International Review of Law, Computers & Technology

ISSN: (Print) (Online) Journal homepage: www.tandfonline.com/journals/cirl20

How should an explanation be? A mapping of technical and legal desiderata of explanations for machine learning models

Alejandra Bringas Colmenarejo, Laura State & Giovanni Comandé

To cite this article: Alejandra Bringas Colmenarejo, Laura State & Giovanni Comandé (02 May 2025): How should an explanation be? A mapping of technical and legal desiderata of explanations for machine learning models, International Review of Law, Computers & Technology, DOI: 10.1080/13600869.2025.2497633

To link to this article: https://doi.org/10.1080/13600869.2025.2497633

Published online: 02 May 2025.

INTERNATIONAL REVIEW OF LAW, COMPUTERS & TECHNOLOGY
https://doi.org/10.1080/13600869.2025.2497633

How should an explanation be? A mapping of technical and legal desiderata of explanations for machine learning models

Alejandra Bringas Colmenarejo [a*], Laura State [b,c*] and Giovanni Comandé [d]

[a]School of Law, University of Southampton, Southampton, United Kingdom; [b]Department of Computer Science, University of Pisa, Pisa, Italy; [c]Scuola Normale Superiore, Pisa, Italy; [d]Sant'Anna School of Advance Studies, Pisa, Italy

ABSTRACT

Machine learning (ML) systems are abundant in our world. However, most of these systems are not understandable, which poses several challenges, including their safety, proper functioning and accountability. Further, ML models are susceptible to social biases, which can lead to unjust and discriminatory situations. The field of eXplainable Artificial Intelligence (XAI) attempts to answer these challenges by providing explanation methods for ML models. However, there is still an open debate about the necessary desiderata of such methods, including the often-missing consideration of the legal side of explanation desiderata. In this work, we put forward a set of five technical and five legal desiderata of XAI and develop a multi-layered mapping encompassing the dynamics among and between the two sets and linking them to actual requirements. From the standpoint of legality, we rely on the European requirements explainability and justificability. In our mapping, we draw the interdependencies and the intersections between the technical and legal desiderata, creating an image that visualises the assessment of the technical and legal driving forces (desiderata matching requirements) in the design and provision of explanations. Ultimately, explainability and justificability desiderata must be systematic; understood as a dynamic, circular and iterative process.

ARTICLE HISTORY
Received 24 April 2024
Accepted 22 April 2025

KEYWORDS
Automated decision-making systems; eXplainable AI; Artificial Intelligence Act

Introduction

Machine learning (ML) models are abundant in everyday life and help us to navigate the world (Willson 2019). They span the range between relatively simple applications, such as movie recommendations on Netflix (Sharma and Dutta 2020) or language translation on DeepL (Kamaluddin et al., 2024), to highly complex and safety-critical applications such as treatment recommendations (Chen et al. 2018; Sahoo et al., 2019), loan applications

CONTACT Alejandra Bringas Colmenarejo Alejandra.Bringas-Colmenarejo@soton.ac.uk; ax.bringas@gmail.com
University of Southampton, Room 2063, Building 4, Highfield Campus, Southampton, SO17 1BJ, United Kingdom
*These two authors share first co-authorship.

(Hurley and Adebayo 2017; Shi et al. 2022), or self-driving cars (Rao and Frtunikj 2018). Another important application is generative models, spearheaded by the GPT-3 model (Gupta et al. 2023). Many of these applications are promising to simplify and improve our lives. However, since most ML models operate in a high-dimensional, non-linear space and depend on a huge number of parameters, it is often impossible to fully comprehend and understand them, and obscuring their use (Burrell 2016; De Laat 2018; Lepri et al. 2018; Pedreschi et al. 2019).

Such opaqueness and lack of understandability pose a challenge for ML models in both day-to-day and highly consequential scenarios (Lyons, Velloso, and Miller 2021). As has been shown in many contexts, ML models are susceptible to social biases (Ntoutsi et al. 2020). For example, biased ML models were identified in a gender recognition software that incorrectly classified female black faces at a much higher rate than male white faces (Buolamwini and Gebru 2018) in a judicial ADM system that unjustly denied parole to black defendants because of the colour of their skin (Angwin et al. 2016), or were used as commercial prediction algorithms which were more likely to refer white people than black people to care programmes intended for patients with complex medical needs, even if they were equally sick (Obermeyer et al. 2019).

The field of *eXplainable Artificial Intelligence* (XAI) (Adadi and Berrada 2018; Barredo Arrieta et al. 2020) investigates possible methods to 'make [AI systems'] behaviour more intelligible to humans by providing explanations' (Gunning et al. 2019). This can be understood as an attempt to answer some of the challenges posed by ML models while maintaining their high performance levels (Adadi and Berrada 2018). XAI is concerned with constructing explanations for ML models and is situated at the intersection between various disciplines, among others, computer science, social science, and law (Longo et al. 2024; Mohseni, Zarei, and Ragan 2021). In essence, XAI attempts to achieve two objectives: (1) 'produce more explainable models, while maintaining a high level of learning performance (prediction accuracy)' (Turek 2018), and (2) 'enable human users to understand, appropriately trust and effectively manage the emerging generation of artificially intelligent partners' (Turek 2018). Achieving both goals is not an easy task, and different concepts shape the landscape and contribute to the field of XAI (Doshi-Velez and Kim 2017).

In this work, we put forward a set of five technical and five legal desiderata[1] of XAI and develop a mapping between these two sides. While there is already some work on technical desiderata of XAI, there is, to the best of our knowledge, little work considering legal desiderata.[2] We attempt to fill this gap and address how legal desiderata can encourage researchers in computer science and practitioners to work on specific explanation methods. We find the development of concrete desiderata to ensure the compliance of XAI methods to the applicable legal framework of high practical relevance and utmost importance.

Contributions

The main contributions of this work are as follows:

(1) We summarise the debate about the technical desiderata of an explanation.

(2) We discuss in-depth the legal desiderata of an explanation and extract five key desiderata from this discussion, linking them where possible to actual compliance duties under the Artificial Intelligence Act. To the best of our knowledge, we are among the first to do this.
(3) We map both technical and legal desiderata onto each other and create a three-layered mapping that can guide researchers and practitioners in the field of XAI or AI regulation.

It is worth stressing that our contributions have gained further relevance after the enactment of the Artificial Intelligence Act (European Parliament and the Council 2024), as some of these explainability desiderata are now part of legal requirements[3] established in European Regulation, applicable under specific circumstances.[4]

Structure of this work

This paper is structured as follows: first, we present a technical introduction to the field of XAI, followed by the presentation of technical desiderata. Then, we introduce the field of XAI from a legal point of view, followed by the set of legal desiderata. We continue by presenting our mapping between both the technical and legal sides. The paper is closed with a discussion of the limits of our work and open challenges and with a short conclusion.

Explainability, ML and technical desiderata

In this section, we will outline our thoughts on the terms explainability and interpretability from a technical perspective. Further, we will introduce a taxonomy of explanations that will help us to navigate the existing body of literature. However, a full survey is beyond the scope of this paper.

Definition

Providing an explanation for a decision that is supported or fully rendered by a machine learning model (an 'Assisted or Automated Decision-Making System', short ADM system[5]) is not only a matter of justifying the decision for legal (Bibal et al. 2021), ethical (Balasubramaniam et al. 2020; Doshi-Velez and Kim 2017), or safety reasons (Doshi-Velez and Kim 2017; Gilpin et al. 2018) but a huge technical challenge (Longo et al. 2024) itself.

But what exactly is an *explanation*, and what is the difference with an *interpretation*? While the (technical) XAI community has not (yet) reached a consensus on that (Lipton 2018; Mittelstadt, Russell, and Wachter 2019), we present here our working definitions for this paper. We understand interpretability as a closely interlinked but different concept to explainability. For interpretability, we adopt the definition by Doshi-Velez and Kim (2017, 2): 'the ability to explain or to present in understandable terms to a human', e.g. a text or a phenomenon (here: an ML model). Opposed, 'explainability is about an interaction, or an exchange of information' (State 2021, 2) and implies a certain degree of a logical connection between the outcome of the ML model and the explanation offered. Thus, while interpretability can be considered a (model) property

displaying the real meaning of the model, an explanation refers to an action that elucidates why the model acts as it acts. It also matters who the specific *end-user* of the explanation is and for what reasons or for which purpose the explanation is given (Doshi-Velez and Kim 2017). For example, an explanation provided to a loan applicant to understand why an application got rejected and to explore actions to change that decision in a future application must look fundamentally different from an explanation given to a data scientist who oversees debugging the model that evaluates the same loan application. As such, explanations are highly *context-dependent* (State 2021).

Finally, an interpretation is supposed to be 'universal' and is not targeted to a specific end-user, while more than one explanation can refer to the same interpretation of a text or phenomenon. An interpretation illustrates something as it is. For example, the expression 'I love you' is interpreted to express feelings towards someone. Yet, it can be explained as expressing parental affection in the context of a parent–children relationship, while it is interpreted differently among partners and explained accordingly. The significance of the interpretations is universal (e.g. it applies to all parent–child relationships). Still, it needs to be explained in different ways according to the end-user (a little child or an adolescent): the interpretations require different explanations to a child or an adult, for example. While this necessarily holds from the legal perspective, it is not strictly for a technical one - here, it depends on which notion of *interpretability* and *explainability* is used. However, from our point of view, an explanation is always specific to the end-user.

Dimensions

The field of XAI follows different approaches. Here, we introduce a taxonomy broadly based on Guidotti et al. (2018) and Molnar (2019), two well-known publications in the field of XAI. While there are considerable efforts to construct and improve the quality of inherently interpretable models (*white boxes* (Molnar 2019)), others focus on the construction of explanations for models that are not interpretable (*black boxes* (Guidotti et al. 2018; Molnar 2019)). According to the same literature (Guidotti et al. 2018; Molnar 2019), only a few machine learning models are inherently interpretable or understandable to a human. These recognised interpretable models are the following: linear models, decision trees, rule lists, and decision sets. These models are also important in the field of XAI - e.g. by often serving as basic building blocks to construct explanations.[6] Regarding explanations, we can distinguish between *model-agnostic* (Adadi and Berrada 2018*;* Molnar 2019*)* and *model-specific* (Adadi and Berrada 2018*;* Molnar 2019) approaches. As its name suggests, model-agnostic approaches can be applied to any type of model, as opposed to model-specific explanations that work only for specific categories of models. Thus, by construction, the first type has a higher degree of usability with respect to the models to which it can be applied. Last, we differentiate between *local* and *global* explanations (Adadi and Berrada 2018; Guidotti et al. 2018; Molnar 2019). Local explanations are only valid for a specific data instance (Adadi and Berrada 2018; Guidotti et al. 2018; Molnar 2019), which corresponds in most cases to the data subject under decision. For a single data instance, it can be assumed that the decision boundary in its close neighbourhood is simple and can be approximated well by the explanation, as opposed to the global decision boundary, which can be arbitrarily complex (Pedreschi

et al. 2019). Global explanations focus on the full model at once; they 'describe the average behaviour of a machine learning model' (Molnar 2019).

The discussion of technical desiderata below was developed with a focus on post-hoc explainability.[7] However, these concepts, in general, scale beyond.[8] Furthermore, with the emergence of foundational models, more specific taxonomies and desiderata are needed and emerging, see, e.g. Birhane et al. (2023), Floridi (2023), and Zini and Awad (2022). Thus, while these models are beyond the scope of this article and require further investigation, we assume that the basic principles of our analysis would apply to them as well.

Technical desiderata

In this section, we summarise the desiderata of an explanation that are given from a technical point of view. Our analysis led to the conclusion that necessary desiderata ('What does the end-user expect from the explanation?') are strongly interlinked with the evaluation of explanations ('Are those expectations met?').[9] Thus, we must also review how each of these desiderata can be measured.

Here, we confine ourselves to the five main desiderata we could identify in our research. Based on three key publications (Chen et al. 2022; Guidotti et al. 2018; Molnar 2019), i.e. widely known publications among technical XAI researchers, we created an initial list of desiderata and then went systematically through related papers. We extended the list with extracted desiderata per paper and decided to present the desiderata in this paper that appeared in the intersection of all papers considered and that relate most closely to the technical side of explanations. However, our overview is *necessarily incomplete,* as a full review was out of scope.[10]

Please note that in this section, we talk strictly about the technical side of explanations; thus, all terms used refer solely to their technical meaning. The mentioned analysis led to identifying the following desiderata.[11]

- Complexity (or comprehensibility/interpretability): How understandable is the explanation to the end-user?[12] (Belle and Papantonis 2021; Chen et al. 2022; Guidotti et al. 2018; Molnar 2019) An example of a typical measure of complexity is the number of premises in an explanatory rule,[13] that is, the number of conditions that need to be satisfied for the consequence of the rule to be valid, with the consequence reflecting the prediction of the explained ML model. A lower number is preferred. This property is also closely linked to insights from the social sciences, demanding that explanations should be 'selective' (Miller 2019; Mittelstadt, Russell, and Wachter 2019).
- Fidelity (or faithfulness): How well does the explanation approximate the machine learning model? (Belle and Papantonis 2021; Chen et al. 2022; Guidotti et al. 2018; Molnar 2019) Fidelity is computed as the ratio of correct predictions over all predictions, thereby comparing the prediction of the explanation and the explained ML model (e.g. Guidotti et al. 2018). High fidelity is always desired, as it corresponds to an explanation that better approximates the ML model, with the best possible fidelity value being one (compared to zero). Because post-hoc explanations approximate the ML model, they can never perfectly mimic it; thus, their fidelity remains below one.

- Accuracy: How well does the explanation work for a novel data point?[14] (Belle and Papantonis 2021; Guidotti et al. 2018; Molnar 2019) Accuracy can be measured by computing the ratio between correct predictions over all predictions, comparing the prediction of the explanation with the original labels (e.g. Guidotti et al. 2018). Questions related to this are: how does the explanation interact with a data point that is out-of-distribution? Does the explanation display probability values? By construction, the accuracy of a post-hoc explanation is always lower than the accuracy of the machine learning model.
- Robustness (or sensitivity, stability): How similar are explanations for two different data points? (Chen et al. 2022; Guidotti et al. 2018; Molnar 2019) This measure depends on a formalised notion of similarity.[15] Intuitively, we do expect that similar data points receive a similar explanation unless there are good reasons not to expect this, e.g. when the predictions are not similar (Molnar 2019).
- Homogeneity: How does faithfulness change across different (sub-)groups?[16] (Chen et al. 2022) This property is closely linked to fairness considerations, especially regarding the explanation itself. For example, a recent study (Balagopalan et al. 2022) found that explanations can show different faithful values for different subgroups. A subgroup is thereby constructed by separating data based on a sensitive feature such as age or gender. If explanations show different faithful values for different subgroups, this can be another source of bias.

All the above-listed desiderata can be quantitatively measured. However, evaluations of explanations must be complemented by qualitative measures - in particular, by user studies (Miller, Howe, and Sonenberg 2017). A survey on user studies can be found here (Rong et al. 2022).

Further, as explanations themselves depend on the context, the exact formulation of the desiderata do as well. A complete evaluation is thus only possible if the explanation is situated within its final context (including the integration of the end-user's needs). The list of the above desiderata should, therefore, be seen as a basic set of desiderata. Concrete formulations of each property, how they should be implemented and evaluated, and the negotiation of additional desiderata or an omission of some must take place in each individual case (Figure 1).

Explainability, ML and legal desiderata

In this section, we will address the notion of explainability[17] from a legal perspective, ultimately considering the legal requirements of explainability and justificability.[18] Further, we will propose a set of legal desiderata, expected to be fulfilled by an explanation to comply with the pertinent legal provision.

Definition

Besides the technical benefits that more interpretable and explainable artificial intelligence algorithms offer, the push towards explainability and interpretability also responds to a legal demand on algorithmic governance (Katzenbach and Ulbricht 2019; Issar and Aneesh 2022). As stated above (see *technical section*), the so-called black-box nature of

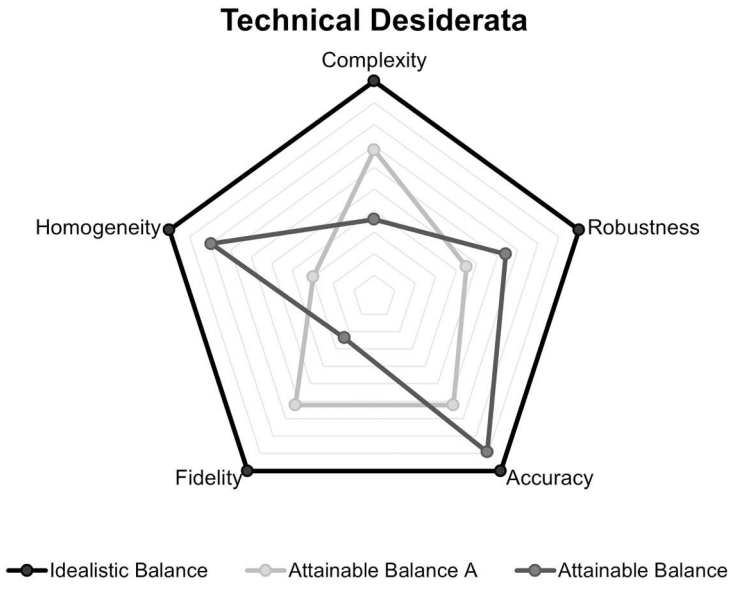

Figure 1. Example of different mappings of technical desiderata. Nodes are labelled with the five identified desiderata. Ideally, all desiderata are fully accomplished (black). In reality, explanations are not ideal. For demonstration, we depict two of such explanations (grey and light grey lines). Here, the degree to which a desideratum is satisfied by the explanation is randomly assigned.

these models raises unprecedented opacity challenges, which encourage calls for accountability and transparency.

In relation to the matter at hand, the European Union developed a legal framework which lays down transparency and understandability at the centre of algorithmic and AI systems' governance.[19] Within this heterogeneous framework, the use of assisted or fully automated decision-making processes based on ML models requires, to some degree, explanations and justifications about the final decision and the decision-making process. For example, the most novel and extent development in this regard is found in the Artificial Intelligence Act, which requires a sufficiently granular level of technical interpretability and explainability to assess the risks of those AI systems presumed high-risk or to demonstrate that presumed high-risk systems are not high-risk under the Artificial Intelligence Act.[20] By extension, assisted or fully automated decision-making processes which fall under the category of high-risk will be expected to respond to more extensive levels of legal explainability and justificability requirements.

Before expounding these requirements in the *dimensions* section, we find it necessary to explain the definitions of explainability and justificability that guide our analysis, as well as the distinction between normative and motivating reasons.

A normative reason 'is a consideration that counts in favour of someone's actions' (Scanlon 2000, 18). In other words, normative reasons justify or make it right for someone to act in a certain way (Logins 2022). Therefore, legally justifying a decision requires proving its correctness, fairness, and lawfulness as referred to in the appropriate laws, norms, and principles. As maintained by Malgieri (2021, 19), a legal *justification* of an

[automated] decision: 'means not merely explaining the logic and the reasoning behind it, but also explaining why it is a legally acceptable (correct, lawful, and fair) decision'.

Motivating reasons, on the contrary, are reasons that either count in favour of the agents' actions or explain their behaviour (Alvarez and Way 2024). We can distinguish between reasons that motivate and reasons that explain. The former addresses the motivations of the decision-maker and their beliefs regarding the reality at hand, while the latter exposes the connection between the knowledge that is available prior to the decision and the following effect (Malgieri 2021). In other words, motivating reasons refer to the subjective knowledge or belief that the decision-maker had about some concrete facts at the moment of making a certain decision. In contrast, explanatory reasons allude to the actual facts and the relationship of cause and effect between those facts and the final result or action (Alvarez and Way 2024). In legal terms, the notion of *explanation* contains both meanings, i.e. the provision of information that 'attempts to render a state of affairs, an event or a process understandable' (Aarnio 1986, 4:22) under a motivating reasons perspective. Therefore, if a decision resulted from an algorithmic decision-making process, its explanation shall disclose the connection between the input data and the final decision or the intentions and objectives that motivated such a decision (Malgieri 2021).

In consequence, explanations are descriptive and intrinsically grounded on the ADM system with the goal of allowing individuals to understand a single decision or the whole system. Meanwhile, justifications are normative and extrinsic, intended to assess the legality and validity of the decision. Justifications can demonstrate that the decision is grounded on the pertinent rule of law, against which the legality and validity of the decision will be assessed (Henin and Le Métayer 2021).

Given the above differentiation between explanations and justifications, we conceive *legal explainability* as a set of legal information requirements that specify the rationale and motivation of ADM systems. Likewise, we envision *legal justificability* as the set of information requirements directed to demonstrate the normativity, lawfulness, and legitimacy of ADM systems as a whole.[21] In other words, explainability is about explaining how an ADM system reached the decision but does not clarify whether that decision was made in a legally compliant way. On the contrary, justificability shows that the applicable legal requirements have been satisfied, both regarding whether the decision was made in a certain way and whether it fulfils the legal reason or conditions for that type of decision. Justificability would require showing, for instance, that the specific AI system complies with the requirements of the Artificial Intelligence Act and that the data used for its training were lawfully collected and used according to other applicable legislations (e.g. GDPR). as referred to in Articles 8 paragraphs 1 and 2 and Article 11 paragraph 1 in the case of the Artificial Intelligence Act.

Another key distinction needs to be highlighted to understand legal explainability and justificability: the timing of the information provided with respect to the decision affecting an individual. *Ex-ante -information- obligations* arise right before any decision is made (Edwards and Veale 2017; Malgieri and Pasquale 2022; Wachter, Mittelstadt, and Floridi 2017), usually covering the normative reasons (justifications) for the use of the ADM system but sometimes also addressing the motives behind the existence of the ADM process. *Ex-post -information- obligations* appear after the decision has been made and implicitly allow individuals to contest the specific decision (Crabtree, Urquhart,

and Chen 2019; Goodman and Flaxman 2017; Gyevnar and Ferguson 2023; Mostowy 2020) with the information received about the decision's causality (explanatory reasons) and motivation (motivating reasons). Although in practice this distinction is not set in stone, whether explainability and justificability relate to a single decision or the whole decision-making process is a matter of serious relevance. For instance, Article 22 of the GDPR and Article 86 of the Artificial Intelligence Act establish, respectively, a right to explanation of individual decisions solely based on automated processing or on the basis of the output from a high-risk AI system (*ex-post -information- obligations*), while information duties under Article 13 and 14 of the GDPR refer to the logic involved in the automated decision-making and the envisaged consequences of such processing (both *ex-ante and ex-post -information- obligations)*.

Dimensions

Even before the European Union enacted the Artificial Intelligence Act in June 2024, calls for explainability and justificability came from specific legal fields covering both public and private law.[22] In recent years, the greatest debate around these requirements arose after the publication of the European General Data Protection Regulation -hereafter GDPR (European Parliament and the Council 2016)- which includes the so-called right to an explanation and right to information.[23] Although in this paper we will refrain from addressing the legal basis of each right, we argue that the former ought to focus on the explainability aspect of the decision, while the latter on its justificability. However, these rights do not apply equally to all decisions that are based on individuals' personal information. The level of automatisation and the severity of the effects of a decision (Binns and Veale 2021) determine the level of explainability and justificability demanded. Hence, the GDPR is a clear example of how different interests and objectives can coexist in a legal norm, dictating different requirements.

Consumer protection law also embraces explainability and justificability requirements with the perspective of empowering individuals in the unbalanced relations between consumers and private actors and enabling individuals with the appropriate actions. Whether the specific activities of the private actors contemplated in consumer protection law should be considered an ADM system would require further discussion. However, what seems arguable is that these activities require, at some level or another, the profiling of the individual and the use of data-driven algorithmic systems, which normally lead to either an assisted or an automated decision that affects the individual.[24]

For instance, the Regulation (EU 2019/1150) on promoting fairness and transparency for business users of online intermediation services compels in Article 5 providers of online intermediation services - envisaged as e-commerce marketplaces, online software application services, and online social media services - to set out 'in their terms and conditions the main parameters determining ranking and the reasons for the relative importance of those main parameters as opposed to other parameters' (European Parliament and the Council 2019b). The same Regulation also obliges providers of online search engines to set out in plain and intelligible language the 'main parameters, which individually or collectively are most significant in determining raking and the relative importance of those main parameters' (European Parliament and the Council 2019b) in an easy and publicly available description. Compliance with these transparency

requirements does not require the provision of any information regarding the algorithms in use that could enable consumer deception or harm through manipulation of search results, as referred to in Article 5, paragraph 6. Furthermore, Recital 25 of the Regulation clarifies that the description of the parameters determining ranking, which might include the number and type of main parameters, should offer enough information to ensure that the consumer obtains an adequate understanding of how the ranking system works.

The Directive (EU 2019/2161) on the better enforcement and modernisation of Union consumer protection rules - The Modernisation Directive - also introduces transparency requirements in the Directive (2005/29/EC) concerning unfair business-to-consumer commercial practices in the internal market and Directive (2011/83/EU) on consumer rights. On the one hand, the new Article 4 (a) of the Directive concerning unfair business-to-consumer commercial practices in the internal market establishes that traders, when providing the consumer with the possibility to search for products offered by different traders or by consumers on the basis of a query, shall be made available general information on the 'main parameters determining the ranking of products presented as a result of the search query and the relative importance of those parameters, as opposed to other parameters' (European Parliament and of the Council 2019). Further, Recital 21 of the Modernisation Directive clarifies that these 'transparency requirements' aim to ensure adequate transparency towards the consumer about the parameters that determine ranking, specifying in Article 22 that those parameters mean 'any general criteria, processes, specific signals incorporated into algorithms or other adjustment, or demotion mechanisms used in connection with the ranking' (European Parliament and of the Council 2019). To ensure that consumers understand ranking functionality, the information shall be provided succinctly, easily, and prominently and shall be directly available, as referred to in Recital 22 of the Directive. Recital 23 determines that traders are not prescribed to present a customised description of each query, nor must they disclose the detailed functioning of their ranking mechanism. A general description of the main parameters determining the ranking that explains the main default parameters used by the trader and their relative importance as opposed to other parameters would be enough to ensure compliance, at least until the AI literacy[25] duties come into force in February 2025 since the AI literacy requested will need to take into account, as per Article 3 paragraph 56 'skills, knowledge and understanding that allow […] affected persons, taking into account their respective rights and obligations in the context of [the Artificial Intelligence Act], to make an informed deployment of AI systems, as well as to gain awareness about the opportunities and risks of AI and possible harm it can cause'.

On the other hand, the amended Article 6 (a) point II of the Directive on consumer rights introduces a new information requirement for distance and off-premises contracts. Concretely, the new provision determines that consumers shall be warned about price personalisation based on automated decision-making before any such type of contract binds them. Recital 45 of the Directive on the better enforcement and modernisation of Union consumer protection rules highlights that consumers should be clearly informed of the existence of price personalisation based on profiling or automated decision-making so that they can ponder the risk in their purchase decision. Likewise, Recital 45 clarifies that this right to information does not conflict with the right not to be subject to automated individual decision-making as referred to in Article 22 of the GDPR.

The Proposal for a regulation on preventing the dissemination of terrorist content online advances a series of transparency obligations towards the automated removal of content. Article 6 of the Proposal recognised as a proactive measure to protect against the dissemination of terrorist content the use of automated tools which can either (1) prevent the re-upload of content which has previously been removed or to which access has been disabled because it is considered to be terrorist content or (2) detect, identify and expeditiously remove or disable access to terrorist content (European Commission 2018). In this regard, proposed Article 8 compels hosting service providers to, where appropriate, set out in their terms and conditions a meaningful explanation of the functioning of proactive measures, including the use of automated tools. In Article 9, the Proposal declares that the use of automated tools shall need to be accompanied by effective and appropriate safeguards ensuring that the decisions concerning the content, particularly those removing or disabling content, are accurate and well-founded. Recital 17 highlights the relevance of avoiding unintended and erroneous decisions when the detection of content has been made using automated means. Recital 26 calls attention to the right of users to ascertain the reasons upon which the content uploaded by them has been removed or to which access has been disabled, according to Article 19 of the Treaty of the European Union and Article 47 of the Charter of fundamental rights of the European Union regarding the rights to an effective remedy and a fair trial. In this sense, providers are asked to make available meaningful information that enables affected users to contest the decision. In its Recitals, the Proposal highlights its intent to contribute to the protection of public security while ensuring the protection of fundamental rights, including 'the rights to respect for private life and to the protection of personal data, the right to effective judicial protection, the right to freedom of expression, including the freedom to receive and impart information, the freedom to conduct a business, and the principle of non-discrimination'.

Regulation (EU 2024/1698) laying down harmonised rules on artificial intelligence and amending Regulations (EC) No 300/2008, (EU) No 167/2013, (EU) No 168/2013, (EU) 2018/858, (EU) 2018/1139 and (EU) 2019/2144 and Directives 2014/90/EU, (EU)2016/797 and (EU) 2020/1828 - The Artificial Intelligence Act - establishes the most extensive explainability and justificability requirements to date for AI systems in the form of AI literacy, transparency and information obligations, and human-oversight, which in consequence establish a quasi-general obligation for interpretable and explainable AI systems – per the definitions offered in the *technical definition section*.

While in the Artificial Intelligence Act the AI literacy requirements -as referred to in Article 4- apply to all AI systems with no exception -extending explainability and justificability requirements well beyond high-risk AI systems-, specific explainability duties attach to high-risk[26] and to General purpose AI models[27] and the 'right to explanation of individual decision-making' are more limited.

Indeed, the Artificial Intelligence Act includes a right to explanation for decisions made on the basis of an output from a specific class of high-risk AI system. Concretely, Article 86 paragraph 1 reads as follows:

> any affected person subject to a decision which is taken by the deployer on the basis of the output from a *high-risk AI system listed in Annex III, with the exception of systems listed under point 2 thereof*, and which produces legal effects or similarly significantly affects that person in

> a way that they consider to have an adverse impact in their health, safety and fundamental rights shall have the right to obtain from the deployer clear and meaningful explanations of the role of the AI system in the decision-making procedure and the main elements of the decision taken (European Parliament and the Council 2024).

Upon this provision, the Artificial Intelligence Act grants affected subjects, but only in these specific instances, with the right to request from the deployer of a high-risk AI system a clear and meaningful explanation of the role of the decision-making procedure, the main parameters of the decision taken and the related input data. Interestingly, Article 86 paragraph 3 asserts that the mentioned right only applies to the extent that the right is not already provided for under other EU legislation. The other legislation mentioned above will, consequently, need to be considered when implementing the Artificial Intelligence Act and its own right to explanation.

The Artificial Intelligence Act established a general requirement for interpretable and explainable AI systems as referred to in the technical section of this paper. Artificial Intelligence Act's Article 4 compels providers and deployers of AI systems to take measures to ensure the sufficient level of AI literacy of their staff, other persons dealing with the operation and use of AI systems, and the individuals affected by them. As per Article 3 paragraph 56 AI literacy means:

> Skills, knowledge and understanding that allow providers, deployers and affected persons, taking into account their respective rights and obligations in the context of this Regulation, to make an informed deployment of AI systems, as well as to gain awareness about the opportunities and risks of AI and possible harm it can cause (European Parliament and the Council 2024).

Accordingly, the measures mentioned in Article 4 to ensure AI literacy must involve interpretability and explainability as understood in technical terms - see *technical definition section*. However, the knowledge and understanding of the AI system will by no means be homogenous, rather the explanations and justifications offered to each actor will depend on the scope of their respective duties and rights as established in the Artificial Intelligence Act. For example, Article 13 of the Artificial Intelligence Act specifies transparency and information duties to deployers of high-risk AI systems to 'ensure that their operation is sufficiently transparent to enable deployers to interpret the system's output and use it appropriately' (European Parliament and the Council 2024). Among other information, high-risk AI systems would need to be accompanied by instructions indicating '[…] its intended purpose, the level of accuracy including its metrics, robustness, and cybersecurity, its performance regarding specific persons or groups of persons on which the system is intended to be used, or specifications about the input data […]' (European Parliament and the Council 2024). Likewise, Article 14 impels high-risk AI systems to be designed and developed in a manner that ensures effective oversight by a natural person - also possible via appropriate human-machine interfaces -, the proper understanding of the relevant capabilities and limitations of the high-risk AI system, and the correct interpretation of its output. Explainability and justificability - as per our legal understanding - are unequivocally linked to these obligations and rights as the persons involved in the operation of AI systems will need to understand the rationale and motivation of the AI system as well as its normativity, lawfulness, and legitimacy.

Moreover, as anticipated, the (limited) right provided by Article 86 'shall apply only to the extent that the right referred to in paragraph 1 is not otherwise provided for under Union law'. Thus, its explainability and justificability require a case-by-case approach considering the information rights provided for by Union law; for instance, the rules mentioned above were applicable to high-risk AI systems referred to in Article 6 paragraph 2.

Yet, per Article 4 of the Artificial Intelligence Act, these duties need to be fine-tuned 'taking into account their [staff and other persons dealing with the operation and use of AI systems on their behalf] technical knowledge, experience, education and training and the context the AI systems are to be used in, and considering the persons or groups of persons on whom the AI systems are to be used'.

When considering the requirements for interpretable and explainable AI systems under European laws, we cannot forget to mention the EU's general principles and primary law, which already set interpretability and explainability constraints for the use of AI systems. For instance, the constitutional value of the rule of law, recognised in Article 2 of the Treaty of the European Union, requires that everyone is treated equally and rightfully by all decision-makers and has the right to challenge the decision through fair proceedings. Therefore, public authorities and institutions are compelled to explain and justify administrative and judicial decisions also when algorithms are fully or partially involved. On the other hand, the European Union and Member States' administrations are bound by the right to offer a reasoned decision (European Commission for Democracy through Law (Venice Commission) 2016), which requires administrative decisions to be accompanied by an explanation of the factual and legal grounds that motivate the judgements. This reasoning needs to be offered clearly and precisely, reflecting the real reasons and motives for the decision (Venice Commission 2016, 26). Article 41 of the Charter of Fundamental Rights of the European Union states that 'every person has the right to have his or her affairs handled impartially, fairly and within a reasonable time by the institutions, bodies, offices, and agencies of the Union. This right includes '[…] the obligation of the administration to give reasons for its decisions' (The European Union 2012). The use of automated decision-making by no means evades the need for such an administrative motivation (Citron 2017; Demková 2023). The Hague District Court's Rechtbank De Haag Case (2020) recognised – as the first Court on an EU member state – that the lack of transparency of an ADM system used by a public actor can put in risk individuals' interests if it does not offer sufficient and verifiable information about the functioning of the system or the risk analysis method. The scoring system considered was an instrument called the Systeem Risico Indicatie, which was used by the Dutch government to detect various forms of fraud (e.g. social benefits, allowances, and taxes fraud). The Hague district court found Systeem Risico Indicatie unlawful given that the Dutch state had not offered enough information to allow individuals to understand the decision or convince them that it was made according to the pertinent laws.

Further, Article 6 of the European Union Charter of Fundamental Rights states that judicial decisions shall respect the rights to a fair trial, due process, and transparency. Besides explaining the factual and legal grounds on which the judicial decision was made, judges need to respond to all the claims and arguments made by the trial parties. Therefore, in the scenario that an ADM system fundamentally contributed to a judgement, the motivation of such a judicial decision would be expected and required as this motivation does

not depend on who made the decision (a human or ADM system) but on the fact that a decision was made.

Beyond the interpretability and explainability requirements that can be drawn from the EU's fundamental principles, rights and values, EU secondary law also delimits the explainability and interpretability of AI systems used by public actors in some instances. The rights to information and an explanation, as referred to in the GDPR, oblige data controllers irrespectively of their public or private nature, whereas Article 5 of the Artificial Intelligence Act explicitly addresses specific instances of the use of AI systems by public actors. For example, the use of 'risk assessments of natural persons in order to assess or predict the risk of a natural person committing a crime offence, based solely on the profiling of a natural person or on assessing their personality traits and characteristics' is defined as a prohibited practice. Yet, when the assessment is not solely based on an AI system, it is considered high-risk as referred to in Annex III paragraph 6 (d). Furthermore, other AI systems which are used in the context of access to enjoyment of essential public services and benefits; law enforcement; migration, asylum and border control management; and the administration of justice and democratic processes, are also classified as high-risk, as stated in Annex III paragraphs 5, 6, 7, and 8 respectively. Examples of these types of high-risk AI systems include those used to evaluate the eligibility of a natural person for healthcare; to profile a natural person in the course of the detection, investigation or prosecution of criminal offences, to examine applications for asylum, visa or residence permits, or to research and interpret facts and the law and apply the law to a concrete set of facts. The Artificial Intelligence Act has, therefore, delimited the permitted use of AI systems by public authorities and has granted a limited right to explanation to decisions made by a public authority based on the output of a high-risk AI system when they consider having an adverse impact on their health, safety or fundamental rights.[28] Likewise, all information and transparency obligations required for the use of AI systems -covered above- are required regardless of the public or private nature of the deployer.

The previous analysis suggests, among other things, that differences in the interpretability and explainability requirements w.r.t the Artificial Intelligence Act are based on the risk classification of the AI system rather than on the private or public nature of the actors involved with it. Although the use of concrete AI systems for public authors is considered high-risk or directly prohibited, the actual requirements on explainability and interpretability do not change if the actor is private or public. However, EU public actors are bound by general principles of law and fundamental rights that set specific explainability and interpretability requirements not applicable to private actors. Hence, we can argue that public actors are expected to comply with an extra layer of explainability and interpretability compared to private actors (Demková 2023).

It is true, however, that -although in need of further research-, another differentiation emerges according to the possible target of an explanation and justification. If the target is a consumer in a business-to-consumer transaction, the legal desiderata for consumer protection and its corresponding general legal principles would step in as requirements, eventually requiring further tailoring of the explanation. Similar distinctions might emerge among public administrations since it is possible to assume that the actions of public administrations would impact fundamental rights in different ways requiring

different levels of explainability[29] (Demková 2023). For instance, the impact of an ADM system is different if a public administration is deciding about the allocation of a right or permission to build or if the administration is adjudicating a right or deciding on a crime.

Legal desiderata

In this section, we present the list of five legal desiderata resulting from our assessment of the European legal framework on explainability and justificability. The desiderata were obtained as follows: we started by observing common desirable desiderata of legal explainability in the work of Bibal et al. (2021); Hacker and Passoth (2022); Lognoul (2020). To the best of our knowledge, these works are the firsts to survey and synthesise requirements on XAI systems in the European legal framework, covering both public and private law instead of limiting their research to a specific area of law, such as health (Amann et al. 2020), public administration (Olsen et al. 2019), or law enforcement (Raaijmakers 2019). Thereupon, we reconsidered Malgieri (2021, 16), who stands up for *just* ADM systems, which are only possible 'through a practical justification statement and process through which the data controller proves', why the AI systems are not unfair, not discriminatory, not obscure, not unlawful, etc. With this distinction between explainability and justificability requirements in mind, we re-examined the European laws addressing ADM systems (see dimensions above) and put forward the legal desiderata. Sector-specific desiderata (e.g. for public administrations and for consumers) are not addressed in this article beyond the laws discussed above, thus the list of desiderata shall be understood as a first approximation, which might need to be re-considered on a case-by-case basis.

- Substantive Desiderata: invoke the rights, duties, obligations, and causes of action derived from legal explainability and justificability requirements.
  - Normativity: every decision is embedded in a context regulated by various fields of law. This norm specification needs to be pondered in the explanation and justification of the decision. This means tailoring the scope of the information that will be provided to the requirements of the law, i.e. to offer general information about the main features, general information about the system functionality and main features, specific information on all the features used in the decision and of their combinations, or specific information on all components of the decision-making system. While these information requirements can seem quite straightforward at first sight, interpreting these formulations in a manner that agrees with technical concepts and approaches towards explainability can be a great challenge. Indeed, one needs to acknowledge the XAI approaches and the distinction between interpretation and explanation set forth in the technical definitions and dimensions sections presented above. Additionally, ADMs do not operate in siloed environments but in situations affected by multiple laws (e.g. data protection, consumer law, finance products, etc.). Thus, the legal interpretations, which are related to the different legal rules, need to be considered before defining the specific legal desiderata in any given case more granularly. This implies that legal desiderata need to functionally reflect the actual legal requirements. The information offered in such scenarios has

to comply with multiple requirements whose coordinated interpretation is a preliminary requirement requesting appropriate legal skills.[30]

- Purposefulness: explainability and justificability requirements are determined by several factors concerning the decision-making process, such as the degree of automation, who is the decision-maker (a public authority or a private firm) (Bibal et al. 2021), who is the individual affected by such a decision, in which context the decision is taken, and which are the potential effects and risks for the individuals and the society (Hacker and Passoth 2022; Lognoul 2020). The combination of these factors in an algorithmic decision-making process brings forth different explainability and justificability requirements which respond to various legal objectives and interests, among which Sovrano et al. (2022), Hacker and Passoth (2022), and Bibal et al. (2021) have identified the protection of individuals towards potential risks and harms, the provision of enabling actions and rights, the compliance with the relevant obligations, the building and increase of trust in machine learning algorithms, the enhancements of market's and sectors' functioning, and the improvement of regulatory oversight. These purposes need to be satisfied by explanations and justifications.[31]

- Procedural / Formal Desiderata: specify the rules and the methods used to ensure explainability and justificability rights and obligations.
  - Truthfulness: The information provided needs to be accurate, truthful, and complete.[32] Explainability and justificability requirements are rightfully constrained by intellectual property rights and legitimate business interests. Likewise, explanations and justifications should be appropriate to the particular ADM system and specific to the norm. These conditions, however, do not excuse the manipulation of the information to achieve scheming or cunning purposes.
  - Intelligibility: the language and formulation used and the presentation chosen for the explanation and justification (text, graphs, images, figures) need to ensure its understandability and plain clarity. This desideratum also relates to the tension between accuracy and interpretability, meaning that a complex, information-rich explanation is often hard to understand for the end user and, therefore fails to fulfil its main purpose.[33] Therefore, an explanation must navigate a trade-off between being easily understandable and sufficiently detailed (Malgieri and Comandé 2017).
  - Accessibility: information with explainability or justificability aims must be easily, prominently, and adequately available. In general, obtaining such information should not be hindered or obstructed, although it can be directly and publicly accessible or only accessible to interested parties by default or upon request.[34]

Contrary to our judgment of technical desiderata, we consider legal desiderata intrinsically qualitative desiderata that do not permit a quantitative analysis or evaluation and yet certainly require context to determine their degree of accomplishment. Accordingly, the proposed desiderata should be considered as principles that need to be assessed in each specific scenario (Figure 2).

Analysis: How should an explanation be?

We started this work by addressing the technical debate over explainability and interpretability, clarifying that the former is an action between the end-user and the explanation,

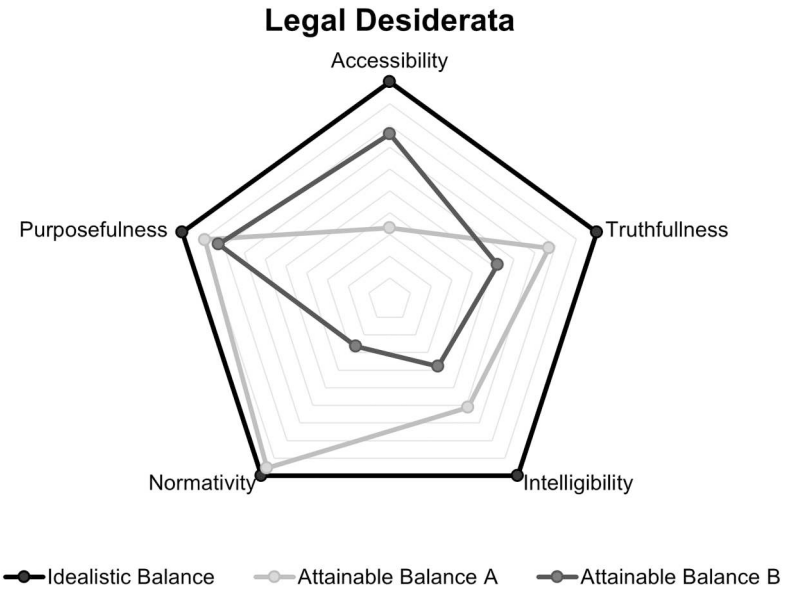

Figure 2. Example of different mappings of legal desiderata. Nodes are labelled with the five identified desiderata. Ideally, all desiderata are fully accomplished (black curve). In reality, explanations are not ideal. For demonstration, we depict two of such explanations (grey and light grey lines). Here, the degree to which a desideratum is satisfied by the explanation is randomly assigned.

with the aim to provide information about the workings of an ML model, whereas the latter is a property of an ML model. With that in mind, we proposed five technical desiderata that need to lead the construction and provision of explanations about ML models. In the subsequent section, we discussed the legal approach toward the explainability of ADM systems, assessing the distinction between explanation and justification requirements. Explanations are descriptive, intending to make the decision-making process and the reached decision understandable for the interested party. Justifications, on the contrary, are normative, aiming at exposing the lawfulness, fairness, and legality of the decision. Henceforth, we put forward a set of two types of legal desiderata (i.e. substantive and procedural) that we contemplate as paramount when considering legal compliance towards such requirements.

In this section, we create a mapping between the technical and legal side of explainability, answering the following questions: Which of the presented desiderata from the technical side aligns with which legal desiderata, and vice versa? How strong is the overlap between these two points of view? How are they connected, and where lie (possible) tensions? We acknowledge, however, that this mapping contains some limitations, starting from how comprehensive the different desiderata are. Legal desiderata encompass both the decision and the decision-making process, while technical desiderata -in many cases- concern only the model used to make a decision within a larger decision-making process. In other words, the legal desiderata that we have identified cover a broader range of circumstances affecting the ADM system than the proposed technical desiderata. Consequently, the overlay between legal and technical desiderata is not identical, nor does it aim to be, as the object addressed by each desideratum is intrinsically different (Figure 3).

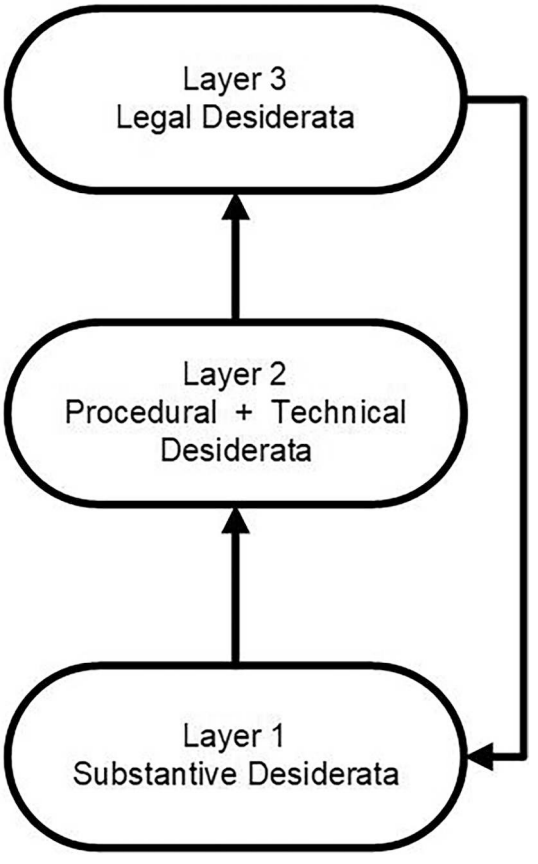

Figure 3. Visualisation of the different layers of desiderata of explanation.

First layer: substantive (legal) desiderata to determine what explainability method should be chosen

The method used to complete a task needs to be decided based on the goal to achieve. Here, we assume that developing and providing explanations about an ADM system responds to some legal requirement on explainability and justificability that shall be fulfilled. Therefore, legal substantive desiderata need to be considered as the first layer of this mapping. Whereas normativity offers the more basic criterion to determine the details that would encompass the explanations, purposefulness delimits the goals to be pursued with such explanations. Consequently, choosing the appropriate explanation method is closely connected to the application context and how the (legal) substantial desiderata materialise. In that sense, substantial desiderata can be seen as the basis for the process of developing an explanation scenario and determining 'what' the content of the final explanation is.

Specifying the information that needs to be provided is also central to choosing the explainability method that best connects all the technical desiderata. Some methods are well-studied and characterised in terms of their (technical) desiderata and limitations (e.g. stability for LIME (Alvarez-Melis and Jaakkola 2018)). Methods are also restricted by

their input data type (not all developed methods work on any type of data), by what they produce as an output (e.g. plots, plain numbers, decision rules), whether they operate locally or globally, are applicable to any model or only some, and at which stages they are applied to the decision pipeline (Guidotti et al. 2018; Molnar 2019; Speith 2022). The output of a method is thereby crucial: e.g. whether it provides the weights of the main features that determine the decision, decision rules of the whole system, or contrastive data examples is important to map to the correct level of normativity.

Resultantly, normativity dictates which methods shall be used and which shall not, as it specifies the desirable information and explanation established in the body of the law.

Second layer: technical and procedural (legal) desiderata to determine how explanations and information should be

We can draw a close connection between (all) technical desiderata and procedural (legal) desiderata since they both aim to answer the question of 'how' an explanation should be, focusing on its concrete content and the conditions and standards it shall meet. However, a one-to-one mapping would fail to live up to the complexity of the described desiderata. Legal explainability requires truthfulness, which in turn demands technical accuracy, fidelity, robustness, and homogeneity of the explanations, as they are indispensable to ensure the correctness of an explanation. That said, legal explainability is also restricted by the need for intelligibility, intimately connected to the technical property of complexity. Both desiderata relate to the understandability of an explanation and the trade-off that an explanation must navigate between being succinct and interpretable but sufficiently complex and accurate to reveal all important details. Further, the notion of understandability is similar to the notion of legibility developed around the information duties of the GDPR concerning automated decision-making (Articles 13, 14, 22) (Malgieri and Comandé 2017) and the notion of AI literacy introduced in the Artificial Intelligence Act (Article 11).

Third layer: legal desiderata to confirm the choice and restart if needed

Once an explanation of an ADM system has been developed and the overlap between legal and technical desiderata has been evaluated, the logical step is to reconsider the decisions made in the light of substantive legal desiderata (see also Figure 3). The technical explanations obtained in the second stage should offer some insights regarding the internal logic of the ADM system, thus, responding partly to the legal explainability requirements. Likewise, some of the information provided through technical means can also unravel some of the intentions or motives behind the decision as well as offer the appropriate reasons to justify the fairness, lawfulness, and correctness of the process and the final decision.

Upon this, one should consider that even in the implausible scenario of finding the perfect balance between the proposed technical desiderata, legal explainability, and justificability desiderata would remain in a dynamic equilibrium. Therefore, legal substantive desiderata must be reconsidered under the light of the developed ADM's explanations and rebalanced to assess which other information - for example, information about the fairness of the decision-making process - would be required to offer the appropriate

explanations and justifications. This is highly coherent with the legal way of reasoning in which added information may change the legal results.

How can an explanation be?

This paper evolved around the question *How should an explanation be?* both from a technical and legal point of view. While certainly based on state-of-the-art (technical) explanation methods and current legal European frameworks, we developed our two sets of desiderata based on an ambitious idea of well-developed and idealistic explanations. This is also needed -otherwise, a mapping of *something desired as essential* does not make sense (Merriam n.d.). However, a critical check of the current reality is also necessary, and we do so by asking *How can an explanation be?* However, it is fair to ask the question only under a certain condition: *given the current state of development*. Here, we must acknowledge that the field of XAI (excluding some work in the 80s on expert systems (Confalonieri et al. 2021)) is a relatively new field, in active development and testing different pathways of which only some can be successful in the long term. Some exemplary problems are an unclear mathematical definition of a neighbourhood for local explanation methods and issues of robustness (see e.g. for LIME (Alvarez-Melis and Jaakkola 2018; Molnar 2019)), and an (often) missing evaluation of explanations on use cases and via user studies (Doshi-Velez and Kim 2017; Miller, Howe, and Sonenberg 2017; Murdoch et al. 2019). A non-technical issue of explanations is fair-washing, i.e. providing explanations that are fairer than the original ML model, as demonstrated for rule lists (Aivodji et al. 2019). Such fair-washing makes it possible, for example, to use explanations towards the profit-oriented interests of a company and not towards the rights and interests of the individual affected by the decisions of that company.

It is also worth mentioning that the legal explainability scheme and desiderata presented in this paper encompass transparency, accountability, and information provisions to those developing, providing, or using ADM systems, as well as the legal framework of individual rights to people impacted by these systems. Although applicable to both perspectives, the balance of each property is highly dependent on the applicable legal framework and the intended end-user of the explanation. Indeed, we have already emphasised the importance of the context and circumstances where the ADM takes place for the provision of appropriate and adequate explanations and justifications, albeit here we want to stress the relevance of who is the provider and the recipient of the information.

Accessibility relates to how findable an explanation is. This property is not inherent to the explanation (as most of the other desiderata), but a consideration on a higher level that might necessitate other technical means (e.g. provide a web interface to the explanation, write an easy-to-understand introduction on the web page). Further, it closely ties to considerations about intellectual property rights, customer rights and the intention a provider of an explanation is pursuing.

In essence, the ideal balances presented in the figures above (see Figures 1 and 2) are not more than unfeasible property mappings that would not respond to any real case. When considering the balance between technical and legal desiderata, real scenarios will end up offering a final image where some desiderata prevail over others, as seen in the attainable examples shown in the figures.

Despite these limitations, the need to bridge legal and technical desiderata for explainability and justificability purposes is urgent to address and becomes even more urgent to approach given the emergence of foundational models trained on even larger amounts of data. Control techniques that emerged so far (including the desiderata put forward in this paper) may need an extension, for example, to also include information on which data a model was trained. Such an extension of our work is left for the future.

The enactment of the Artificial Intelligence Act has furthered the need to ensure that legal and technical definitions of explainability align and complement each other. Previously established legal requirements for the explainability of ADM systems -i.e. the GDPR or the Modernisation Directive- were highly focused on individuals' rights to information and an explanation. The - at the publication time of this article - the newly adopted Artificial Intelligence Act established a general obligation for AI literacy and understandability, which undoubtedly requires the interpretability and explainability of the system - from a technical point of view - as well as its explainability and justificability - from a legal perspective. The ponderance and balance of the desiderata presented in this paper, then, turn out to be a legal obligation rather than a desirable action.

In this article, we focus on desiderata for explanations, with the specific starting point of post-hoc explanations from the technical perspective. While the Artificial Intelligence Act is putting forward a limited 'duty to explain for high-risk AI systems', it does not provide many details about how such an explanation should look or which (technical) XAI method should be used. This is in line with Walke et al. (2023).

Still, the Artificial Intelligence Act promotes largely the concept of AI literacy that applies to all AI systems and not only to high-risk systems. AI literacy must be ensured by both providers and deployers for 'staff and other persons dealing with the operation and use of AI systems on their behalf' and 'considering the persons or groups of persons on whom the AI systems are to be used'. The AI literacy concept assumes building the competencies needed to understand the explanations and is, therefore, indirectly influencing explainability.

We can identify two consequences from the approach towards AI explainability and interpretability adopted by the Artificial Intelligence Act. On the one hand, some requirements in the Artificial Intelligence Act that are specified for the AI system (and not explicitly for the explanation) may be also applicable to the explanation (or the integrated system of explanation and AI). On the other hand, the technical desiderata specified above may be interpreted more broadly. Further, these desiderata would need to be extended to the new interpretation of explanations and AI systems that are integrated with the explanation, and some desiderata put forward may be no longer applicable (e.g. the 'fidelity' of an explanation is only applicable to post-hoc explanations). This broader interpretation of an explanation, its desiderata as well as such an associated reading of the Artificial Intelligence Act, however, remains an instance of future work.

Our analysis *How should an explanation be?* suggest that the explainability and justificability of ADM systems need to be systemic and iterative. The interplay between the legal and technical desiderata leads to the conclusion that their interactions can change the desiderata themselves. For instance, a technical explanation may reveal the need, possibility, or impossibility of relying on a specific legal basis for the given action (e.g. data processing) or trigger different rights and safeguards (e.g. right to information or to an explanation). Similarly, legal desiderata which impose a certain degree of explainability

or justificability with certain levels of intelligibility (Malgieri and Comandé 2017) may compel the selection of a different or more specific explanation method (e.g. global or local explanation) or might even stimulate the development of a new method. In essence, there are potential interactions between the ADM systems and the law that require explainability and justificability to be systematic; understood as a dynamic, circular, and iterative process. This idea gives way to the potential development of a management system allowing the interplay between technical and legal desiderata for the explainability and justificability of ADM systems. This is, however, an avenue that will need to be addressed in future research and that might be useful in the actual implementation of the risk management approach required by the Artificial Intelligence Act.

Our analysis proposes research opportunities for the XAI community, including the development of better transparent-by-design models and explanations that follow our above-defined technical and legal mapping of desiderata. In the end, more interpretable ADM systems and novel explainability methods can redress the balance of the desiderata and enhance compliance with explainability and justificability requirements. Such a development will also help to overcome the argument that the importance and usefulness of legal provisions on explainability and justificability is negligible, given their lack of technical feasibility.

Conclusion

In this work, we put forward a set of five technical and a set of five legal desiderata of explanations, answering the question *How should an explanation be?* From a technical side, we found the following desiderata: complexity, fidelity, accuracy, stability, and homogeneity. We studied the European framework of explainability and justificability and extracted two substantive and three procedural desiderata (i.e. five legal desiderata), namely normativity and purposefulness, as well as truthfulness, intelligibility, and accessibility. We developed a mapping between these two points of view and found that these desiderata complement each other in a multi-layered fashion. Ultimately, explainability and justificability requirements must be systematic; understood as a dynamic, circular and iterative process.

A great challenge in the XAI community is to work towards interdisciplinary goals. Here, a critical point is to correctly communicate with each other, with a special focus on common terminology and similar ad-hoc objectives, since different disciplines tend to associate different meanings with the same term, consequently diverging in their intentions and purposes. An example of this unfortunate disagreement has been presented in this paper in relation to the term explanation. While in the technical domain, an explanation about ADM systems is relatively straightforward and can fulfil different purposes (understanding, trust, debugging, legal obligations), from a legal perspective, it is not only of high importance to make a distinction between explanation and justification (and between normative and motivating reasons) but also acknowledging that borders between one and the other might not be as clear-cut as wished for. In essence, part of the challenges existing within the XAI community is this miscommunication between disciplines where the use of the same terminology might not entail the understanding of the same definitions or the achievement of the same goals. The

mapping of, and necessary balance between, technical and legal desiderata in real-life scenarios can thus be convoluted if this discordance is not overturned.

Notes

1. We use 'desiderata' to describe how (here) the explanation should ideally be, and which aspects and features it should fulfil.
2. See the work of Bibal et al. (2021); Hacker and Passoth (2022); Lognoul (2020).
3. We use 'functional/legal requirement' to refer to any obligation imposed by law in regard with explainability. If no functional legal basis can be identified, we rely on the term desiderata.
4. We pinpoint to footnotes 11, 12, 14, and 16 for a proper list of articles in the Artificial Intelligence Act that now demand technical desiderata as legal requirements.
5. In the technical literature, 'ADM system' can refer both to a fully or partly automated decision-making system. Sometimes, 'ADM system' is used but the details are not defined. Preference may be also given to the specific name of the system/model (e.g., naming the type of ML model) over the general term 'ADM system'. Our definition is similar to Pedreschi et al. (2019); 'black box AI systems for automated decision making, often based on machine learning over (big) data'). From a legal perspective, we refer to The Information Commissioner Officer (ICO) that defines automated decision-making as 'the process of making a decision by automated means without any human involvement. These decisions can be based on factual data, as well as on digitally created profiles or inferred data' (n.d.). By default, an assisted decision-making is the type of information system that provides support to individuals or organizations in making a decision. The Guideline on Automated Decision-Making and Profiling of Article 29 Working Party furthers the scope of the concept of an automated decision-making and clarifies its correct interpretation in regard to Article 22 of the GDPR (Article 29 Working Party 2020). For a more detail and practical discussion on the differences between assisted and automated decision-making see Binns and Veale (2021).
6. Examples are LIME (Ribeiro, Singh, and Guestrin 2016) (based on a linear model) and LORE (Guidotti et al. 2019) (based on a decision tree).
7. Post-hoc explanations can be defined as follows: 'explaining a (plausibly opaque) model after it was trained' (Speith 2022, 2).
8. The property that is unique to post-hoc explainability is 'fidelity'. As we state also below, these desiderata are meant as a starting point, thus deviations/extensions are possible. The focus on post-hoc explainability is based, among others, on its popularity (e.g. Wachter, Mittelstadt, and Russell 2018).
9. Questions 'What does the end-user expect from the explanation?' and 'Are those expectations met?' were formulated by the authors to fulfil the goal of this paper.
10. We also point the interested reader towards a couple of related survey papers: (Adadi and Berrada 2018; Barredo Arrieta et al. 2020; Belle and Papantonis 2021; Chen et al., n.d.; Confalonieri et al. 2021; Guidotti et al. 2018; Langer et al. 2021; Molnar 2019; Murdoch et al. 2019; Sokol and Flach 2020; Vilone and Longo 2021a).
11. Few specific connections between these desiderata and the Artificial Intelligence Act can be identified. We point to Article 11, 13, 14, 53 and 86, as well as Annex IV, providing some general connections to technical explainability. Below, we indicate via footnotes the match we found between a technical desideratum and a statement in the Artificial Intelligence Act.
12. This desideratum can be broadly matched with Recital 72, Article 13 paragraph 2 and Article 11 paragraph 1 from the Artificial Intelligence Act.
13. For example, Setzu et al. (2021) used this measure to evaluate a novel explanation method, Vilone and Longo (2021b) discussed the measure more broadly for rule-based explanations.
14. Accuracy can be broadly matched with Article 13 paragraph 2 from the Artificial Intelligence Act.

15. There are different ways to formalize similarity, the adopted definition will depend on various factors such as the type of data.
16. Homogeneity can be broadly matched with Article 13 paragraph 3 (v) from the Artificial intelligence Act, which requires high-risks AI systems to be accompanied by instructions containing information about 'when appropriate, its performance regarding specific persons or groups of persons on which the system is intended to be used'.
17. We argue that technical interpretability and explainability do not have direct counterparts in the European legal framework and legal academic literature. To our view, interpretability is an intrinsic technical term that is sometimes used interchangeably with that of 'understandability' and even 'transparency' in legal and social science contexts. Moreover, when the 'understandability' and 'transparency' of a system is expected or demanded, the law does not seem to be concerned with whether it is achieved through technical interpretability or explainability. In this paper we focus on the desiderata of explanations for ML systems, rather than in desiderata for ML systems in general. For this reason, we consider that the appropriate comparison should be made between technical interpretability and explainability and legal of explainability, englobing the latter the notions of explainability and justificability. We dwell on the differentiation of these last two concepts in the *definitions section*.
18. Later on defined in the *definition section*, we envision *legal justificability* as the set of information requirements directed to demonstrate the normativity, lawfulness, and legitimacy of ADM systems as whole.
19. For example, Annex IV paragraph 2 (c) of the Artificial Intelligence Act requires for the technical documentation for high-risk AI systems to provide 'the description of the system architecture explaining how software components build on or feed into each other and integrate into the overall processing; the computational resources used to develop, train, test and validate the AI system'. Likewise, Annex XI paragraph 2 (3) requires for general purpose AI providers to deliver technical documentation to 'where applicable, a detailed description of the system architecture explaining how software components build or feed into each other and integrate into the overall processing'.
20. Limited to the cases under Article 6 paragraph 3 of the Artificial Intelligence Act.
21. Our definition of *justificability* foster from the notion of *just* algorithms introduced by Malgieri (2021) on the basis of which society want a sustainable environment of desirable AI systems, we should aim not only at transparent explainable, fair, lawful, and accountable algorithms, but we also should seek for 'just' algorithms, that is, automated decision-making systems that include all the above-mentioned qualities (transparency, explainability, fairness, lawfulness, and accountability). We understand that legal justificability requirements for ADM systems seek to create a framework to assess the legality and validity of the decision, inevitable demanding *just* algorithms and ADM systems.
22. The scope of this article is limited to ML systems in the form of ADM systems and the explainability requirements existing around decisions resulted from their use. However, we cannot delimit the analysis of the law – *legal dimensions section*- only to ML systems as such, rather we need to acknowledge that machine learning is just one type of artificial intelligence and that the technology behind an ADM system does not necessarily need to be machine learning. For this reason, in the *legal dimension section* we refer to the particular terminology used by the pertinent law whether it is, for instance, 'ADM', 'AI system', 'ranking product', or 'profiling and automated decision making'. We understand that under any of these notions, ADM systems based on machine learning are inevitably included, whereas the opposite cannot be said. To avoid any inaccuracy and misunderstanding we will use the particular wording used in each law as we consider they can all apply to an ADM system as referred in this article.
23. An extensive academic debate took place around the existence, enforceability, and effectiveness of the right to an explanation for individual automated decisions and its differences and similarities to the right to information about automated decision-making processes, see for reference Edwards and Veale (2017), Malgieri and Comandé (2017), Mendoza and Bygrave

(2017), Pedreschi et al. (2019); Selbst and Powles (2017), Wachter, Mittelstadt, and Floridi (2017), and Kaminski and Urban (2021).

24. The Guideline on Automated Decision-Making and Profiling of Article 29 Working Party (Article 29 Working Party 2020) engaged in the distinction between profiling and automated decision-making. Although concerned with the notion of profiling and automated decision-making as referred to in Article 22 of the GDPR, the discussion and analysis could be transposed to our argument. See also, *Case C-634/21, SCHUFA Holding (Scoring)* [2023] ECLI EU:C:2023:957, Request for a Preliminary Ruling.
25. Pursuant to recital 20 of the Artificial Intelligence Act, AI literacy 'should equip providers, deployers and affected persons with the necessary notions to make informed decisions regarding AI systems. Those notions may vary with regard to the relevant context and can include understanding the correct application of technical elements during the AI system's development phase, the measures to be applied during its use, the suitable ways in which to interpret the AI system's output, and, in the case of affected persons, the knowledge necessary to understand how decisions taken with the assistance of AI will have an impact on them […]'.
26. For instance, Article 13 paragraph 3 established a transparency and information provision to deployers, which includes, among other things, the high-risk AI system's technical capabilities and characteristics that are relevant to explain its outputs -as per paragraph 3 (b)(iv).
27. As referred to, for instances, in Annexes XI and XII of the Artificial Intelligence Act.
28. The right to explanation as per Article 86 of the Artificial Intelligence Act prescribes the right to obtain from the deployer clear and meaningful explanations of the role of the AI system in the decision-making procedure and the main elements of the decision taken.
29. Providers of an AI system that in accordance to Article 6 of the Artificial Intelligence Act is considered high-risk can argue otherwise if they believe their systems does not pose a significant risk to people's health, safety and rights. An assessment and documentation is required before the sell and use of the system.
30. For instance, just focussing on the normativity requirements of the Artificial Intelligence Act we could mentioned Article 12 paragraph (c) requiring logging capabilities of high-risk systems to provide, at minimum 'the input data for which the search has led to a match', Article 13 paragraph 2 which demands high-risk to be accompanied by instructions for use […] that include concise, complete, correct and clear information that is relevant, accessible and comprehensible to deployers, or Article 13 paragraph 3 (b) which requires such instructions of high-risk AI systems to contain, among other things, 'its intended purpose' or 'when appropriate, information to enable deployers to interpret the output of the high-risk AI system and use it appropriately'. Likewise, Annex IV paragraph 2 (b) of the Artificial Intelligence Act stipulates as part of the technical documentation referred to in Article 11 paragraph 1, 'the design specifications of the system, namely the general logic of the AI system and of the algorithms; the key design choices including the rationale and assumptions made, including with regard to persons or groups of persons in respect of who, the system is intended to be used; the main classification choices; what the system is designed to optimise for, and the relevance of the different parameters; the description of the expected output and output quality of the system; the decisions about any possible trade-off made regarding the technical solutions adopted to comply with the requirements set out in Chapter III, Section 2'. Annex VIII of the Artificial Intelligence Act specifies the information to be submitted upon registration of the high-risk AI systems, including among other 'a description of the intended purpose of the AI system and of the components and functions supported through this AI system' or 'a basic and concise description of the information used by the system (data, inputs) and its operating logic'. Likewise, Annex XI paragraph 2(b) requires information of general-purpose AI models, including 'the design specifications of the model and training process, including training methodologies and techniques, the key design choices including the rationale and assumptions made; what the model is designed to optimise for and the relevance of the different parameters, as applicable'.

31. For instance, the limited right to an explanation as referred to in Article 86 of the Artificial Intelligence Act 'should be clear and meaningful and should provide a basis on which the affected persons are able to exercise their right' -as per Recital 171-. Recital 71 of the Artificial Intelligence Act specified how 'having comprehensible information on how high-risk AI systems have been developed and how they perform through their lifetime is essential to enable traceability of those systems, verify compliance with the requirements under this Regulation, as well as monitoring of their operations and post market monitoring'. Articles 13 paragraph (b)(viii), Article 14 paragraph (4) and Article 53 paragraph 1 (b) (i) further specify the purpose of the explainability and justificability requirements established around AI systems in accordance with the Artificial Intelligence Act.
32. Article 11 and 53 of the Artificial Intelligence Act, for example, require technical documentation to demonstrate the compliance of an AI system with the requirements set in the Regulation, hence indirectly demanding the truthfulness and intelligibility of the explanations that could be used to prove the conformity with the law. Moreover, Article 13 of the Regulation requires high-risk systems to be accompanied by instructions including 'concise, complete, correct and clear information that is relevant, accessible and comprehensible to deployers'. Such information could not be provided without truthful information and explanations about the high-risk AI systems, nor could be the technical documentations mentioned in Annex IV paragraph 2 (b).
33. A requirement for intelligibility inevitable rise from the AI literacy required in Article 4 of the Artificial Intelligence. The same could be said, for instance, in regard to Article 4, 13, and 14 of the Artificial Intelligence Act, which sought to ensure the interpretability and understandability of the AI system for third parties involved in its used. Likewise, intelligibility is expected from Article 86 in regard to how explanations of the role of the AI system in the decision-making procedure shall be 'clear and meaningful'.
34. For instance, Article 22 of the GDPR and Article 86 of the Artificial Intelligence Act impose to set up a procedure to address the exercise of the right to an explanation in accordance with the specific circumstances establish is each Regulation.

Acknowledgements

This work has received funding from the European Union's Horizon 2020 research and innovation programme under Marie Sklodowska-Curie Actions (G.A. 860630) for the project 'NoBIAS – Artificial Intelligence without Bias' and (G.A. 956562) for the project 'Legality Attentive Data Scientist'. Furthermore, the work of L. State has been partly funded by PNRR - M4C2 – Investimento 1.3, Partenariato Esteso PE00000013 – ''FAIR - Future Artificial Intelligence Research'' – Spoke 1 ''Human-centered AI'', funded by the European Commission under the NextGeneration EU programme. This work reflects only the authors' views and the European Research Executive Agency (REA) is not responsible for any use that may be made of the information it contains.

Funding

This work was supported by European Commission [grant numbers: 860630, PE00000013 and 956562].

Author contribution

Alejandra Bringas Colmenarejo: major intellectual contribution to the conception of the paper, editing, and revision for paper's intellectual content, and final approval of the version to be published; **Laura State**: first idea and drafting, major intellectual contribution to the conception of the paper, editing and revision for paper's intellectual

content, and final approval of the version to be published; **Giovanni Comandé**: editing and revision for paper's intellectual content, paper's supervision, and final approval of the version to be published.

ORCID

Alejandra Bringas Colmenarejo http://orcid.org/0000-0002-7968-9853
Laura State http://orcid.org/0000-0001-8084-5297
Giovanni Comandé http://orcid.org/0000-0003-2012-7415

References

Aarnio, Aulis. 1986. *The Rational as Reasonable: A Treatise on Legal Justification*. vol. 4. Dordrecht: D. Reidel Publishing Company.

Adadi, Amina, and Mohammed Berrada. 2018. "Peeking Inside the Black-box: A Survey on Explainable Artificial Intelligence (XAI)." *IEEE Access* 6:52138–60. https://doi.org/10.1109/ACCESS.2018.2870052.

Aivodji, Ulrich, Hiromi Arai, Olivier Fortineau, Sébastien Gambs, Satoshi Hara, and Alain Tapp. 2019. "Fairwashing: The Risk of Rationalization." In *Proceedings of the 36th International Conference on Machine Learning*, 161–70. Long Beach, CA: PMLR. https://proceedings.mlr.press/v97/aivodji19a.html.

Alvarez-Melis, David, and Tommi S. Jaakkola. 2018. "On the Robustness of Interpretability Methods." *CoRR abs/1806.08049*. http://arxiv.org/abs/1806.08049.

Alvarez, Maria, and Jonathan Way. 2024. "Reasons for Action: Justification, Motivation, Explanation." In *The Stanford Encyclopedia of Philosophy*, edited by Edward N. Zalta and Uri Nodelman, Fall 2024. Metaphysics Research Lab, Stanford University. https://plato.stanford.edu/cgi-bin/encyclopedia/archinfo.cgi? entry=reasons-just-vs-expl&archive=win2017.

Amann, Julia, Alessandro Blasimme, Effy Vayena, Dietmar Frey, Vince I Madai, and Precise4Q Consortium. 2020. "Explainability for Artificial Intelligence in Healthcare: A Multidisciplinary Perspective." *BMC Medical Informatics and Decision Making* 20 (1): 1–9. https://doi.org/10.1186/s12911-019-1002-x.

Angwin, Julia, Jeff Larson, Surya Mattu, and Lauren Kirchner. 2016. "Machine Bias: There's Software Used across the Country to Predict Future Criminals. And It's Biased against Blacks." *ProPublica* 23:77–91.

Article 29 Working Party. 2020. "Guidelines on Automated Individual Decision-Making and Profiling for the Purposes of Regulation 2016/679". *Number WP251rev.01*.

Balagopalan, Aparna, Haoran Zhang, Kimia Hamidieh, Thomas Hartvigsen, Frank Rudzicz, and Marzyeh Ghassemi. 2022. "The Road to Explainability Is Paved with Bias: Measuring the Fairness of Explanations." In *FAccT '22 Proceedings of the 2022 ACM Conference on Fairness, Accountability, and Transparency*, 1194–1206. New York, NY: Association for Computing Machinery. https://doi.org/10.1145/3531146.3533179.

Balasubramaniam, Nagadvya, Marjo Kauppinen, Sari Kujala, Kari Hiekkanen, M. Morisio, M. Torchiano, and A. Jedlitschka. 2020. "Ethical Guidelines for Solving Ethical Issues and Developing AI Systems." In *Product-Focused Software Process Improvement, Lecture Notes in Computer Science*. Vol. 12562, edited by Maurizio Morisio, Marco Torchiano, and Andreas Jedlitschka, 331–346. Cham: Springer. https://doi.org/10.1007/978-3-030-64148-1_21.

Barredo Arrieta, Alejandro, Natalia Díaz-Rodríguez, Javier Del Ser, Adrien Bennetot, Siham Tabik, Alberto Barbado, Salvador Garcia, et al. 2020. "Explainable Artificial Intelligence (XAI): Concepts, Taxonomies, Opportunities and Challenges toward Responsible AI." *Information Fusion* 58 (June): 82–115. https://doi.org/10.1016/j.inffus.2019.12.012.

Belle, Vaishak, and Ioannis Papantonis. 2021. "Principles and Practice of Explainable Machine Learning." *Frontiers in Big Data* 4:688969. https://doi.org/10.3389/fdata.2021.688969.

Bibal, Adrien, Michael Lognoul, Alexandre de Streel, and Benoît Frénay. 2021. "Legal Requirements on Explainability in Machine Learning." *Artificial Intelligence and Law* 29 (2): 149–69. https://doi.org/10.1007/s10506-020-09270-4.

Binns, Rubens, and Michael Veale. 2021. "Is That Your Final Decision? Multi-Stage Profiling, Selective Effects, and Article 22 of the GDPR." *International Data Privacy Law* 11 (4): 319–332. https://doi.org/10.1093/idpl/ipab020.

Birhane, Abeba, Atoosa Kasirzadeh, David Leslie, and Sandra Wachter. 2023. "Science in the Age of Large Language Models." *Nature Reviews Physics* 5(5): 277–280. https://doi.org/10.1038/s42254-023-00581-4.

Buolamwini, Joy, and Timnit Gebru. 2018. "Gender Shades: Intersectional Accuracy Disparities in Commercial Gender Classification." In *Proceedings of the 1st Conference on Fairness, Accountability, and Transparency*. Vol. 81, edited by Sorelle A. Friedler and Christo Wilson, 77–91. *Proceedings of Machine Learning Research*. New York: PMLR.

Burrell, Jenna. 2016 "How the Machine 'Thinks': Understanding Opacity in Machine Learning Algorithms." *Big Data & Society* 3(1): 1–12. https://doi.org/10.1177/2053951715622512.

Chen, Jianguo, Kenli Li, Huigui Rong, Kashif Bilal, Nan Yang, and Keqin Li. 2018. "A Disease Diagnosis and Treatment Recommendation System Based on Big Data Mining and Cloud Computing." *Information Sciences* 435:124–149. https://doi.org/10.1016/j.ins.2018.01.001.

Chen, Zixi, Varshini Subhash, Marton Havasi, Weiwei Pan, and Finale Doshi-Velez. 2022. "What Makes a Good Explanation? A Harmonized View of Properties of Explanations." In *Workshop on Trustworthy and Socially Responsible Machine Learning, NeurIPS 2022*. https://doi.org/10.48550/arXiv.2211.05667.

Citron, Danielle Keats. 2017. "Technological Due Process" *Wash UL Review* 85(6): 1249–1313.

Confalonieri, Roberto, Ludovik Coba, Benedikt Wagner, and Tarek R. Besold. 2021. "A Historical Perspective of Explainable Artificial Intelligence." *WIREs Data Mining and Knowledge Discovery* 11 (1): 1–21. https://doi.org/10.1002/widm.1391.

Crabtree, Andy, Lachlan Urquhart, and Jiahong Chen. 2019. "Right to an Explanation Considered Harmful." *SSRN Electronic Journal*. https://doi.org/10.2139/ssrn.3384790.

De Laat, Paul B. 2018. "Algorithmic Decision-making Based on Machine Learning from Big Data: Can Transparency Restore Accountability?" *Philosophy & Technology* 31 (4): 525–541. https://doi.org/10.1007/s13347-017-0293-z.

Demková, Simona. 2023. *Automated Decision-making and Effective Remedies: The New Dynamics in the Protection of EU Fundamental Rights in the Area of Freedom, Security and Justice*. Cheltenham: Edward Elgar Publishing.

Doshi-Velez, Finale, and Been Kim. 2017. "Towards a Rigorous Science of Interpretable Machine Learning." *arxiv:1702.08608*.

Edwards, Lilian, and Michael Veale. 2017. "Slave to the Algorithm? Why a Right to an Explanation is Probably Not the Remedy You are Looking For." *Duke Law & Technology Review* 16:18–84. https://doi.org/10.2139/ssrn.2972855.

European Commission. 2018. *Proposal for a Regulation of the European Parliament and of the Council on Preventing the Dissemination of Terrorist Content Online*. 2018/0331 (COD).

European Commission for Democracy through Law (Venice Commission). 2016. *The Rule of Law Checklist*.

European Parliament and of the Council. 2019. *Directive (EU) 2019/2161 of the European Parliament and of the Council of 27 November 2019 Amending Council Directive 93/13/EEC and Directives 98/6/EC, 2005/29/EC and 2011/83/EU of the European Parliament and of the Council as Regards the Better Enforcement and Modernisation of Union Consumer Protection Rules. OJ L328*. 2019/2161.

European Parliament and the Council. 2016. *Regulation (EU) 2016/679 on the Protection of Natural Persons with Regard to the Processing of Personal Data and on the Free Movement of Such Data, and Repealing Directive 95/46/EC (General Data Protection Regulation) (2016)*. OJ L 119/1. EU/2016/679

European Parliament and the Council. 2019. Regulation (EU) 2019/1150 of the European Parliament and of the Council of 20 June 2019 on Promoting Fairness and Transparency for Business Users of Online Intermediation Services. OJ L 186/57. *EU/2019/1150*.

European Parliament and the Council. 2024. Regulation (EU) 2024/1689 of the European Parliament and of the Council of 13 June 2024 laying down Harmonised Rules on Artificial Intelligence (Artificial Intelligence Act) and Amending Regulations (EC) No 300/2008, (EU) No 167/2013, *(EU) No 168/2013, (EU)2018/858, (EU) 2018/1139 and (EU) 2019/2144 and Directives 2014/90/EU, (EU)2016/797 and (EU) 2020/1828 (Artificial Intelligence Act)* (2024) *OJ L, EU/2024/1689*

European Union. 2012. Charter of Fundamental Rights of the European Union. *OJ C C326/391. 2012/C 326/02*.

Floridi, Luciano. 2023. "AI as Agency Without Intelligence: On ChatGPT, Large Language Models, and Other Generative Models." *Philosophy & Technology* 36 (1): 15. https://doi.org/10.1007/s13347-023-00621-y.

Gilpin, Leilani H., David Bau, Ben Z. Yuan, Ayesha Bajwa, Michael Specter and Lalana Kagal. 2018. "Explaining Explanations: An Overview of Interpretability of Machine Learning," *2018 IEEE 5th International Conference on Data Science and Advanced Analytics (DSAA)*. Turin, Italy: 80–89. https://doi.org/10.1109/DSAA.2018.00018.

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

Guidotti, Riccardo, Anna Monreale, Fosca Giannotti, Dino Pedreschi, Ssalvatore Ruggieri, and Franco Turini. 2019. "Factual and Counterfactual Explanations for Black Box Decision Making". *IEEE Intell. Syst*. 34(6): 14–23. https://doi.org/10.1109/MIS.2019.2957223.

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

Gunning, David, Mark Stefik, Jaesik Choi, Timothy Miller, Simone Stumpf, and Guang-Zhong Yang. 2019. "XAI—Explainable Artificial Intelligence." *Science Robotics* 4 (37): eaay7120. https://doi.org/10.1126/scirobotics.aay7120.

Gupta, Bulbul, Tabish Mufti, Shahab Saquib Sohai and Dag Øivind Madsen. 2023. "ChatGPT: A Brief Narrative Review". *Cogent Business & Management*, *10* (3): 1–17. https://doi.org/10.1080/23311975.2023.2275851.

Gyevnar, Balint, and Nick Ferguson. 2023. "Aligning Explainable AI and the Law: The European Perspective." *arXiv*. http://arxiv.org/abs/2302.10766.

Hacker, Philipp, and Jan-Hendrik Passoth. 2022. "Varieties of AI Explanations Under the Law. from the GDPR to the AIA, and Beyond." In *International Workshop on Extending Explainable AI Beyond Deep Models and Classifiers*, edited by Andreas Holzinger, Randy Goebel, Ruth Fong, Taesup Moon, Klaus-Robert Muller, and Wojciech Samek, 343–73. Cham: Springer. https://doi.org/10.1007/978-3-031-04083-2_17.

The Hague District Court. 2020. Case SyRI (Systeem Risico Indicatie), ECLI:NL:RBDHA:2020:865. Rechtbank Den Haag (The Hague District Court).

Henin, Clément, and Daniel Le Métayer. 2021. "A Framework to Contest and Justify Algorithmic Decisions." *AI and Ethics* 1 (4): 463–76. https://doi.org/10.1007/s43681-021-00054-3.

Hurley, Mikella, and Julius Adebayo. 2017. "Credit Scoring in the Era of Big Data." *Yale JL & Tech* 18 (5): 148–216.

Information Commissioner Officer. n.d. "What is Automated Individual Decision-making and Profiling?" *Information Commissioner Officer*. Accessed January, 10, 2025. https://ico.org.uk/for-organisations/uk-gdpr-guidance-and-resources/individual-rights/automated-decision-making-and-profiling/what-is-automated-individual-decision-making-and-profiling/.

Issar, Shiv, and Aneesh Aneesh. 2022. "What is Algorithmic Governance?" *Sociology Compass* 16 (1): 1–14. https://doi.org/10.1111/soc4.12955.

Kamaluddin, Mohamad Ihsan, Moch Wildan Khoerul Rasyid, Fourus Huznatul Abqoriyyah, Andang Saehu. 2024. Accuracy Analysis of DeepL: Breakthroughs in Machine Translation Technology. *Journal of English Education Forum (JEEF))*. 4(2): 122–126. https://doi.org/10.29303/jeef.v4i2.681.

Kaminski, Margot E. and Jennifer M. Urban. 2021. "The Right to Contest AI". *Columbia Law Review*, *121*(7): 1957–2048. https://columbialawreview.org/content/the-right-to-contest-ai/.

Katzenbach, Christian, and Lena Ulbricht. 2019. "Algorithmic Governance." *Internet Policy Review* 8 (4): 1–18. https://doi.org/10.14763/2019.4.1424.

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

Lepri, Bruno, Nuria Oliver, Emmanuel Letouzé, Alex Pentland, and Patrick Vinck. 2018. "Fair, Transparent, and Accountable Algorithmic Decision-Making Processes." *Philosophy & Technology* 31 (4): 611–627. https://doi.org/10.1007/s13347-017-0279-x.

Lipton, Zachary C. 2018. "The Mythos of Model Interpretability." *Communications of the ACM* 61 (10): 36–43. https://doi.org/10.1145/3233231.

Logins, Arturs. 2022. *Normative Reasons: Between Reasoning and Explanation*. New York, NY: Cambridge University Press.

Lognoul, Michael. 2020. "Explainability of AI Tools in Private Sector: An Attempt for Systemization." *SSRN Electronic Journal*. https://doi.org/10.2139/ssrn.3685906.

Longo, Luca, Mario Brcic, Federico Cabitza, Jaesik Choi, Roberto Confalonieri, Javier Del Ser, Riccardo Guidotti, et al. 2024. "Explainable Artificial Intelligence (XAI) 2.0: A Manifesto of Open Challenges and Interdisciplinary Research Directions". *Information Fusion,* 106: 1–22. https://doi.org/10.1016/j.inffus.2024.102301.

Lyons, Henrietta, Eduardo Velloso, and Tim Miller. 2021. "Conceptualising Contestability: Perspectives on Contesting Algorithmic Decisions." In *Proceedings of the ACM on Human-Computer Interaction 5 (CSCW1)*: 1–25. https://doi.org/10.1145/3449180.

Malgieri, Gianclaudio. 2021. "'Just' Algorithms: Justification (Beyond Explanation) of Automated Decisions Under the General Data Protection Regulation." *Law and Business* 1 (1): 16–28. https://doi.org/10.2478/law-2021-0003.

Malgieri, Gianclaudio, and Giovanni Comandé. 2017. "Why a Right to Legibility of Automated Decision-Making Exists in the General Data Protection Regulation." *International Data Privacy Law* 7 (4): 243–65. https://doi.org/10.1093/idpl/ipx019.

Malgieri, Gianclaudio, and Frank Pasquale. 2022. "From Transparency to Justification: Toward Ex-ante Accountability for AI." *Brooklyn Law School, Legal Studies Paper, Brussels Privacy Hub Working Paper*, 712 (33). https://doi.org/10.2139/ssrn.4099657.

Mendoza, Isak, and Lee A. Bygrave. 2017. "The Right Not To Be Subject to Automated Decisions Based on Profiling." In *EU Internet Law: Regulation and Enforcement*, edited by Tatiana-Eleni Synodinou, Philippe Jougleux, Christiana Jougleux, Christiana Markou, and Thalia Prastitou, 77–98. Cham: Springer. https://doi.org/10.1007/978-3-319-64955-9_4.

Merriam Webster. n.d. "Dictionary: Desideratum" Merriam-Webster. Accessed September 25, 2024. https://www.merriam-webster.com/dictionary/desiderata.

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

Miller, Tim, Piers Howe, and Liz Sonenberg. 2017. "Explainable AI: Beware of Inmates Running the Asylum Or: How I Learnt to Stop Worrying and Love the Social and Behavioural Sciences." *arXiv:1712.00547*.

Mittelstadt, Brent, Chris Russell, and Sandra Wachter. 2019. "Explaining Explanations in AI." In *FAT* '19 Proceedings of the Conference on Fairness, Accountability, and Transparency*, 279–288. New York , NY: Association for Computing Machinery. https://doi.org/10.1145/3287560.3287574.

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

Molnar, Christoph. 2019. "Interpretable Machine Learning: A Guide for Making Black Box Models Explainable."

Mostowy, Walter A. 2020. "Explaining Opaque AI Decisions, Legally." *Berkeley Technology Law Journal* 35: 1291. https://doi.org/10.2307/27121777.

Murdoch, W. James, Chandan Singh, Karl Kumbier, Reza Abbasi-Asl, and Bin Yu. 2019. "Interpretable Machine Learning: Definitions, Methods, and Applications." *CoRR abs/1901.04592*. https://doi.org/10.48550/arXiv.1901.04592.

Ntoutsi, Eirini, Pavlos Fafalios, Ujwal Gadiraju, Vasileios Iosifidis, Wolfgang Nejdl, Maria-Esther Vidal, et al. 2020. Bias in Data-driven Artificial Intelligence Systems—An Introductory Survey. *WIREs Data Mining and Knowledge Discovery* 10 (3): 1–14. https://doi.org/10.1002/widm.1356.

Obermeyer, Ziad, Brian Powers, Christine Vogeli, and Sendhil Mullainathan. 2019. "Dissecting Racial Bias in an Algorithm Used to Manage the Health of Populations." *Science* 366 (6464): 447–53. https://doi.org/10.1126/science.aax2342.

Olsen, Henrik Palmer, Jacob Livingston Slosser, Thomas Troels Hildebrandt, and Cornelius Wiesener. 2019. "What's in the Box? The Legal Requirement of Explainability in Computationally Aided Decision-Making in Public Administration." https://core.ac.uk/download/pdf/322818589.pdf.

Pedreschi, Dino, Fosca Giannotti, Riccardo Guidotti, Anna Monreale, Salvatore Ruggieri, and Franco Turini. 2019. "Meaningful Explanations of Black Box AI Decision Systems." *Proceedings of the AAAI Conference on Artificial Intelligence* 33 (1): 9780–9784. https://doi.org/10.1609/aaai.v33i01.33019780.

Raaijmakers, Stephan. 2019. "Artificial Intelligence for Law Enforcement: Challenges and Opportunities." *IEEE Security & Privacy* 17 (5): 74–77. https://doi.org/10.1109/MSEC.2019.2925649.

Rao, Qing, and Jelena Frtunikj. 2018 "Deep Learning for Self-Driving Cars: Chances and Challenges." Proceedings of the 1st international workshop on software engineering for AI in autonomous systems SEFAIS: 35–38. https://doi.org/10.1145/3194085.3194087.

Ribeiro, Marco Tulio, Sameer Singh, and Carlos Guestrin. 2016. ""Why Should I Trust You?": Explaining the Predictions of Any Classifier." In *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining KDD*, 1135–1144. New York, NY: ACM. https://doi.org/10.1145/2939672.2939778.

Rong, Yao, Tobias Leemann, Thai-trang Nguyen, Lisa Fiedler, Tina Seidel, Gjergji Kasneci, and Enkelejda Kasneci. 2022. "Towards Human-centered Explainable AI: User Studies for Model Explanations." *CoRR abs/2210.11584*.

Sahoo, Abhaya Kumar, et al. 2019. "Intelligence-based Health Recommendation System Using big Data Analytics." In *Big Data Analytics for Intelligent Healthcare Management*, edited by Nilanjan Dey, Himansu Das, Bighnaraj Naik, and Himansu Sekhar Behera, 227–246. India: Academic Press.

Scanlon, Thomas M. 2000. *What We Owe to Each Other*. Cambridge, MA: Harvard University Press.

Selbst, A. and Julia Powles. 2017. "Meaningful Information and the Right to Explanation". *International Data Privacy Law 7*(4): 233–242. https://doi.org/10.1093/idpl/ipx022.

Setzu, Mattia, Riccardo Guidotti, Ann Monreale, Franco Turini, Dino Pedreschi, and Fosca Giannotti. 2021. "Glocalx - from Local to Global Explanations of Black Box AI Models". *Artificial Intelligence* 294: 1–20. https://doi.org/10.1016/j.artint.2021.103457.

Sharma Nisha and Mala Dutta. 2020. "Movie Recommendation Systems: A Brief Overview". In *Proceedings of the 8th International Conference on Computer and Communications Management (ICCCM '20). Association for Computing Machinery*, New York, NY, USA: 59–62. https://doi.org/10.1145/3411174.3411194.

Shi, S., R. Tse, W. Luo, Stefano D'Addona, and Giovanni Pau. 2022. "Machine Learning-Driven Credit Risk: A Systemic Review." *Neural Computing and Applications* 34 (17): 14327–14339. https://doi.org/10.1007/s00521-022-07472-2.

Sokol, Kacper, and Peter A. Flach. 2020. "Explainability Fact Sheets: A Framework for Systematic Assessment of Explainable Approaches." In *FAT**, 56–67. New York, NY, USA: Conference on Fairness, Accountability, and Transparency (FAT* '20). https://doi.org/10.1145/3351095.3372870.

Sovrano, Francesco, Salvatore Sapienza, Monica Palmirani, and Fabio Vitali. 2022. "Metrics, Explainability and the European AI Act Proposal." *J* 5 (1): 126–38. https://doi.org/10.3390/j5010010.

Speith, Timo. 2022. "A Review of Taxonomies of Explainable Artificial Intelligence (XAI) Methods." In *2022 ACM Conference on Fairness, Accountability, and Transparency, (FAccT '22)*, 2239–2250. New York, NY: ACM. https://doi.org/10.1145/3531146.3534639.

State, Laura. 2021. "Logic Programming for XAI: A Technical Perspective." In *ICLP Workshops. Vol. 2970. CEUR Workshop Proceedings. CEUR-WS.org*.

Turek, Matt. 2018. Explainable Artificial Intelligence (XAI)." *Defense Advanced Research Projects Agency*. https://www.darpa.mil/research/programs/explainable-artificial-intelligence

Vilone, Giulia and Luca Longo. 2021a. "A Quantitative Evaluation of Global, Rule-Based Explanations of Post-Hoc, Model Agnostic Methods." *Frontiers in Artificial Intelligence* 4: 1–20. https://doi.org/10.3389/frai.2021.717899.

Vilone, Giulia, and Luca Longo. 2021b. "Notions of Explainability and Evaluation Approaches for Explainable Artificial Intelligence." *Information Fusion* 76:89–106. https://doi.org/10.1016/j.inffus.2021.05.009.

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

Wachter, Sandra, Brent Mittelstadt, Chris Russell. 2018. "Counterfactual Explanations Without Opening the Black Box: Automated Decisions and the GDPR." *Harvard Journal of Law & Technology (Harvard JOLT)*, 31(2): 841–888.

Walke, Fabian, Lars Bennek, Till J. Winkler, and J. Till. 2023. "Artificial Intelligence Explainability Requirements of the AI Act and Metrics for Measuring Compliance." In *Wirtschaftsinformatik 2023 Proceedings*. Vol. 77. AISeL. https://aisel.aisnet.org/cgi/viewcontent.cgi?article=1076&context=wi2023.

Willson, Michele. 2019. "Algorithms (and the) Everyday." In *The Social Power of Algorithms*, edited by David Beer, 137–150. Routledge.

Zini, Julia El, and Mariette Awad. 2022. "On the Explainability of Natural Language Processing Deep Models." *ACM Computing Surveys* 55 (5): 1–31. https://doi.org/10.1145/3529755.