The Tower of Babel in Explainable Artificial Intelligence (XAI)

David Schneeberger[1](✉), Richard Röttger[4], Federico Cabitza[3], Andrea Campagner[3], Markus Plass[1], Heimo Müller[1], and Andreas Holzinger[1,2]

[1] Medical University of Graz, Graz, Austria
david.schneeberger@medunigraz.at
[2] University of Natural Resources and Life Sciences, Vienna, Austria
[3] University of Milano-Bicocca, Milan, Italy
[4] South Denmark University (SDU), Odense, Denmark

Abstract. As machine learning (ML) has emerged as the predominant technological paradigm for artificial intelligence (AI), complex black box models such as GPT-4 have gained widespread adoption. Concurrently, explainable AI (XAI) has risen in significance as a counterbalancing force. But the rapid expansion of this research domain has led to a proliferation of terminology and an array of diverse definitions, making it increasingly challenging to maintain coherence. This confusion of languages also stems from the plethora of different perspectives on XAI, e.g. ethics, law, standardization and computer science. This situation threatens to create a "tower of Babel" effect, whereby a multitude of languages impedes the establishment of a common (scientific) ground. In response, this paper first maps different vocabularies, used in ethics, law and standardization. It shows that despite a quest for standardized, uniform XAI definitions, there is still a confusion of languages. Drawing lessons from these viewpoints, it subsequently proposes a methodology for identifying a unified lexicon from a scientific standpoint. This could aid the scientific community in presenting a more unified front to better influence ongoing definition efforts in law and standardization, often without enough scientific representation, which will shape the nature of AI and XAI in the future.

Keywords: Artificial Intelligence · AI · Machine Learning · ML · Explainable AI · XAI · explainability · interpretability · transparency · ethics · law · GDPR · DSA · Artificial Intelligence Act · standardization · ISO · IEC · IEEE

1 Introduction and Motivation

With the (nearly) ubiquitous spread of complicated black box models like GPT-4, explainable AI (XAI) has gained importance in both science and industry as a counterbalancing force. XAI refers to the development of artificial intelligence

A. Holzinger et al. (Eds.): CD-MAKE 2023, LNCS 14065, pp. 65–81, 2023.
https://doi.org/10.1007/978-3-031-40837-3_5

(AI) systems that can provide clear, understandable, and interpretable explanations for their advice and decisions. The very definition of explanation, and of its mentioned desirable properties, is, however, often not straightforward from a scientific point of view, leaving intuitive understanding aside.

Indeed, with the expansion of this research area the definition of terms and the variety of definitions is growing so fast that it is becoming extremely difficult to follow. This confusion of languages also stems from the plethora of different perspectives on XAI, e.g. ethics, law, standardization and computer science. There is no community-based agreement about central terms like explanation, explainability or interpretability and, in the scientific domain, the context of these definitions is often not clear. We are therefore facing, as mentioned in the Introduction, the threat of a "tower of Babel" effect, i.e. a confusion of languages and terminologies which makes it hard to find common (scientific) ground.

To counter this linguistic ambiguity, this paper maps the perspectives of ethics guidelines, law and standardization and in these fields. In comparison to the scientific perspective, these fields are often driven by the quest for standardized, uniform definitions. It shows that despite this goal, there is still no common vocabulary in these fields. Subsequently, it proposes a method to focus the diverging perspectives in the XAI field in the search for a common "vocabulary", i.e. a unified lexicon from a scientific standpoint. Such a unified lexicon could aid the scientific community in presenting a more unified front to better influence ongoing definition efforts in law and standardization, which will shape the nature of AI and XAI in the future but are often marred by a lack of scientific participation and democratic legitimacy.

2 Ethics Guidelines and XAI

Law (e.g. the Artificial Intelligence Act, see Sect. 3.3) and standards are often informed by relevant documents and reports, i.e. soft law or ethics guidelines. For example, the OECD (Organisation for Economic Co-operation and Development) [46] defines the principle of transparency and explainability in the following way: AI actors "should provide meaningful information, appropriate to the context [...] to foster a general understanding of AI systems, to make stakeholders aware of their interactions with AI systems [...] to enable those affected by an AI system to understand the outcome, and, to enable those adversely affected by an AI system to challenge its outcome based on plain and easy-to-understand information on the factors, and the logic that served as the basis [...]".

This illustrates that terms like transparency and explainability are often used without drawing clear boundaries. Documents often refer to them as umbrella terms comprising several distinct elements, i.e. more general information (e.g. information on the interaction with an AI system), but also elements, which could necessitate the implementation of XAI approaches (e.g. "information on the factors and the logic that served as basis"). This muddled language makes it harder to derive clear implementation measures for XAI.

In contrast, the ethics guidelines of the high-level expert group on AI [28], set up by the European Commission, differentiate between several elements of transparency, which itself is linked with the principle of explicability, i.e. traceability (concerning the documentation of data sets, algorithms, and the processes that yield the decision), explainability (mainly concerning the ability to explain both the technical processes of an AI system and the related human decisions; information of the degree to which an AI system influences and shapes the organisational decision-making process, design choices of the system, and the rationale for deploying it) and communication (i.e. humans have the right to be informed that they are interacting with an AI system; capabilities and limitations should be communicated). Mainly the second element, explainability concerning the technical process, is linked with the implementation of XAI but again does not state concrete measures.

This problem of the use of vague umbrella terms, illustrated by the OECD example above, also exists on a macro level. As meta studies on ethics guidelines [39] show, "transparency" is the most often mentioned principle, but the interpretation, what transparency entails, varies widely in these guidelines, concerning what should be transparent (e.g. data use, human-AI-interaction, automated decisions, purpose of data use/application of the AI system) or the goal of transparency (e.g. minimize harm, improve AI, legal reasons, foster trust, principle of democracy). To achieve transparency, disclosure of information is often suggested but there is no agreement what should be disclosed (e.g. use of AI, source code, data use, evidence base, limitations, laws, responsibility for AI, investments, impact).

Ienca and Vayena [31] differentiate between two main thematic families of transparency mentioned in guidelines: Firstly, transparency of algorithms and data processing methods (which refers to the implementation of XAI approaches) and secondly transparency of human practices related to the design, development and deployment of AI systems (e.g. disclosing relevant information to data subjects, avoiding secrecy, forbidding conflicts of interest between AI actors and oversight bodies).

These divergent interpretations of transparency lead to divergences in the implementation strategies proposed to achieve transparency. Generally, a major problem lies in deducing concrete technological implementations from the very abstract ethical values and principles described in ethics guidelines [25].

As a brief mapping of these guidelines has illustrated, they seem to contribute to the "tower of Babel" effect concerning XAI terms as they often - which partly lies in the nature of ethics guidelines - only set out abstract principles without describing concrete implementation strategies.

3 Law and XAI

3.1 GDPR

Switching to the perspective of law and XAI, as AI specific regulation has only recently come into the focus of national and international legislators, the legal framework currently does not contain explicit legal definitions of "explainability" or “transparency”. This could change when the proposed Artificial Intelligence Act (AIA) comes into force (see Sect. 3.3).

Of course, at the EU and the national level there are (older) laws, which were not written with AI in mind, but which are also applicable to AI systems and contain transparency obligations (with further references [3,48]).

For example, the General Data Protection Regulation (GDPR) [21] has wide implications for the use of AI and it has become a model law for AI regulation. The processing of data in the context of (fully) automated individual decision-making, i.e. without (substantial) human involvement, is principally forbidden by Art. 22 GDPR - which has been in the center of the “right to an explanation” debate (with further references [6,40,45,49]) – but fully automated decision-making is allowed if one of three exceptions (necessary for entering into/performance of a contract, authorisation by EU/member state law, explicit consent) applies.

In such a case, specific information has to be proactively provided (Art. 13, 14) and the data subject also has a right to access this information on request (Art. 15). This includes information about the “existence of automated decision-making”, about “the logic involved” and “the significance and the envisaged consequences”.

The passage “the logic involved” has been interpreted in different ways, e.g. as a subject-specific local explanation of a specific decision [24,44,50] or as variant of a general (global) explanation (mainly concerning the features employed) [59]. Explaining the logic involved could therefore necessitate the implementation of a feature-importance based XAI approach.

A recent opinion (16 March 2023, C-634/21, ECLI:EU:C:2023:220) (with further references [47,54]) of the attorney general Pikamäe could clarify the interpretation. These opinions are often but not always adopted by the European Court of Justice. The opinion states that the “logic involved” does not necessitate the disclosure of the algorithm used. According to the opinion only “general information, in particular on the factors taken into account in the decision-making process and their weighting at an aggregated level”, i.e. a form of a global feature-importance explanation, has to be provided. But as the opinion also states that “sufficiently detailed explanations on the method used to calculate the score and on the reasons that led to a certain result” have to be provided, this seems contradictory as the wording “a certain result” seems to imply a local explanation. This contradiction will have to be clarified by the court of Justice but it seems more likely that “logic involved” will be interpreted as a more general (global) explanation, mainly based on aggregated features.

Recital 71 also mentions a right “to obtain an explanation of the decision reached after such assessment” as part of suitable measures to safeguard the data subject (Art. 22 para. 3) but this right is only mentioned in the recital. Recitals mainly function as guidelines on how to interpret law but can not create law themselves. Therefore, the existence and the content of a “right to (an) explanation” is still disputed in scholarship (e.g. [49,59]).

3.2 Digital Services Act (DSA)

The new Digital Services Act (DSA) [22], which for example comes into play if an information society service provider (e.g. a social network) uses AI to moderate content (for an overview, see [18,42]), also contains transparency provisions. Providers of intermediary services have to include information "on any policies, procedures, measures and tools used for the purpose of content moderation, including algorithmic decision-making" in their terms and conditions (Art. 14 para. 1 DSA). They are also subject to yearly transparency public reporting obligations on content moderation. This includes information on "any use made of automated means for the purpose of content moderation" (Art. 15 para. 1(e) DSA). These obligations do not seem to directly relate to the implementation of XAI methods, but they require transparency on an abstract, global level, i.e. a qualitative description and information about the purpose and performance metrics (i.e. accuracy and error rates) of these systems.

Online platforms displaying advertising must also ensure that the recipients of the service can identify meaningful information "about the main parameters used to determine the recipient to whom the advertisement is presented and, where applicable, about how to change those parameters" (Art. 26 para. 1(d) DSA). This requires a form of explanation on the main features used in displaying advertisements, i.e. a feature-importance explanation, which seems to have a local ("used to determine the recipient") and a counterfactual element ("how to change those parameters"). This obligation could therefore necessitate the implementation of a XAI approach, which provides this local feature-importance and counterfactual information.

3.3 The (Proposed) Artificial Intelligence Act (AIA)

In April 2021, the European Commission proposed the so-called Artificial Intelligence Act (AIA) [19]. Since then several amendments have been suggested by the EU co-legislators, the Council [11] and the European Parliament [20]. Even though there were some remaining issues (e.g. AI definition, regulation of general-purpose AI/foundational models like GPT-4) the European Parliament held a positive plenary vote on 14 June 2023 [60]. Therefore, the final phase of the law-making process, the so-called trilogue, has started.

The AIA (for a general introduction see [57]) follows a risk-based approach. AI systems with an "unacceptable risk" (Art. 5 AIA e.g. social scoring modelled on China) will be banned, while high-risk AI systems will be subjected to strict regulation and must undergo an ex-ante conformity assessment. Concerning systems which pose a limited risk, these are subject to specific transparency obligations (Art. 52 AIA, e.g. chatbots must identify themselves).

The AIA addresses two different forms of high-risk AI systems (Art. 6 AIA): First, AI systems that are products or a safety component of a product already covered by EU harmonisation legislation requiring a third-party conformity assessment (e.g. medical devices). Second, in Annex III AIA eight categories of stand-alone AI systems are listed which are also considered high-risk (e.g. migration, asylum and border control management).

The AIA contains a specific transparency obligation for high-risk AI systems. According to Art. 13 para. 1 AIA high-risk AI systems must be "be designed and developed in such a way to ensure that their operation is sufficiently transparent to enable users to interpret the system's output and use it appropriately". The appropriate type and degree of transparency seems to be relative, its goal is achieving compliance with (other) relevant obligations of the AIA (recital 47: "a certain degree of transparency").

Crucially, the AIA does not offer (legal) definitions (Art. 3 AIA) for the central terms "sufficiently transparent" or "to interpret". The AIA does not mention the concept of "explainability" and therefore does not differentiate between interpretability and explainability [17]. This lack of definitions could lead to legal uncertainty and the mentioned "tower of Babel" effect.

As has been stated in legal scholarship, this leaves the interpretation of Art. 13 para. 1 AIA and the level of transparency/interpretability required unclear [4,15]. Therefore, it has been argued that the question of how to make AI systems interpretable is left to the discretion of the AI system provider, i.e. the AI developer [17].

In conclusion, this leaves the interpretation, whether Art. 13 AIA necessitates the implementation of XAI techniques and which approach has to be chosen, e.g. if a local or global explanation is required, open. It can also be argued that only a general form of transparency, mainly through the provision of "instructions", which have to be proved according to Art. 13 para. 2 seq., will suffice to satisfy this requirement. These instructions must for example contain the purpose, the level of accuracy, robustness, circumstances, which may lead to risks, performance metrics regarding the use groups, specification for the input data or on training/validation/testing data.

For example according to [5] Art. 13 para. 1 AIA does not imply the necessity of explainability in the sense that the way in which data have been processed must be entirely traceable, but a more general form of transparency of the system's functioning and output generation. Furthermore, a study [52] on request of the European Commission stated that XAI techniques are not the "only means available to understand and interpret AI systems outputs" and therefore not required for all high-risk AI systems. Instead "documentation approaches, scenarios, principles of operations, as well as interactive training materials" will fulfill the requirements of Art. 13 AIA. This indicates that the implementation of XAI approaches is not a core component of this transparency obligation.

Several attempts to define the terminology used in Art. 13 AIA illustrate the struggle to find uniform definitions, which shape how XAI will be used in the future. For example, the Council [11] proposed to simplify this obligation, i.e. to use the term "understand" instead of "interpret", which in our opinion is equally vague and has no real benefits.

The second co-legislator, the European Parliament [20], also tries to fill this vague terminology with life. In the version of the Parliament, AI systems must be "sufficiently transparent to enable providers and users to reasonably understand the system's functioning." In our opinion the addition of "functioning" suggests a more general level of transparency, which also "shall be ensured in accordance with the intended purpose of the AI system", again indicating that the level of transparency is context sensitive. As a very important step in the direction of a precise terminology, the Parliament suggested to define "transparency", which shall "mean that, at the time the high-risk AI system is placed on the market, all technical means available in accordance with the generally acknowledged state of art are used to ensure that the AI system's output is interpretable by the provider and the user." As this refers to the state of the art, which is always in flux, this could mean that XAI approaches will become mandatory as they become state of the art and if they provide a clear benefit in helping the user interpret the output. On the other hand, the Parliament in our opinion seems to suggest a high-level, global form of transparency, based on a simplified understanding of the system and the features used ("The user shall be enabled to understand and use the AI system appropriately by generally knowing how the AI system works and what data it processes [...]"). This reduced obligation of "generally knowing" does not seem to necessitate the implementation of XAI techniques. This should in turn allow "the user to explain the decisions taken by the AI system to the affected person [...]". In our opinion this clarification is an important step in the right direction as it minimizes legal uncertainty regarding how "transparency" must be interpreted.

As a point of criticism, in the original AIA proposal the output must be interpretable only for the (professional) user (i.e. a doctor) and not the person who is affected by an AI system (i.e. a patient). But professional users are seldom the only ones put at risk by AI systems [7]. Therefore, Art. 13. AIA is sometimes referred to as a form of "user-empowering explainability" [53]. Critically, people who are affected by high-risk AI systems, are left without a new right to information [17]. This lack of a "human-centred approach" has been a major point of criticism [55].

To solve this oversight, the European Parliament [20] proposed the introduction of "A right to explanation of individual decision-making" (Art. 68c AIA). This would give "[a]ny affected person subject to a decision which is taken [...] on the basis of the output from an high-risk AI system" (e.g. a diagnosis by a doctor) "which produces legal effects or similarly significantly affects him or her" (e.g. it affects the health of a patient) a "right to request [...] clear and meaningful explanation [...] on the role of the AI system in the decision-making procedure, the main parameters of the decision taken and the related input data." In our opinion, this suggests a form of a local feature-importance explanation (main parameters of the decision, related input data), which could necessitate the implementation of XAI approaches, and additionally an explanation of the role of the AI system (e.g. diagnostic aid). This explanation must also be target appropriate (recital 84b "[...] they should take into account the level of expertise and knowledge of the average consumer or individual"). If this focus on the explanation of an individual decision is held up in the trilogue, this could necessitate the implementation of a XAI approach, which can produce a local feature-importance explanation.

Thematically linked, Art. 14 AIA on human oversight also requires the implementation of measures that enable the individuals, to whom human oversight is assigned, to "be able to correctly interpret the high-risk AI system's output". In this regard, "the characteristics of the system and the interpretation tools and methods available", i.e. the implementation of XAI techniques, have to be taken into account.

Even though the amendments by the European Parliament described above are a step in the right direction and could lead to a more precise terminology, there is still a high level of legal uncertainty in interpreting these transparency obligations. This leads to economic risk for AI providers, who have to interpret the provision themselves when assessing the conformity with the AIA. Of course, the jurisprudence of the European Court of Justice could lead to clarification, but this will only be on a case-to-case basis and will take years. Therefore, the third layer, standardization, could play an important role in defining these abstract concepts set out by law.

4 Standardization and XAI

As law, even AI-specific regulation, must be applicable to many different categories of automated/autonomous software systems, these instruments must be in a sense "technology-agnostic" as law can not be easily amended in lockstep with every novel technological development. Therefore, legal rules are by-design often written from an abstract perspective, i.e. they only set out high-level principles and goals like "security" or "transparency". The concrete technical implementation is often defined by standards, which are (often) developed by (private) organizations, so-called SDOs (Standards Development Organizations). To ensure a uniform level of AI safety, several SDOs are drafting AI standards to fill existing regulatory gaps.

At the international and EU level, the most important SDOs are the International Organization for Standardization (ISO), the International Electrotechnical Commission (IEC), the Institute of Electrical and Electronics Engineers (IEEE), the International Telecommunication Union (ITU), the Internet Engineering Task Force (IETF), the European Committee for Standardization (CEN), the European Committee for Electrotechnical Standardization (CENELEC) and the European Telecommunications Standards Institute (ETSI) [16].

In the upcoming part of the paper, we aim to give a brief overview of the standards concerning explainability/interpretability. As a caveat, most of these standards are still in development and as (most) of the drafts can not be publicly accessed, we do not aim to give an in-depth analysis.

ISO and IEC created the joint technical committee JTC 1/SC 42 which serves as "the focus and proponent [...] (for the) standardization program on Artificial Intelligence". Several working groups exist which are focused on different aspects (e.g. WG 1 foundational standards; WG 2 data; WG 3 trustworthiness) [37].

On the one hand, ISO/IEC AWI 12792, which is still in development, aims to create a transparency taxonomy describing "the semantics of the information elements and their relevance to the various objectives of different AI stakeholders" [34].

On the other hand, the technical specification ISO/IEC AWI TS 6254 "Objectives and approaches for explainability of ML models and AI systems", which is also still in a drafting state, "describes approaches and methods that can be used to achieve explainability objectives of stakeholders with regards to ML models and AI systems' behaviours, outputs, and results" [36].

It identifies characteristics of explainability (explanation needs, form, approaches, and technical constraints) and uses them to categorise existing approaches. As a limitation, according to a report [52], it does not discuss or compare the technological maturity and known limitations of the methodologies (i.e. if methods are trustworthy and reflect the actual decision-making process).

The ongoing discussions about these two standards illustrate the central aim and struggle of defining "transparency" and "explainability", which are the cornerstones of these standards [1]. Transparency was broadly defined as the "availability in relation to stakeholders of meaningful, faithful, comprehensive, accessible and understandable information about a relevant aspect of an AI system". XAI approaches could help in generating this necessary information. Interpretability concerning algorithms was defined as the "ease with which a stakeholder can comprehend in a timely manner the objective of an AI system, the reasons for the system's behavior, and whether it is working given its purpose and in line with stakeholder expectations, and how different inputs could lead to different outcomes". Interpretability can be reached through technical approaches like explainability methods or other analysis or visualization methods. Similar to the ethics guidelines of the high-level expert group on AI (see Sect. 2) two levels of explainability were differentiated. Explainability concerning policy as the "ability to provide stakeholders of an AI system with concise, accessible, sufficient and useful explanatory information beyond the AI system's results", which refers to the wider socio-economic context of an AI system, and explainability concerning algorithms as the "capability of an AI system to correctly produce the reasons for its own behavior in a timely manner, allowing scrutiny of whether it is working given its purpose and in line with stakeholder expectations, and how different inputs could lead to different outcomes", which refers to the implementation of XAI techniques.

Additionally, the terms explainability and/or interpretability are also mentioned in ISO/IEC 22989:2022 [33] on "Artificial intelligence concepts and terminology" and in ISO/IEC AWI TS 29119-11 [35] concerning testing of AI systems and in the ISTQB (International Software Testing Qualifications Board) syllabus [38] for "Certified Tester AI Testing" [13].

At the level of the IEEE, the P7000 series of standards is being developed as part of the Global Initiative on Ethics of Autonomous and Intelligent Systems. In contrast to more traditional standards, these standards aim to address "specific issues at the intersection of technological and ethical considerations" [30].

Regarding transparency, the already published standard IEEE P7001 [29] sets out transparency requirements without defining how to achieve them, i.e. which XAI techniques or solution to use. It (only) describes different levels of transparency with an increasing range of sophistication and complexity [52].

At the national level, the German SDOs DIN (Deutsches Institut für Normung) and DKE (Deutsche Kommission für Elektrotechnik Elektronik Informationstechnik) have released the second version of an extensive "Standardization Roadmap AI", which maps the existing standards and analyses the need for new AI standards [13]. As the roadmap states, there is a need to specify formal requirements for XAI methods (i.e. formulation of concrete operationalizable/testable requirements). It also states that additional basic research in XAI is required because available methods have not yet been fully and widely researched and applied. To fill these gaps, DIN is also working on a standard concerning explainability [12].

Besides these standards for explainability/interpretability, a whole range of standards for AI systems and related technologies is being developed at the national and international level (see [13,16]).

In comparison to the perspectives of ethics and law, the field of standardization illustrates even better the quest for a standardized, uniform terminology, which is still ongoing. But as the mapping above indicates, the contours of central terms are becoming sharper and sharper.

5 The Link Between Law and Standardization

Law and standardization are thematically interlinked. As a study regarding the AIA states: "Standards are set to bring the necessary level of technical detail into the essential requirements prescribed in the legal text, defining concrete processes, methods and techniques that AI providers can implement in order to comply with their legal obligations" [52]. Co-Regulation through standardization based on the new Legislative Framework (NLF) is a cornerstone of the AIA. The essential requirements contained in law are given concrete form by standards [16].

Instead of interpreting obligations like the transparency obligation Art. 13 AIA discussed in Sect. 3.3, which could take time and expertise and also lead to legal risk, AI providers can mitigate uncertainty and follow (harmonized) standards. This leads to the presumption, that an AI system conforms with the requirements of the AIA. Therefore, in practice (harmonised) standards will play an important role in shaping the technical requirements and therefore the XAI landscape.

These harmonised standards are developed on demand of the European Commission and are published in the official journal of the EU. At the EU level CEN, CENELEC and ETSI (see Sect. 4) function as the SDOs which can either transpose existing standards into European standards if they comply with European values, standards and legislation, or they can develop own standards. At the moment of writing the European Commission has already started the process to adopt a standardization request providing a formal mandate to European SDOs to develop the necessary standards [52].

Even though these standards could bring the necessary clarity to high-level obligations contained in ethics guidelines or the AIA by defining essential XAI terms, this heightened role of standardization has some disadvantages. Besides other general problems of standardizing AI (e.g. rapid change of the underlying technology, ongoing debate on ethical and legal questions [2]) there are numerous points for criticism: SDOs like the IEC and ISO typically work on a subscription model and retain copyright [56], creating a monetary barrier especially for small AI developers to access these standards. The standardization process is susceptible to lobbying [56] and large, global players could therefore try to influence the definition of central terms to shape the XAI landscape. Regulation by standards shifts the law-making power to private bodies, which, compared to national or EU legislation, lack in options for democratic control and participation [16,17,23,43,57]. This also reduces the possibility for the scientific community to influence the ongoing AI governance discussion.

The European Parliament has seemingly recognized his problem in their AIA amendments stating that it is necessary "to ensure a balanced representation of interests by involving all relevant stakeholders in the development of standards." (Recital 61) Therefore, the Commission must consult with the AI Office and the Advisory Forum (Art. 40 AIA, Recital 61a), which "should ensure varied and balanced stakeholder representation and should advise the AI Office" (Recital 76).

6 A Proposed Solution

As our analysis of ethics guidelines, law and standardization has shown, the quest for a precise terminology is still ongoing. In turn, XAI scientists cannot rely on the vague, partially contradictory, and overly numerous definitions. Furthermore, especially in standardization there is often very low participation of representatives of academy and scientific researchers. Methods of democratic representation are often lacking.

A first step to counter this development is to be aware of the definition problem and to create sensitivity about the opacity of the standards drafting mechanism. This position papers aims to contribute in building such an awareness in the scientific community.

As a second step, we then pose the opposite problem: how can scientists and XAI scholars inform the process of law-making and standardization so as to provide guidance for the conformity assessment that will be so crucial in evaluating the legality of the next AI systems disseminated to the general public or adopted in sensitive areas such as health care or public safety?

We therefore created a simple and feasible method so that, at least the community of scholars who are most interested in these issues, can converge in a lexicographic and definitional effort that brings order and gains the necessary visibility and credibility to inform standard and policy making.

In recent years, scientists active in the field of XAI have produced several reviews (e.g., [8,10,14,26,27,32,41,58]), both systematic and more narrative and exploratory ones, to understand the lexical and definition variety in the field and, in some ways, help reduce the linguistic babel, since this is seen as an obstacle for the diffusion and wide adoption of successful design patterns, and sound evaluation methods. Nonetheless, while all of these contributions primarily consist of taxonomies or similar hierarchical categorizations that attempt to represent, and somehow systematize, the above mentioned variety, we note that their aims (and, thus, the set of concepts and definitions they document and attempt to map out) differ. Indeed, while some of the referenced surveys [10,26,32] largely aimed at categorizing existing XAI techniques from the methodological point of view, with a consequent focus on notions related to presentation modality or explanation type; others have also considered a more user-oriented perspective, and thus focused on definitions and notions related to the evaluation, validation and effects of explanations [14,27]; or also to a more general investigation of the understanding of the notion of explanation itself [8,58]. Thus, it is easy to see that the above mentioned contributions can only be understood as a starting point for our proposed initiative, which is still far from being an exhausted topic.

What we are proposing, indeed, is to activate a truly communal initiative that can lead a set of representative scholars to 1) collect all the major definitions proposed in the highest impact articles or most comprehensive reviews 2) invite all the authors of these articles and registered participants at major conferences in the field (e.g. the International Conference on eXplainable Artificial Intelligence, the IJCAI Workshop on Explainable Artificial Intelligence, the Actionable Explainable AI Session at the Cross Domain Conference for Machine Learning and Knowledge Extraction, CD-MAKE) to vote about the precision, clarity and comprehensiveness of definitions of concepts such as explanation, explainability, transparency, causability, understandability on opportune ordinal scales, 3) to aggregate the results with state-of-the art methods, such as the one used in [9]; and 4) to return the results to the community, possibly iterating a few times so as to reduce variability and facilitate consensus building, in a manner not unlike a Delphi method involving the most motivated people in the field and mediated by asynchronous collaboration tools such as online questionnaires [51] and shared papers.

7 Conclusion

This paper mapped the ongoing efforts to define central XAI terms in ethics, law and standardization. It illustrates that the quest for a common vocabulary is still ongoing but there is the danger that the essential vocabulary and therefore the XAI landscape could be defined by efforts marred by a lack of scientific participation. After describing these challenges, the authors propose to start a consolidation process at the Cross Domain Conference for Machine Learning and Knowledge Extraction, CD-MAKE conference and systematically close the gap between scientific publications on one side and ethics guidelines, law and standards on the other side. A unified lexicon could aid the scientific community in presenting a more unified front to better influence ongoing definition efforts which will shape the nature of AI and XAI in the future. Instead, all areas should strengthen each other and learn from each other.

Acknowledgements. The authors declare that there are no conflict of interests. This work does not raise any ethical issues. This work has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No. 826078 (Feature Cloud). This publication reflects only the authors' view and the European Commission is not responsible for any use that may be made of the information it contains. Parts of this work have been funded by the Austrian Science Fund (FWF), Project: P-32554 (explainable Artificial Intelligence). This paper has been made open access CC-BY, freely accessible to the international research community. We are grateful for the valuable reviewer comments.

References

1. AI Standards Hub: Output from workshop on ISO/IEC standards for AI transparency and explainability. https://aistandardshub.org/forums/topic/output-from-workshop-on-iso-iec-standards-for-ai-transparency/-and-explainability/
2. Beining, L.: Vertrauenswürdige KI durch Standards? (2020). https://www.stiftung-nv.de/sites/default/files/herausforderungen-standardisierung-ki.pdf
3. Bibal, A., Lognoul, M., de Streel, A., Frénay, B.: Legal requirements on explainability in machine learning. Artif. Intell. Law **29**, 149–169 (2021). https://doi.org/10.1007/s10506-020-09270-4
4. Bomhard, D., Merkle, M.: Europäische KI-Verordnung. Recht Digit. **1**(6), 276–283 (2021)
5. Bordt, S., Finck, M., Raidl, E., von Luxburg, U.: Post-hoc explanations fail to achieve their purpose in adversarial contexts. In: FAccT 2022: 2022 ACM Conference on Fairness, Accountability, and Transparency, pp. 891–905. ACM, New York (2022). https://doi.org/10.1145/3531146.3533153
6. Brkan, M.: Do algorithms rule the world? Algorithmic decision-making and data protection in the framework of the GDPR and beyond. Int. J. Law Inf. Technol. **27**(2), 91–121 (2019). https://doi.org/10.1093/ijlit/eay017
7. Busuioc, M., Curtin, D., Almada, M.: Reclaiming transparency: contesting the logics of secrecy within the AI act. Eur. Law Open **2**, 1–27 (2022). https://doi.org/10.1017/elo.2022.47
8. Cabitza, F., et al.: Quod erat demonstrandum?-Towards a typology of the concept of explanation for the design of explainable AI. Expert Syst. Appl. **213**, 118888 (2023)
9. Cabitza, F., Ciucci, D., Locoro, A.: Exploiting collective knowledge with three-way decision theory: cases from the questionnaire-based research. Int. J. Approximate Reasoning **83**, 356–370 (2017)
10. Cambria, E., Malandri, L., Mercorio, F., Mezzanzanica, M., Nobani, N.: A survey on XAI and natural language explanations. Inf. Process. Manage. **60**(1), 103111 (2023)
11. Council: Proposal for a regulation of the European Parliament and of the Council laying down harmonised rules on Artificial Intelligence (Artificial Intelligence Act) and amending certain union legislative acts - general approach, 14954/22 (2022). https://www.kaizenner.eu/post/aiact-part3

12. DIN: SPEC 92001–3 Künstliche Intelligenz - Life Cycle Prozesse und Qualitätsanforderungen - Teil 3: Erklärbarkeit. https://www.din.de/de/forschung-und-innovation/din-spec/alle-geschaeftsplaene/wdc-beuth:din21:354291453
13. DIN, DKE: Normungsroadmap Künstliche Intelligenz: Version 2 (2022). https://www.dke.de/de/arbeitsfelder/core-safety/normungsroadmap-ki
14. Ding, W., Abdel-Basset, M., Hawash, H., Ali, A.M.: Explainability of artificial intelligence methods, applications and challenges: a comprehensive survey. Inf. Sci. (2022)
15. Ebers, M.: Standardisierung Künstlicher Intelligenz und KI-Verordnungsvorschlag. Recht Digit. **2**, 588–597 (2021)
16. Ebers, M.: Standardizing AI: the case of the european commission's proposal for an 'artificial intelligence act'. In: The Cambridge Handbook of Artificial Intelligence: Global Perspectives on Law and Ethics, pp. 321–344. Cambridge University Press, Cambridge (2022). https://doi.org/10.1017/9781009072168.030
17. Ebers, M., Hoch, V.R.S., Rosenkranz, F., Ruschemeier, H., Steinrötter, B.: The European Commission's proposal for an Artificial Intelligence Act- a critical assessment by members of the Robotics and AI Law Society (RAILS). J **4**(4), 589–603 (2021). https://doi.org/10.3390/j4040043
18. Eifert, M., Metzger, A., Schweitzer, H., Wagner, G.: Taming the giants: the DMA/DSA package. Common Mark. Law Rev. **58**(4), 987–1028 (2021). https://doi.org/10.54648/cola2021065
19. European Commission: Proposal for a Regulation of the European Parliament and of the Council laying down harmonised rules on Artificial Intelligence (Artificial Intelligence Act) and amending certain union legislative acts, COM(2021) 206 final. https://eur-lex.europa.eu/legal-content/EN/TXT/?uri=celex:52021PC0206
20. European Parliament: Amendments adopted by the european parliament on 14 June 2023 on the proposal for a regulation of the European Parliament and of the Council on laying down harmonised rules on Artificial Intelligence (Artificial Intelligence Act) and amending certain union legislative acts (COM(2021) 0206 - C9–0146/2021 - 2021/0106(COD))1. https://www.kaizenner.eu/post/aiact-part3
21. European Parliament, Council: Regulation (EU) 2016/679 of the European Parliament and of the Council of 27 April 2016 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), OJ L 2016/119, 1. https://eur-lex.europa.eu/eli/reg/2016/679/oj
22. European Parliament, Council: Regulation (EU) 2022/2065 of the European Parliament and of the Council of 19 October 2022 on a Single Market For Digital Services and amending Directive 2000/31/EC (Digital Services Act), OJ L 2022/277, 1. https://data.europa.eu/eli/reg/2022/2065/oj
23. Guijarro Santos, V.: Nicht besser als nichts: Ein Kommentar zum KI-Verordnungsentwurf. Zeitschrift Digitalisierung Recht **3**(1), 23–42 (2023)
24. Hacker, P., Passoth, J.H.: Varieties of AI explanations under the law. From the GDPR to the AIA, and beyond. In: Holzinger, A., Goebel, R., Fong, R., Moon, T., Müller, K.R., Samek, W. (eds.) xxAI 2020. LNCS, vol. 13200, pp. 343–373. Springer, Cham (2022). https://doi.org/10.1007/978-3-031-04083-2_17
25. Hagendorff, T.: The ethics of AI ethics: an evaluation of guidelines. Mind. Mach. **30**(1), 99–120 (2020). https://doi.org/10.1007/s11023-020-09517-8
26. Hanif, A., Zhang, X., Wood, S.: A survey on explainable artificial intelligence techniques and challenges. In: 2021 IEEE 25th International Enterprise Distributed Object Computing Workshop (EDOCW), pp. 81–89. IEEE (2021)

27. Haque, A.B., Islam, A.N., Mikalef, P.: Explainable artificial intelligence (XAI) from a user perspective: a synthesis of prior literature and problematizing avenues for future research. Technol. Forecast. Soc. Chang. **186**, 122120 (2023)
28. High-level Expert Group on AI: Ethics guidelines for trustworthy AI. https://op.europa.eu/en/publication-detail/-/publication/d3988569-0434-11ea-8c1f-01aa75ed71a1
29. IEEE: IEEE 7001–2021: IEEE standard for transparency of autonomous systems. https://standards.ieee.org/ieee/7001/6929/
30. IEEE: The IEEE global initiative on ethics of autonomous and intelligent systems. https://standards.ieee.org/industry-connections/ec/autonomous-systems/
31. Ienca, M., Vayena, E.: AI ethics guidelines: European and global perspectives. In: Towards Regulation of AI Systems, pp. 38–60. Council of Europe (2020). https://rm.coe.int/prems-107320-gbr-2018-compli-cahai-couv-texte-a4-bat-web/1680a0c17a
32. Islam, M.R., Ahmed, M.U., Barua, S., Begum, S.: A systematic review of explainable artificial intelligence in terms of different application domains and tasks. Appl. Sci. **12**(3), 1353 (2022)
33. ISO, IEC: 22989:2022: Information technology - artificial intelligence - artificial intelligence concepts and terminology. https://www.iso.org/standard/74296.html
34. ISO, IEC: AWI 12792: Information technology - artificial intelligence - transparency taxonomy of AI systems. https://www.iso.org/standard/84111.html
35. ISO, IEC: AWI TS 29119-11: Software and systems engineering - software testing - part 11: Testing of AI systems. https://www.iso.org/standard/84127.html
36. ISO, IEC: AWI TS 6254: Information technology - artificial intelligence - objectives and approaches for explainability of ML models and AI systems. https://www.iso.org/standard/82148.html
37. ISO, IEC: JTC 1/SC 42: Artificial intelligence. https://www.iso.org/committee/6794475.html
38. ISTQB: Certified tester AI testing (CT-AI). https://www.istqb.org/certifications/artificial-inteligence-tester
39. Jobin, A., Ienca, M., Vayena, E.: The global landscape of AI ethics guidelines. Nat. Mach. Intell. **1**(9), 389–399 (2019). https://doi.org/10.1038/s42256-019-0088-2
40. Kaminski, M.E.: The right to explanation, explained. Berkeley Technol. Law J. **34**, 189–218 (2019). https://doi.org/10.15779/Z38TD9N83H
41. Kargl, M., Plass, M., Müller, H.: A literature review on ethics for AI in biomedical research and biobanking. Yearb. Med. Inform. **31**(01), 152–160 (2022)
42. Knyrim, R., Urban, L.: DGA, DMA, DSA, DA, AI-Act, EHDS - ein Überblick über die europäische Datenstrategie (Teil I). Dako **3**, 55–58 (2023)
43. Laux, J., Wachter, S., Mittelstadt, B.: Three pathways for standardisation and ethical disclosure by default under the European Union Artificial Intelligence Act. Elsevier Preprint SSRN (2023). https://doi.org/10.2139/ssrn.4365079
44. Malgieri, G., Comandé, G.: Why a right to legibility of automated decision-making exists in the general data protection regulation. Int. Data Priv. Law **7**(4), 243–265 (2017). https://doi.org/10.1093/idpl/ipx019
45. Malgieri, G.: Automated decision-making and data protection in Europe. In: Research Handbook on Privacy and Data Protection Law, pp. 433–448. Edward Elgar, Cheltenham/Northampton (2022). https://doi.org/10.4337/9781786438515
46. OECD: Transparency and explainability. https://oecd.ai/en/dashboards/ai-principles/P7

47. Palmiotto Ettorre, F.: Is credit scoring an automated decision? The opinion of the AG Pikamäe in the case C-634/21 (2023). https://digi-con.org/is-credit-scoring-an-automated-decision-the-opinion-of-the-ag-//pikamae-in-the-case-c-634-21/
48. Schneeberger, D.: Der Einsatz von Machine Learning in der Verwaltung und die Rolle der Begründungspflicht. Ph.D. thesis, Graz (2023)
49. Schneeberger, D., Stöger, K., Holzinger, A.: The European legal framework for medical AI. In: Holzinger, A., Kieseberg, P., Tjoa, A.M., Weippl, E. (eds.) CD-MAKE 2020. LNCS, vol. 12279, pp. 209–226. Springer, Cham (2020). https://doi.org/10.1007/978-3-030-57321-8_12
50. Selbst, A.D., Powles, J.: Meaningful information and the right to explanation. Int. Data Priv. Law **7**(4), 233–242 (2017). https://doi.org/10.1093/idpl/ipx022
51. Shinners, L., Aggar, C., Grace, S., Smith, S.: Exploring healthcare professionals' perceptions of artificial intelligence: validating a questionnaire using the e-Delphi method. Digit. Health **7**, 20552076211003430 (2021)
52. Soler Garrido, J., et al.: AI watch: artificial intelligence standardisation landscape update (2023). https://publications.jrc.ec.europa.eu/repository/handle/JRC131155
53. Sovrano, F., Sapienza, S., Palmirani, M., Vitali, F.: Metrics, explainability and the European AI Act proposal. J **5**(1), 126–138 (2022). https://doi.org/10.1093/idpl/ipx022
54. Strassemeyer, L.: Externes Scoring kann, muss aber nicht unter Art. 22 Abs. 1 DSGVO fallen. Datenschutz-Berater (4), 102–106 (2023)
55. Van Kolfschooten, H.: EU regulation of artificial intelligence: challenges for patients' rights. Common Mark. Law Rev. **59**(1), 81–112 (2022). https://doi.org/10.54648/cola2022005
56. Veale, M., Matus, K., Robert, G.: AI and global governance: modalities, rationales, tensions. Annu. Rev. Law Soc. Sci. (2023). https://doi.org/10.31235/osf.io/ubxgk
57. Veale, M., Zuiderveen Borgesius, F.: Demystifying the draft EU artificial intelligence act. Comput. Law Rev. Int. **22**, 97–112 (2021). https://doi.org/10.9785/cri-2021-220402
58. Vilone, G., Longo, L.: Notions of explainability and evaluation approaches for explainable artificial intelligence. Inf. Fusion **76**, 89–106 (2021)
59. Wachter, S., Mittelstadt, B., Floridi, L.: Why a right to explanation of automated decision-making does not exist in the general data protection regulation. Int. Data Priv. Law **7**(2), 76–99 (2017). https://doi.org/10.1093/idpl/ipx005
60. Zenner, K.: Documents and timelines: the artificial intelligence act (part 3) (2023). https://www.kaizenner.eu/post/aiact-part3