XAI Desiderata for Trustworthy AI: Insights from the AI Act

Martin Krutský[1,*], Jiří Němeček[1], Jakub Peleška[1], Paula Gürtler[2] and Gustav Šír[1]

[1]Department of Computer Science, Czech Technical University, Karlovo náměstí 13, Prague, Czech Republic
[2]Center for Environmental and Technology Ethics, Czech Academy of Sciences, Celetná 988/38, Prague, Czech Republic

Abstract

Explainable AI (XAI) is an actively growing field. When choosing a suitable XAI method, one can get overwhelmed by the number of existing approaches, their properties, and taxonomies. In this paper, we approach the problem of navigating the XAI landscape from a practical perspective of emerging regulatory needs. Particularly, the recently approved AI Act gives users of AI applications classified as "high-risk" the *right to explanation*. We propose a practical framework to navigate between these high-risk domains and the diverse perspectives of different explainees' roles via six core XAI desiderata. The introduced desiderata can then be used by stakeholders with different backgrounds to make informed decisions about which explainability technique is more appropriate for their use case. By supporting context-sensitive assessment of explanation techniques, our framework contributes to the development of more trustworthy AI systems.

Keywords

explainable AI, AI governance, AI Act, trustworthy AI, XAI taxonomy

1. Introduction

Achieving transparency of models remains a significant challenge in the quest for trustworthy AI. While fully interpretable (understandable by a person) models are desirable for many reasons, as thoroughly argued by Rudin [1], there are incentives for choosing uninterpretable (black-box) models, mainly due to the empirically observed trade-off between performance and interpretability [2]. To obtain insight into the inner workings of a black-box model, one can turn to Explainable AI (XAI). XAI is a vast field of research, focused on explaining why a model reached some decision (so-called local explanation) or approximating the whole black-box with an interpretable model (i.e., a global explanation).

Obtaining such insight would be particularly useful in the area of AI safety, where contemporary efforts struggle with developing reliable, generalizable solutions for aligning AI systems to human values. Black-box deep learning methods exhibit critical failures like reward hacking [3], highlighting the infeasibility of some safe universal value function optimization. Widespread techniques based on such optimization over user data, such as Reinforcement Learning from Human Feedback (RLHF) [4], thus remain unreliable. Generally, as AI systems operate in increasingly complex environments, unforeseen failure modes are inevitable. Since it is impossible to predefine and test for all risks in advance, continuous human oversight is essential. We argue that *explainability* should serve as the foundation for such oversight, enabling ongoing assessment and intervention as new challenges emerge.

In addition, the development of cutting-edge AI models remains highly centralized, limiting external stakeholder influence. This further limits transparency, as stakeholders (e.g., deployers, users, or auditors) lack insight into how these models operate. Explainability is, therefore, crucial for *democratizing oversight*, enabling external actors to scrutinize, challenge, and, in turn, trust the AI decision-making [5].

TRUST-AI: The European Workshop on Trustworthy AI. Organized as part of the European Conference of Artificial Intelligence—ECAI 2025. October 2025, Bologna, Italy.

*Corresponding author.
martin.krutsky@cvut.cz (M. Krutský); nemecek.jiri@fel.cvut.cz (J. Němeček); jakub.peleska@cvut.cz (J. Peleška); guertler@flu.cas.cz (P. Gürtler); gustav.sir@cvut.cz (G. Šír)
https://martin-krutsky.github.io/ (M. Krutský); https://nemecekjiri.cz/ (J. Němeček); https://jakubpeleska.cz/ (J. Peleška); https://cetep.eu/paula-gurtler/ (P. Gürtler); https://gustiks.github.io/ (G. Šír)
0009-0000-9710-1147 (M. Krutský); 0009-0005-0585-1642 (J. Němeček); 0009-0000-8561-8106 (J. Peleška); 0000-0002-7604-3213 (P. Gürtler); 0000-0001-6964-4232 (G. Šír)

While the recent EU regulation [6] represents progress with the *right to explanation* for high-risk systems, it needs to be supported with suitable XAI methods for its implementation. Even though there is a broad body of technical explainability research, it mostly considers narrow metrics such as fidelity or attribution accuracy. These metrics fail to capture the broader ethical context, where explainability is not merely a technical concern but a socio-technical bridge between AI systems, governance structures, and users.

2. Selecting XAI methods

In light of the abundance of explainability (XAI) methods [7, 8, 9, 10], choosing the right one is not trivial. We argue that it should not depend only on the technical properties of the XAI method (e.g., faithfulness) or the AI model (e.g., differentiability, data modality) but also, crucially, on the specific goals of the explanation process. This entails social [11] and philosophical [12] aspects, but also the regulatory point of view [13, 14, 15, 16], as addressed in this paper. While there is plenty of research in each domain separately, their interaction, although essential, is rarely interrogated.

2.1. XAI Evaluation and Taxonomies

There are many taxonomies and surveys of XAI, as described in the survey of XAI surveys by Schwalbe and Finzel [10]. Existing XAI evaluation frameworks focus primarily on technical aspects [17], and in the Human-Centered XAI literature, user experience and performance are prioritized [18, 19, 20].

We focus on the evaluation of XAI methods with governance in mind. A crucial work by Nannini [15] argued for XAI from the legal perspective, stressing the diverse needs of different stakeholders. Panigutti et al. [14] then discussed limitations of XAI methods. Nauta et al. [21] presented 12 "Co-properties" of explanations which can serve as a comprehensive set of evaluation dimensions or desiderata for XAI methods. Importantly, Fresz et al. [16] extended the set with 5 more to 17 Co-properties, as motivated by legal practice. Our work can be considered an extension of these efforts. We distill the various existing dimensions to a more manageable set of 6 dimensions while establishing a link between the contemporary regulatory perspective (AI Act) and the diverse audiences of the explanations.

2.2. AI Act

As mentioned above, explanations for AI system outputs gain further relevance under the AI Act [6]—the first comprehensive regulation on safety related to AI systems, which has entered into force on 1 August 2024 in the European Union. Article 86 of this regulation establishes the *right to explanation* for AI systems legally defined as high-risk. The AI Act serves as product safety legislation, aiming to establish ex-ante rules for transparency and product safety that allow for the seamless functioning of the single market and prevent breaches of fundamental rights. In order to ensure an appropriate regulatory burden, the AI Act defines different levels of risk, which correspond to different obligations. First, it defines a number of unacceptable use cases, which are prohibited uses of AI. Second, it identifies eight high-risk areas which "pose a significant risk of harm to the health, safety or fundamental rights of natural persons" (Art 6(3)).[1] It is in this category of risk that most obligations for transparency, documentation, risk management, data governance, and the right to explanation apply. Further, there is a risk classification for general-purpose AI models in those with and those without Systemic Risk (Art 51(1)). Our proposed mapping requires knowing the application of a model, meaning that general-purpose AI could be considered only after it is put to a specific task.

While it is disputed whether the AI Act legally requires XAI methods [13, 14], there are strong reasons to strive for explainability from a regulatory perspective, nonetheless. For example, in the enforcement of the AI Act, compliance checks via auditors will be an important cornerstone. Such auditors need

[1]The 8 areas are roughly biometrics, education and vocational training, employment, access to essential services, law enforcement, migration, administration of justice, and critical infrastructure. The last-mentioned area is the only one exempt from the right to explanation.

Table 1
Definitions of the six desiderata.

Desideratum	An explanation should be...
Faithfulness	...faithful to the model's behavior.
Robustness	...resilient to small input perturbations, gene-ralize well.
Intuitiveness	...understandable to lay users.
Verifiability	...comparable to prior expert knowledge.
Actionability	...able to suggest actions that can be taken to change the model or its output.
Scalability	...practically computable for larger models.

to understand how systems work in order to assess risk levels. In the open question of AI liability, Ebers [13] demonstrates that "both contract and tort law can provide incentives to develop and use XAI systems" [13, p. 125] because it allows deployers of AI systems to point to the specific cause of harm and thus facilitate accountability. Consequently, identifying appropriate XAI methods under the AI Act, which can potentially inform technical standards and common practices among AI developers and deployers in the EU, is a worthwhile subject.

3. Proposed framework

We now describe our main contribution—the mapping from an explanation audience and the application area to a set of desiderata, which can guide the selection of the most appropriate XAI method. The entire framework, visualized in Figure 1, consists of two branches connected to a set of desiderata (dimensions on which to evaluate XAI methods). The left branch centers on the application, using the high-risk AI system categories as a guide, and the right one takes the point of view of the audience of the explanation. We first describe the choice of the six desiderata, then each branch, and finally discuss the choice of the appropriate XAI method based on this framework.

3.1. XAI desiderata

To make the framework visualizable and possible to navigate, we propose a unifying reduction of the many existing proposals for classifying properties (i.e., evaluation dimensions) of XAI methods. Starting from the previously proposed comprehensive set of 17 Co-properties [16] (an extension of the original 12 [21]), we narrow them into 6 main XAI desiderata, listed in Table 1.

Faithfulness First and foremost is faithfulness (also referred to as correctness [16, 21] or fidelity [22]). It is the ability of an XAI method to correctly represent the explained model behavior. In relation to the 17 Co-properties, in addition to the above-mentioned Correctness, we include Confidence and Completeness [21], because an explanation with lower confidence (or without information about confidence) can be considered less faithful to the model. Similarly, an explanation that only explains a part of the model's behavior can be considered unfaithful to the model regarding the remaining parts. Faithfulness is essential in all XAI uses. When an explanation is not faithful to the model, it can be misleading, even deceptive. Unfaithful explanation is useless at best, and possibly harmful.

Robustness Robustness is another essential technical property. Many XAI methods are unstable with respect to small input (or model) changes, or even to changing the random seed. Robustness includes Consistency (determinism and implementation invariance) and Continuity (generalization and continuity of the explanation) [21]. Indeed, an explanation that changes dramatically with a small change of the input or fails on out-of-distribution examples is non-robust and can be considered less desirable in some cases, e.g., in safety-critical systems.

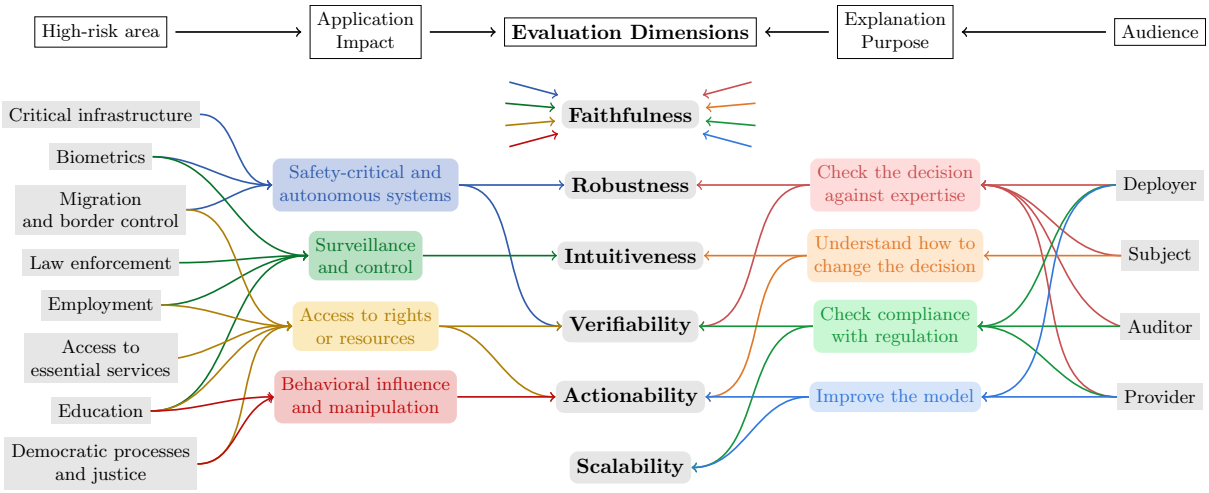

Figure 1: Proposed governance-oriented flowchart, identifying key evaluation dimensions for XAI methods. The structure reflects the AI Act's classification of high-risk applications (left side) and intended explanation audiences (right side). We shortcut arrows from each impact and purpose to faithfulness for visual clarity.

Intuitiveness Pulling away from the more technical properties, there is intuitiveness (sometimes referred to as interpretability [22]), the understandability of the explanation to a non-expert stakeholder. It entails Covariate complexity, Compactness, and Compositionality. Reduced complexity, including only necessary information, and good presentation, respectively, improve the intuitiveness of an explanation. Especially for lay users, intuitiveness is key to a better understanding of the model.

Verifiability A related property to intuitiveness, meaning that the explanation can be used to assess whether the model behaves according to prior expert knowledge. It loosely includes Coherence, i.e., comparability to and accordance with prior expert knowledge. Verifiability is a more technical property related to the presence of objective facts. As a result, an unintuitive explanation can still be verifiable.

Actionability Most closely related to Contrastivity (whether the explanation is discriminative w.r.t. other events or targets), actionability is the extent to which one can act on the explanation provided, i.e., to change the next prediction outcome. Controlability (or interactiveness) is also included in this dimension, since a controllable explanation provides the user with more fine-grained actions.

Scalability From the five extending Co-properties [16], we consider only Computations as relevant and include it in Scalability. It is the ability to provide an explanation in a practical time (Computations) and for larger models. This is an essential desideratum, heavily influencing the choice of an appropriate XAI method, because there is usually a trade-off between speed and quality of an explanation.

Other dimensions We have disregarded four Co-properties [16] for the following reasons. Context (relevance to the user needs) is encapsulated in the right side of our framework, where the audience's needs are specified. Consilience (whether more XAI methods should be used) and Counterability (whether a process for user objections should exist) are related more to the system as a whole, rather than the explanation method. Constancy (requirement of a storable explanation format) is irrelevant, since all computer-bound information can also be saved. Coverage (providing a local or a global explanation) represents an important distinction, but this is a property parallel to our framework.

3.2. Selecting most relevant desiderata

We now explain the mapping from the explanation audience and High-risk application areas to the six desiderata specified above. While all desiderata are relevant for each use case, some may be more

important in a given application or to a given explainee (e.g., auditor, subject). The full mapping is in Figure 1.

The mapping process can be explained using an example from Panigutti et al. [14]. Imagine a student proctoring AI system that uses a camera feed to estimate whether a student is cheating. This system clearly belongs to the Education area of High-risk systems, according to Annex 3 (3d) of the AI Act [6]. We then follow the arrows and evaluate the weight of the three relevant impact categories, where student proctoring is mainly about student surveillance. Following the arrow from surveillance to intuitiveness, we can say that intuitiveness should be prioritized when choosing an XAI method. Next, we follow the chart from the other side, taking the perspective of a stakeholder. Imagine we are a school considering the use of this proctoring tool. As a *deployer*, we select the most relevant of the three purposes and follow the respective arrow. In the early stages of the project, we want to check that the model behaves as expected, i.e., check the decisions of the model against our own domain expertise in spotting cheating students. For this consideration, we see that we should add *robustness* and *verifiability* to the list. This process leaves us with a subset of desiderata, which we can prioritize in the search for appropriate XAI methods.

3.2.1. Impact-oriented desiderata

There are many AI applications that belong to one of the eight high-risk areas according to the AI Act. We group them into four categories, based on their main impact, respectively. Note that we include Critical infrastructure for completeness, despite this area being exempt from the right to explanation.

Safety-critical and autonomous systems When human safety is in question, the system belongs to this category. All systems of critical infrastructure belong in this category, since they are critical for safety by definition. For the area of biometrics, one can imagine an access gate using biometric data. And from the area of border control, an example could be a fraudulent document detection.[2] Additional applications could be autonomous driving systems or medical diagnosis.

We link Safety-critical systems to *robustness* because when safety is in question, we must be able to get reliable explanations. We also link them to *verifiability* since, to implement such a system in practice, one would like the systems to allow for checking compliance with expert reasoning.

Surveillance and control Here, the main goal is to oversee a process or detect an event, while not being critical to safety. A facial analysis at a store checkout belongs to the area of biometrics and is used to surveil the checkout process. In the Law enforcement area, all applications belong to this category, from face recognition in a search for the suspect to predictive policing that controls which district to focus on. In employment opportunities, one might like to monitor worker efficiency, falling into the surveillance category. Similarly, in education, we can consider the student proctoring system outlined above. Another system belonging to this category could be emotion recognition tools.

The Surveillance and control category implies a focus on *intuitiveness*, as one would like to provide the surveilled subjects with human-understandable explanations of the decisions made about them.

Access to rights or resources This category contains all applications unrelated to safety where a decision is being reached. An example from the Migration could be an evaluation of a visa application. In employment, this could be a system filtering resumes. By definition, all systems utilized for access to essential services belong to this category. An automated grading system belongs here from the education area. And finally, in justice, one can take almost any system, e.g., a system summarizing evidence and law, which helps decide about access to the fundamental right of freedom. Systems evaluating one's access to a job, university, loan, or welfare all belong here.

[2]While it could be argued that this belongs to the access to rights category, the main impact of this application is indeed to detect threats to safety.

As exemplified, access to resources is another domain where domain expert knowledge plays a significant role and, thus, *verifiability* is an important property. Further, when rejecting access to someone, proposing *actionable* feedback is highly desired.

Behavioral influence and manipulation This final category comprises all uses that are primarily focused on influencing or manipulating human behavior. An example from the area of education could be a system curating the learning path of a student. In the democratic process, one might use an AI-based curator of an online campaign for elections. Most uses of chatbots, recommender systems, and other personalized systems would also fall into this domain.

When faced with such a system, the most important desideratum is *actionability*. When human behavior is being influenced, one might be naturally interested in how to steer the model or avoid some influences altogether.

3.2.2. Purpose-oriented desiderata

We now turn to the audience perspective on the right side of the Figure 1. There, we consider the point of view of four different explainees, based on their role in the system, similarly to Langer et al. [23]. This involves the provider, i.e., the entity that develops an AI model. Then comes an auditor who evaluates the model for compliance. A deployer decides to use the model and sets it up. Finally, a subject is then the person subjected to some outcome of the model. Each role has one main explanation purpose, positioned on the left of it respectively (but can also have multiple different ones, as per the arrows).

Check the decision against expertise The main concern that the deployer might have before using a model is to check the decision against what they know to be true in their domain of expertise. We assume that every role can have some prior knowledge that should align with the explanation of the decision (e.g., a doctor's experience with diagnosis).

To validate a decision, the two main desiderata to consider are *robustness*, as expert decision-making is often stable under minor (e.g., human-imperceptible) perturbations, and *verifiability*, since a proper format and information content are necessary for rigorous explanation validation.

Understand how to change the decision Only when assuming the role of the subject of some decision does it make sense to consider how one might change it. This contrastivity is known to be desirable to lay users [11].

Clearly, it is important that such an explanation is *intuitive*. Otherwise, a layperson might struggle to understand it, making it irrelevant. At the same time, one would like it to be *actionable*, suggesting some actions that could be taken to change the decision.

Check compliance with regulation When an auditor (or a provider, or a deployer) examines the model to evaluate compliance, it brings a substantially different perspective. Here, we assume that the regulation does not merely suggest that there should be some explanation, but rather that the model should behave a certain way, and the explanations are used as a means of evaluation.

In this scenario, we require *verifiability*, which is essential in compliance checking, enabling evaluation by an expert auditor. The model behavior will have to be examined for models of all sizes, implying the link to *scalability*.

Improve the model Finally, we consider the main scenario of the provider. Explanations are being used not only to validate the AI model, but also to improve it. This is of interest only to the provider and possibly the deployer, if they have access to the model itself.

To improve a model, one would again like an *actionable* explanation, i.e., one that suggests how the model should change to modify its behavior appropriately. Finally, one would like to improve models of all sizes, rendering the *scalability* desideratum important as well.

3.3. Selecting the right XAI method

Clearly, there is more to the selection of the right XAI method than our usage-oriented desiderata. What remains are technical constraints, such as data modality or the choice of the prediction model. There are many existing taxonomies of XAI methods based on such technical properties [10]. The intended use of our framework is to take the subset of desiderata, and check which XAI method complies with the most of them, while being usable under the technical constraints imposed by the used implementation. Alternatively, one can first perform the mapping and then influence the choice of the AI model to better enable the use of some XAI method with more desirable properties. Finally, this framework can be used to find blind spots in XAI research—a set of desiderata with no suitable existing XAI method.

In our earlier student proctoring example, we found *faithfulness*, *robustness*, *verifiability*, and *intuitiveness* as the most desirable. Now, if we were using an end-to-end Convolutional Neural Network, we might be constrained to use saliency maps [24] for explanations, which are interpretable, verifiable, and can be faithful. We would then prioritize selecting the most robust method faithful to the model. Alternatively, if we were deciding what AI model to use, we might see that counterfactual explanations [25] are faithful, verifiable, and highly interpretable; we could choose a robust method for the counterfactual generation and develop an AI model that would allow the method to be used. For example, a 2-phase model first extracting facial features and then using them for the cheater detection (as described by Panigutti et al. [14]), enabling the natural use of counterfactual explanations in the second step.

4. Conclusion

We have proposed a framework for mapping from high-risk uses of AI and the roles of different stakeholders, as specified by the AI Act, to six core XAI desiderata. With a running example, we showed that the framework can be used by non-experts to make a more informed decision when choosing appropriate XAI tools from a given subset that satisfies external technical constraints. The proposed framework could be used even for applications outside the high-risk category, as long as one is able to map them to one of the four impact categories.

Acknowledgment

This work has received funding from the European Union's Horizon Europe Research and Innovation program under the grant agreement TUPLES No. 101070149.

Declaration on Generative AI

In this work, the authors used generative AI tools for grammar and spell checks. After using these tools, the authors revised the content as needed and take full responsibility for the publication's content.

References

[1] C. Rudin, Stop explaining black box machine learning models for high stakes decisions and use interpretable models instead, Nature Machine Intelligence 1 (2019) 206–215.

[2] A. Assis, J. Dantas, E. Andrade, The performance-interpretability trade-off: a comparative study of machine learning models, Journal of Reliable Intelligent Environments 11 (2024) 1.

[3] L. L. Di Langosco, J. Koch, L. D. Sharkey, J. Pfau, D. Krueger, Goal misgeneralization in deep reinforcement learning, in: International Conference on Machine Learning, PMLR, 2022, pp. 12004–12019.

[4] S. Griffith, K. Subramanian, J. Scholz, C. L. Isbell, A. L. Thomaz, Policy shaping: Integrating human feedback with reinforcement learning, Advances in neural information processing systems 26 (2013).

[5] N. Aoki, The importance of the assurance that "humans are still in the decision loop" for public trust in artificial intelligence: Evidence from an online experiment, Computers in Human Behavior 114 (2021) 106572.
[6] Regulation (EU) 2024/1689, Regulation (EU) 2024/1689 of the European Parliament and of the Council, 2024.
[7] A. Adadi, M. Berrada, Peeking inside the black-box: a survey on explainable artificial intelligence (xai), IEEE access 6 (2018) 52138–52160.
[8] F. K. Došilović, M. Brčić, N. Hlupić, Explainable artificial intelligence: A survey, in: 2018 41st International convention on information and communication technology, electronics and microelectronics (MIPRO), IEEE, 2018, pp. 0210–0215.
[9] R. Dwivedi, D. Dave, H. Naik, S. Singhal, R. Omer, P. Patel, B. Qian, Z. Wen, T. Shah, G. Morgan, et al., Explainable ai (xai): Core ideas, techniques, and solutions, ACM Computing Surveys 55 (2023) 1–33.
[10] G. Schwalbe, B. Finzel, A comprehensive taxonomy for explainable artificial intelligence: a systematic survey of surveys on methods and concepts, Data Mining and Knowledge Discovery 38 (2024) 3043–3101.
[11] T. Miller, Explanation in artificial intelligence: Insights from the social sciences, Artificial Intelligence 267 (2019) 1–38.
[12] J. Lu, D. Lee, T. W. Kim, D. Danks, Good Explanation for Algorithmic Transparency, 2019.
[13] M. Ebers, Explainable ai in the european union: an overview of the current legal framework (s), Nordic yearbook of law and informatics 2020–2021: law in the era of artificial intelligence (2022).
[14] C. Panigutti, R. Hamon, I. Hupont, D. Fernandez Llorca, D. Fano Yela, H. Junklewitz, S. Scalzo, G. Mazzini, I. Sanchez, J. Soler Garrido, et al., The role of explainable ai in the context of the ai act, in: Proceedings of the 2023 ACM conference on fairness, accountability, and transparency, 2023.
[15] L. Nannini, Habemus a Right to an Explanation: so What? – A Framework on Transparency-Explainability Functionality and Tensions in the EU AI Act, Proceedings of the AAAI/ACM Conference on AI, Ethics, and Society 7 (2024) 1023–1035.
[16] B. Fresz, E. Dubovitskaya, D. Brajovic, M. F. Huber, C. Horz, How Should AI Decisions Be Explained? Requirements for Explanations from the Perspective of European Law, Proceedings of the AAAI/ACM Conference on AI, Ethics, and Society 7 (2024) 438–450.
[17] S. Mohseni, N. Zarei, E. D. Ragan, A multidisciplinary survey and framework for design and evaluation of explainable ai systems, ACM Transactions on Interactive Intelligent Systems (TiiS) 11 (2021) 1–45.
[18] Q. V. Liao, K. R. Varshney, Human-centered explainable ai (xai): From algorithms to user experiences, 2021. arXiv:2110.10790.
[19] P. Lopes, E. Silva, C. Braga, T. Oliveira, L. Rosado, Xai systems evaluation: a review of human and computer-centred methods, Applied Sciences 12 (2022) 9423.
[20] J. Kim, H. Maathuis, D. Sent, Human-centered evaluation of explainable ai applications: a systematic review, Frontiers in Artificial Intelligence 7 (2024) 1456486.
[21] M. Nauta, J. Trienes, S. Pathak, E. Nguyen, M. Peters, Y. Schmitt, J. Schlötterer, M. Van Keulen, C. Seifert, From anecdotal evidence to quantitative evaluation methods: A systematic review on evaluating explainable ai, ACM Computing Surveys 55 (2023) 1–42.
[22] M. A. Islam, M. F. Mridha, M. A. Jahin, N. Dey, A unified framework for evaluating the effectiveness and enhancing the transparency of explainable ai methods in real-world applications, 2024. arXiv:2412.03884.
[23] M. Langer, D. Oster, T. Speith, H. Hermanns, L. Kästner, E. Schmidt, A. Sesing, K. Baum, What do we want from Explainable Artificial Intelligence (XAI)? – A stakeholder perspective on XAI and a conceptual model guiding interdisciplinary XAI research, Artificial Intelligence 296 (2021) 103473.
[24] J. Adebayo, J. Gilmer, M. Muelly, I. Goodfellow, M. Hardt, B. Kim, Sanity Checks for Saliency Maps, in: Advances in Neural Information Processing Systems, volume 31, Curran Associates, Inc., 2018.
[25] S. Wachter, B. Mittelstadt, C. Russell, Counterfactual Explanations Without Opening the Black Box: Automated Decisions and the GDPR, SSRN Electronic Journal (2017).