer simple explanations [78]. Furthermore, when more complex decisions need to be explained, information could be staggered by degree of complexity. Research on privacy policies, for example, suggests that multi-layered information may bridge the gap between diverging processing capacities of different actors [83].Hence, simple and concise explanations could be given first, with more detailed, expert-oriented explanations provided on a secondary level upon demand. For investment information, this has already been implemented with the mandate on a Key Investor Document in EU Regulation 1286/2014 (PRIIPS Regulation) (see also [54, p.540]). Finally, empirical research again shows that actionable explanations tend to be contrastive, a concept increasingly explored in AI explanations as well [64, 65].

Hence, there are no one-size-fits-all explanations; rather, they need to be adapted to different contexts and addressees. What the now classic literature on privacy policies suggests is that providing information is only one element of a more general privacy awareness and privacy-by-design strategy [44] that takes different addressees, practical needs and usable tools into account: A browser-plugin notifying about ill-defined or non-standard privacy settings can be more helpful for individual consumers than a detailed and descriptive walk-through of specific privacy settings. A machine-readable and standardized format for reviewing and monitoring privacy settings, however, is helpful for more technical reviews by privacy advocacy organizations. The 'ability to respond' to different contexts and addressees therefore is a promising path towards 'response-able' [51] AI. One particular strategy might be to let affected persons choose foils

(within reasonable constraints) and generate contrastive explanations bridging the gap between fact and foil.

4.1.2 Goal orientation Beyond these general observations for cognitive optimization, actionable explanations should be clearly linked to the respective goals of the explanations. If the objective is to enable an understanding of the decision by affected persons and to permit the exercise of rights or meaningful review (rights- or review-enabling transparency), shortlists of the most relevant features for the decision ought to be required [79][12, for limitations].This facilitates, inter alia, checks for plausibility and discrimination. Importantly, such requirements have, in some areas, already been introduced into EU law by recent updates of consumer and business law. Under the new Art. 6a of the Consumer Rights Directive and the new Art. 7(4a) of the Unfair Commercial Practices Directive, online marketplaces will shortly need to disclose the main parameters for any ranking following a search query, and their relative importance. Art. 5 of the P2B Regulation 2019/1150 equally compels online intermediaries and search engines to disclose the main parameters of ranking and their relative importance. However, these provisions require global, not local explanations [37, p.52][14, p.161].

This not only generates technical difficulties for more complex AI systems, but the risk that consumers will flatly ignore such global explanations is arguably quite high. Rather, in our view, actionable information should focus on local explanations for individual decisions. Such information not only seems to be technically easier to provide, but it is arguably more relevant, particularly for the exercise of individual rights. From a review-enabling perspective, local information could be relevant as well for NGOs, collective redress organizations and supervisory authorities seeking to prosecute individual rights violations. In this sense, a collective dimension of individual transparency emerges (cf. also [46]). On the downside, however, local feature relevance information may produce a misleading illusion of simplicity; in non-linear models, even small input changes may alter principal reason lists entirely [57, 12].

If, therefore, the goal is not to review or challenge the decision, but to facilitate market decisions and particularly to create spaces for behavioral change of affected persons (decision-enabling transparency), for example to improve their credit score, counterfactual or contrastive information might serve the purpose better [97, 65]. In the example of credit scoring, this could set applicants toward the path of credit approval. Such information could be problematic, however, if the identified features merely correlate with creditworthiness, but are not causal for it. In this case, the risk of applicants trying to 'game the system' by artificially altering non-causal features are significant (e.g., putting felt tips under furniture as predictors of creditworthiness [85, p.71]). Moreover, in highly dimensional systems with many features, many counterfactuals are possible, making it difficult to choose the most relevant one for the affected person [97, p.851]. In addition, some counterfactually relevant features may be hard or impossible

to change (e.g., age, residence) [50]. In these cases, local shortlists of the most relevant features [79] or minimal intervention advice [50] might be more helpful.

Overall, research for the type of explanation with the best fit for each context will have to continue; it will benefit from cross-fertilization with social science research on the effectiveness of information more generally and explanations more particularly [64] as well as with research in science & technology studies on organizational, institutional and cultural contextualization of decision support, explanations, and accountability. Ultimately, a context-dependent, goal-oriented mix of explanations (e.g., relevance shortlist combined with counterfactual explanation) might best serve the various purposes explanations have to fulfil in concrete settings. In this, a critical perspective drawing on the limitations of the disclosure paradigm in EU market law (see, e.g., [11, 39]) should be helpful to prevent information overload and to limit disclosure obligations to what is meaningfully oriented to the respective goals of the explanations.

4.2 Connections to Algorithmic Fairness

Transparency, and explanations such as disclosure of the most relevant features of an AI output, may serve yet another goal: non-discrimination in algorithmic decision making. A vast literature deals with tools and metrics to implement non-discrimination principles at the level of AI models to facilitate legal compliance [76, 106, 52]. Explanations may reinforce such strategies by facilitating bias detection and prevention, both by affected persons and review institutions. For example, in the case of credit scoring, disclosure of the most important features (local explanations) could help affected persons determine to what extent the decision might have been driven by variables closely correlated with protected attributes [3]. Such cross-fertilization between bias detection and explanations could be termed 'fairness-enabling transparency' and should constitute a major research goal from a legal and technical perspective.

In a similar vein, Sandra Wachter and colleagues have convincingly advocated for the disclosure of summary statistics showing the distribution of scores between different protected groups [98]. As one of the authors of this contribution has argued, such disclosures might in fact already be owed under the current GDPR disclosure regime (Art. 13(2)(f), Art. 14(2)(g), Art. 15(1)(h) GDPR: information about the 'significance and envisaged consequences' of processing, see [40, p.1173-1174]). In addition, Art. 13(3)(b)(iv) AIA proposes the disclosure of a high-risk AI system's 'performance as regards the persons or groups of persons on which the system is intended to be used". While one could interpret this as a mandate for differential statistics concerning protected groups, such an understanding is unlikely to prevail, in the current version of the AIA, as a reference to protected attributes in the sense of antidiscrimination law is patently lacking. Fairness-enabling transparency, such as summary statistics showing distributions between protected groups, to the extent available, thus constitutes an area that should be included in the final version of the AIA.

4.3 Quality Benchmarking

Finally, technical and protective transparency closely relates to (the disclosure of) quality standards for AI systems. These metrics, in turn, also enable regulatory review and are particularly important, as seen, in banking law [54, p.561-563]. Two aspects seem to stand out at the intersection of explanations and quality benchmarking:

First, an absolute quality control, such as the one installed in Art. 174/185 CRR, could be enshrined for all AI applications, at least in medium- and high-stakes settings (transcending the ultimately binary logic of the AIA with respect to risk classification). In these settings, quality assurance might be considered as important as, or even more important than, mere explainability. Quality control would include, but not be limited to, explanations facilitating decisions about the generalizability of the model (e.g., local explanations). Importantly, the disclosure of performance metrics would also spur workable competition by enabling meaningful comparison between different AI systems. Notably, relevant quality assurance provisions in the AIA (Art. 10/15 AIA) are limited to high-risk applications. An update of the AIA might draw inspiration from banking law in working toward a quality assurance regime for algorithmic decision making in which the monitoring of field performance and the assessment of the generalizability of the model via explainability form an important regulatory constraint not only for high-risk but also for medium-risk applications, at least.

Second, understanding the risks and benefits of, and generating trust in, AI systems should be facilitated by testing the quality of AI models against the benchmark of traditional (non-AI-based) methods (relative quality control). For example, a US regulator, the Consumer Financial Protection Bureau, ordered a credit scoring startup working with alternative data to provide such an analysis. The results were promising: according to the analysis, AI-based credit scoring was able to deliver cheaper credit and improved access, both generally and with respect to many different consumer subgroups [30][35, p.42]. To the extent that the analysis is correct, it shows that AI, if implemented properly and monitored rigorously, may provide palpable benefits not only to companies using it, but to consumers and affected persons as well. Communicating such benefits by benchmarking reports seems a sensible way to enable more informed market decisions, to facilitate review and to generate trust – strengthening three important pillars of any explainability regime for AI systems.

4.4 Interventions and Co-Design

Such ways of going beyond the already existing and currently proposed forms of transparency obligations by developing formats and methods to produce actionable explanations, by connecting transparency and explainability issues to questions of algorithmic fairness and new or advanced forms of quality benchmarking and control are, as favorable as they are, mainly ex post mechanisms aiming at helping affected persons, users, NGOs or supervisory authorities to evaluate and act upon the outcomes of AI systems in use. They can inform market decisions,

help affected persons to claim rights or enable regular oversight and supervision, but they do not intervene in the design and implementation of complex AI systems. Linking to two distinct developments of inter- and transdisciplinary research can help to further develop forms of intervention and co-design:

First, methods and formats for 'values-in-design' [70, 53] projects have been developed in other areas of software engineering, specifically in human computer interaction (HCI) and computer-supported collaborative work (cscw) setups that traditionally deal with heterogenous user groups as well as with a diverse set of organizational and contextual requirements due to the less domain-specific areas of application of these software systems (see [32] for an overview). Formats and methods include the use of software engineering artifacts to make normative requirements visible and traceable or the involvement of affected persons, stakeholders, or spokespersons in requirements engineering, evaluation and testing [32, 75]. Technical transparency as discussed above can support the transfer and application of such formats and methods to the co-design of AI systems [2] with global explanations structuring the process and local explanations supporting concrete co-design practices.

Second, these methodological advances have been significantly generalized and advanced under the 2014-2020 Horizon 2020 funding scheme, moving from 'co-design to ELSI co-design' [56] and leading to further developing tools, methods and approaches designed for research on SwafS ('Science with and for Society') into a larger framework for RRI ('Responsible Research and Innovation') [28]. In AI research, specifically in projects aiming to improve accountability or transparency, a similar, but still quite disconnected movement towards 'Responsible AI' [24] has gained momentum, tackling very similar questions of stakeholder integration, formats for expert/non-expert collaboration, domain-knowledge evaluation or contestation and reversibility that have been discussed within the RRI framework with a focus on energy technologies, biotechnologies or genetic engineering. This is a rich resource to harvest for further steps towards XAI by adding addressee orientation, contestability criteria or even, reflexively, tools to co-design explanations through inter- and transdisciplinary resarch [63, 69].

5 Conclusion

This paper has sought to show that the law, to varying degrees, mandates or incentivizes different varieties of AI explanations. These varieties can be distinguished based on their respective functions or goals. When affected persons are the addressees, explanations should be primarily rights-enabling or decision-enabling. Explanations for operators or producers, in turn, will typically facilitate technical improvements and functional review, fostering the mitigation of legally relevant risks. Finally, explanations may enable legal review if perceived by third parties, such as NGOs, collective address organizations or supervisory authorities.

The GDPR, arguably, subscribes to a rights-enabling transparency regime under which local explanations may, depending on the context, have to be provided to individual affected persons. Contract and tort law, by contrast, strive for technical and protective transparency under which the potential trade-off between performance and explainability takes center stage: any potentially reduced accuracy or utility stemming from enforcing explanations must be weighed against the potential safety gains such explanations enable. Explanations are required only to the extent that this balance is positive. Banking law, finally, endorses a quality assurance regime in which transparency contributes to the control of systemic risk in the banking sector. Here, even global explanations may be required. The proposal for the AIA, in turn, is primarily geared toward compliance-oriented transparency for professional operators of AI systems. From a rights-enabling perspective, this is a significant limitation.

These legal requirements, however, can be interpreted to increasingly call for actionable explanations. This implies moving beyond mere laundry lists of relevant features toward cognitively optimized and goal-oriented explanations. Multi-layered or contrastive explanations are important elements in such a strategy. Tools, methods and formats from various values-in-design approaches as well as those developed under under the umbrella term of 'responsible research and innovation' can help co-designing such systems and explanations.

Finally, an update of the AIA should consider fairness-enabling transparency, which seeks to facilitate the detection of potential bias in AI systems, as well as broader provisions for quality benchmarking to facilitate informed decisions by affected persons, to enable critical review and the exercise of rights, and to generate trust in AI systems more generally.

References

1. Acquisti, A., Taylor, C., Wagman, L.: The economics of privacy. Journal of Economic Literature **54**(2), 442–92 (2016)
2. Aldewereld, H., Mioch, T.: Values in Design Methodologies for AI. In: Polyvyanyy, A., Rinderle-Ma, S. (eds.) Advanced Information Systems Engineering Workshops. pp. 139–150. Lecture Notes in Business Information Processing, Springer International Publishing, Cham (2021)
3. Arrieta, A.B., Díaz-Rodríguez, N., Del Ser, J., Bennetot, A., Tabik, S., Barbado, A., García, S., Gil-López, S., Molina, D., Benjamins, R., Chatila, R., Herrera, F.: Explainable Artificial Intelligence (XAI): Concepts, taxonomies, opportunities and challenges toward responsible AI. Information Fusion **58**, 82–115 (2020)
4. Avgouleas, E.: Governance of global financial markets: the law, the economics, the politics. Cambridge University Press (2012)
5. Bachmann, G.: Commentary on § 241 BGB, in: Münchener Kommentar zum BGB. BECK, Munich, 8th ed. edn. (2019)
6. Bäcker, M.: Commentary on Art. 13 GDPR, in: Kühling/Buchner, DS- GVO Commentary. BECK, Munich, 3rd ed. edn. (2020)
7. BaFin: Rolle der Aufsicht bei der Verwendung von Kreditscores. BaFin Journal pp. 22–24 (Mar 2019)
8. Bakos, Y., Marotta-Wurgler, F., Trossen, D.D.: Does anyone read the fine print? Consumer attention to standard- form contracts. The Journal of Legal Studies **43**, 1–35 (2014)
9. Bambauer, J., Zarsky, T.: The algorithm game. Notre Dame L. Rev. 94, 1 (2018)
10. Bar-Gill, O.: Smart disclosure: Promise and perils. Behavioural Public Policy (2021)
11. Bar-Gill, O., Ben-Shahar, O.: Regulatory techniques in consumer protection: a critique of European consumer contract law. Common Market Law Review **50**, 109–126 (2013)
12. Barocas, S., Selbst, A.D., Raghavan, M.: The hidden assumptions behind counterfactual explanations and principal reasons. Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency pp. 80–89 (Jan 2020)
13. Ben-Shahar, O., Chilton, A.S.: Simplification of privacy disclosures: An experimental test. The Journal of Legal Studies **45**(S2), S41–S67 (2016)
14. Bibal, A., Lognoul, M., de Streel, A., Frénay, B.: Legal requirements on explainability in machine learning. Artificial Intelligence and Law **29**, 149–169 (2021)
15. Biran, O., Cotton, C.V.: Explanation and justification in machine learning: A survey. IJCAI-17 workshop on explainable AI (XAI) **8**(1), 8–13 (Aug 2017)
16. Breiman, L.: Random forests. Machine learning **45**(1), 5–32 (2001)
17. Brownsword, R.: From Erewhon to AlphaGo: for the sake of human dignity, should we destroy the machines? Law, Innovation and Technology **9**(1), 117–153 (2017)
18. Burrell, J.: How the machine 'thinks': Understanding opacity in machine learning algorithms. Big Data & Society **3**(1) (2016)
19. Cabral, T.S.: Liability and artificial intelligence in the EU: Assessing the adequacy of the current Product Liability Directive. Maastricht Journal of European and Comparative Law **27**(5), 615–635 (2020)
20. Casey, B., Farbangi, A., Vogl, R.: Rethinking Explainable Machines: The GDPR's 'Right to Explanation' Debate and the Rise of Algorithmic Audits in Enterprise. Berkeley Technology Law Journal **34**, 143 (2019)

21. CEBS (Committee of the European Banking Supervisors): Guidelines on the implementation, validation and assessment of Advanced Measurement (AMA) and Internal Ratings Based (IRB) Approaches (2006)
22. Chen, J.M.: Interpreting Linear Beta Coefficients Alongside Feature Importances. Machine Learning (2021)
23. Citron, D.K., Pasquale, F.: The scored society: Due process for automated predictions. Washington Law Review **89**(1) (2014)
24. Dignum, V.: Responsible Artificial Intelligence: How to Develop and Use AI in a Responsible Way. Artificial Intelligence: Foundations, Theory, and Algorithms, Springer International Publishing, Cham (2019)
25. Dumitrescu, E.I., Hué, S., Hurlin, C.: Machine Learning or Econometrics for Credit Scoring: Let's Get the Best of Both Worlds. Working Paper (2021)
26. EBA (European Banking Authority): Guidelines on loan origination and monitoring (2020)
27. Ebers, M., Hoch, V.R., Rosenkranz, F., Ruschemeier, H., Steinrötter, B.: The european commission's proposal for an artificial intelligence act—a critical assessment by members of the robotics and ai law society (rails). J **4**(4), 589–603 (2021)
28. European Commission: Responsible research and innovation Europe's ability to respond to societal challenges. (2012)
29. Expert Group on Liability and New Technologies: New Technologies Formation, Liability for Artificial Intelligence and other emerging digital technologies. Tech. rep. (2019)
30. Fickling, P.A., Watkins, P.: An update on credit access and the Bureau's first No - Action Letter (Aug 2019)
31. Fisher, A.J., Rudin, C., Dominici, F.: All Models are Wrong, but Many are Useful: Learning a Variable's Importance by Studying an Entire Class of Prediction Models Simultaneously. Journal of Machine Learning Research **20**(177), 1–81 (2019)
32. Friedman, B., Hendry, D.G., Borning, A.: A Survey of Value Sensitive Design Methods. Foundations and Trends in Human-Computer Interaction **11**(2), 63–125 (Nov 2017)
33. Froomkin, A.M., Kerr, I.R., Pineau, J.: When AIs outperform doctors: confronting the challenges of a tort-induced over-reliance on machine learning. Ariz. Law Rev. **61**, 33 (2019)
34. Fuster, A., Goldsmith-Pinkham, P., Ramadorai, T., Walther, A.: "Predictably unequal? the effects of machine learning on credit markets.". Working Paper (Oct 2020)
35. Gillis, T.B., is,Talia B.: The Input Fallacy. Minnesota Law Review (forthcoming) (2021)
36. Goodman, B., Flaxman, S.: EU regulations on algorithmic decision-making and a "right to explanation". WHI (Jun 2016)
37. Grochowski, M., Jabłonowska, A., Lagioia, F., Sartor, G., Sartor, G., Francesca, Sartor, G.: Algorithmic Transparency and Explainability for EU Consumer Protection: Unwrapping the Regulatory Premises. Critical Analysis of Law **8**(1), 43–63 (Apr 2021)
38. Hacker, P.: Manipulation by Algorithms. Exploring the Triangle of Unfair Commercial Practice, Data Protection, and Privacy Law. European Law Journal (forthcoming), doi: https://doi.org/10.1111/eulj.12389
39. Hacker, P.: The Behavioral Divide: A Critique of the Differential Implementation of Behavioral Law and Economics in the US and the EU. European Review of Contract Law **11**(4), 299–345 (2015)

40. Hacker, P.: Teaching fairness to artificial intelligence: existing and novel strategies against algorithmic discrimination under EU law. Common Market Law Review **55**(4), 1143–1186 (2018)
41. Hacker, P.: Europäische und nationale Regulierung von Künstlicher Intelligenz. NJW (Neue Juristische Wochenschrift) pp. 2142–2147 (2020)
42. Hacker, P., Krestel, R., Naumann, F., Grundmann, S.: Explainable AI under contract and tort law: legal incentives and technical challenges. Artificial Intelligence and Law **28**, 415–439 (2020)
43. Hagendorff, T.: "The ethics of AI ethics: An evaluation of guidelines.". Minds and Machines **30**, 99–120 (2020)
44. Hansen, M.: Data protection by design and by default à la european general data protection regulation. In: IFIP International Summer School on Privacy and Identity Management. pp. 27–38. Springer (2016)
45. High-Level Expert Group on Artificial Intelligence: Ethics guidelines for trustworthy AI (2019)
46. Hildebrandt, M.: Privacy as protection of the incomputable self: From agnostic to agonistic machine learning. Theoretical Inquiries in Law **20**(1), 83–121 (2019)
47. Holzinger, A., Biemann, C., Pattichis, C.S., Kell, D.B.: What do we need to build explainable AI systems for the medical domain? arXiv preprint arXiv:1712.09923 (2017)
48. Jolls, C.: Debiasing through law and the first Amendment. Stanford Law Review **67**, 1411 (2015)
49. Kaminski, M.E.: The Right to Explanation, Explained. Berkeley Technology Law Journal **34**, 189 (2019)
50. Karimi, A.H., Schölkopf, B., Valera, I.: Algorithmic recourse: from counterfactual explanations to interventions. In Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency (Mar 2021)
51. Kenney, M.: Fables of Response-ability: Feminist Science Studies as Didactic Literature. Catalyst: Feminism, Theory, Technoscience **5**(1), 1–39 (Apr 2019)
52. Kleinberg, J., Ludwig, J., Mullainathan, S., Rambachan, A.: Algorithmic fairness. AEA Papers and Proceedings **108**, 22–27 (May 2018)
53. Knobel, C., Bowker, G.C.: Values in design. Communications of the ACM **54**(7), 26–28 (2011)
54. Langenbucher, K.: Responsible AI-based Credit Scoring – A Legal Framework. European Business Law Review **31**(4), 527–572 (2020)
55. Lapuschkin, S., Wäldchen, S., Binder, A., Montavon, G., Samek, W., Müller, K.R.: Unmasking clever hans predictors and assessing what machines really learn. Nature communications **10**(1), 1–8 (2019)
56. Liegl, M., Oliphant, R., Buscher, M.: Ethically aware IT design for emergency response: from co-design to ELSI co-design. Proceedings of the ISCRAM 2015 Conference (2015)
57. Lipton, Z.C.: The Mythos of Model Interpretability: In machine learning, the concept of interpretability is both important and slippery. Queue **16**(3), 31–57 (2018)
58. Loch, F.: Art. 174, Boos/Fischer/Schulte- Mattler (eds.), VO (EU) 575/2013, 5th ed. (2016)
59. Lombrozo, T.: The structure and function of explanations. Trends in cognitive sciences **10**(10), 464–470 (2006)
60. Lundberg, S.M., Lee, S.I., Lundberg: A Unified Approach to Interpreting Model Predictions. Advances in Neural Information Processing Systems 30 pp. 4765–4774 (2017)

61. Malgieri, G., Comandé, G.: Why a right to legibility of automated decision-making exists in the general data protection regulation. International Data Privacy Law (2017)
62. Malle, B.F.: How the mind explains behavior: Folk explanations, meaning, and social interaction. MIT Press (2004)
63. Mendez Fernandez, D., Passoth, J.H.: Empirical software engineering. From discipline to interdiscipline. Journal of Systems and Software **148**, 170–179 (2019)
64. Miller, T.: Explanation in Artificial Intelligence: Insights from the Social Sciences. Artificial Intelligence **267**, 1–38 (2019)
65. Miller, T.: Contrastive explanation: A structural-model approach. arXiv preprint arXiv:1811.03163 (2020)
66. Mittelstadt, B.D., Allo, P., Taddeo, M., Wachter, S., Floridi, L.: "The ethics of algorithms: Mapping the debate.". Big Data and Society 3.2 (2016)
67. Moore, J.D., Swartout, W.: Explanation in Expert Systems: A Survey, Information Sciences Institute Tech Report. Tech. Rep. ISI/RR-88-228 (1988)
68. Müller, H., Mayrhofer, M.T., Van Veen, E.B., Holzinger, A.: The Ten Commandments of Ethical Medical AI. Computer **54**(07), 119–123 (2021)
69. Müller, P., Passoth, J.H.: Engineering Collaborative Social Science Toolkits. STS Methods and Concepts as Devices for Interdisciplinary Diplomacy. In: Karafillidis, A., Weidner, R. (eds.) Developing Support Technologies: Integrating Multiple Perspectives to Create Assistance that People Really Want, pp. 137–145. Biosystems & Biorobotics, Springer International Publishing, Cham (2018)
70. Nissenbaum, H.: Values in the Design of Computer Systems. Computers in Society (March), 38–39 (1998)
71. N.N.: Editorial, Towards Trustable Machine Learning. Nature Biomedical Engineering **2**, 709–710 (Oct 2018)
72. Obar, J.A., Oeldorf-Hirsch, A.: The biggest lie on the internet: Ignoring the privacy policies and terms of service policies of social networking services. Information, Communication & Society **23**(1), 128–147 (2020)
73. Paal, B., Hennemann, M.: Commentary on Art. 13, in Paal/Pauly (eds.), Datenschutz-Grundverordnung. Kommentar. BECK, Munich, 3rd ed. edn. (2021)
74. Pasquale, F.: The black box society. Harvard University Press (2015)
75. Passoth, J.H.: Die Demokratisierung des Digitalen. Konrad Adenauer Stiftung: Analysen & Argumente (424), 1–13 (2021)
76. Pessach, D., Shmueli, E.: Algorithmic fairness. arXiv preprint arXiv:2001.09784 (2020)
77. Rathi, S.: Generating counterfactual and contrastive explanations using SHAP. arXiv preprint arXiv:1906.09293 (2019)
78. Read, S.J., Marcus-Newhall, A.: Explanatory coherence in social explanations: A parallel distributed processing account. Journal of Personality and Social Psychology **65**(3), 429 (1993)
79. Ribeiro, M.T., Singh, S., Guestrin, C.: " Why should i trust you?" Explaining the predictions of any classifier. Proceedings of the 22nd ACM SIG KDD international conference on knowledge discovery and data mining pp. 1135–1144 (2016)
80. Ronan, H., Junklewitz, H., Sanchez, I.: Robustness and Explainability of Artificial Intelligence. JRC Technical Report 13 (2020)
81. Rudin, C.: Stop explaining black box machine learning models for high stakes decisions and use interpretable models instead. Nature Machine Intelligence **1**(5), 206–215 (2019)

82. Samek, W., Montavon, G., Vedaldi, A., Hansen, L.K., Müller, K.R. (eds.): Explainable AI: Interpreting, Explaining and Visualizing Deep Learning, vol. 11700. Springer Nature (2019)
83. Schaub, F., Balebako, R., Durity, A.L.: A design space for effective privacy notices. In: Eleventh Symposium On Usable Privacy and Security ({SOUPS} 2015). pp. 1–17 (2015)
84. Schneeberger, D., Stöger, K., Holzinger, A.: The European legal framework for medical AI. In: International Cross-Domain Conference for Machine Learning and Knowledge Extraction. pp. 209–226. Springer, Cham (2020)
85. Schröder, T.: Programming Fairness. MaxPlanckResearch pp. 68–73 (2019)
86. Seehafer, A., Kohler, J.: Künstliche Intelligenz: Updates für das Produkthaftungsrecht? EuZW pp. 213–218 (2020)
87. Selbst, A.D.: Negligence and AI's human user. BUL Rev. **100**, 1315 (2020)
88. Selbst, A.D., Barocas, S.: The intuitive appeal of explainable machines. Fordham Law Review **87**, 1085 (2018)
89. Selbst, A.D., Powles, J.: Meaningful information and the right to explanation. International Data Privacy Law **7**(4), 233 (2017)
90. Simonyan, K., Vedaldi, A., Zisserman, A.: Deep inside convolutional networks: Visualising image classification models and saliency maps. In: In Workshop at International Conference on Learning Representations (2014)
91. Smuha, N.A., Ahmed-Rengers, E., Harkens, A., Li, W., MacLaren, J., Piselli, R., Yeung, K.: How the eu can achieve legally trustworthy ai: A response to the european commission's proposal for an artificial intelligence act. Available at SSRN (2021)
92. Strandburg, K.J.: Adjudicating with Inscrutable Decision Tools. MIT Press (forthcoming) (2021)
93. Sunstein, C.R.: Simpler: The Future of Government. Simon & Schuster (2013)
94. Toke, M.J.: Restatement (Third) of Torts and Design Defectiveness in American Products Liability Law. Cornell Journal of Law and Public Policy **5**(2), 239 (1996)
95. Veale, M., Borgesius, F.Z.: Demystifying the draft eu artificial intelligence act—analysing the good, the bad, and the unclear elements of the proposed approach. Computer Law Review International **22**(4), 97–112 (2021)
96. Wachter, S., Mittelstadt, B., Floridi, L.: 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 (2017)
97. Wachter, S., Mittelstadt, B., Russell, C.: "Counterfactual explanations without opening the black box: Automated decisions and the GDPR.". Harvard Journal of Law & Technology **31**, 841 (2018)
98. Wachter, S., Mittelstadt, B., Russell, C.: Why Fairness Cannot Be Automated: Bridging the Gap Between EU Non-Discrimination Law and AI. Computer Law & Security Review (forthcoming) (2021)
99. Wagner, G.: Robot Liability. In Liability for Artificial Intelligence and the Internet of Things. Nomos Verlagsgesellschaft mbH & Co. KG (Feb 2019)
100. Wagner, G.: Commentary on § 3 ProdHaftG, in: Münchener Kommentar zum BGB. BECK, Munich, 8th ed. edn. (2020)
101. Wagner, G.: Commentary on § 823 BGB, in: Münchener Kommentar zum BGB. BECK, Munich, 8th ed. edn. (2020)
102. Wendehorst, C.: Strict liability for AI and other emerging technologies. Journal of European Tort Law **11**(2), 150–180 (2020)

103. Wischmeyer, T.: "Artificial intelligence and transparency: opening the black box". In Regulating artificial intelligence, pp. 75–102. Springer, Cham (2020)
104. Zarsky, T.Z.: Transparent Predictions. U. Ill. L. Rev. p. 1503 (2013)
105. Zech, H.: Künstliche Intelligenz und Haftungsfragen. ZfPW pp. 198–219 (2019)
106. Zehlike, M., Hacker, P., Wiedemann, E.: Matching code and law: achieving algorithmic fairness with optimal transport. Data Mining and Knowledge Discovery **34**(1), 163–200 (2020)