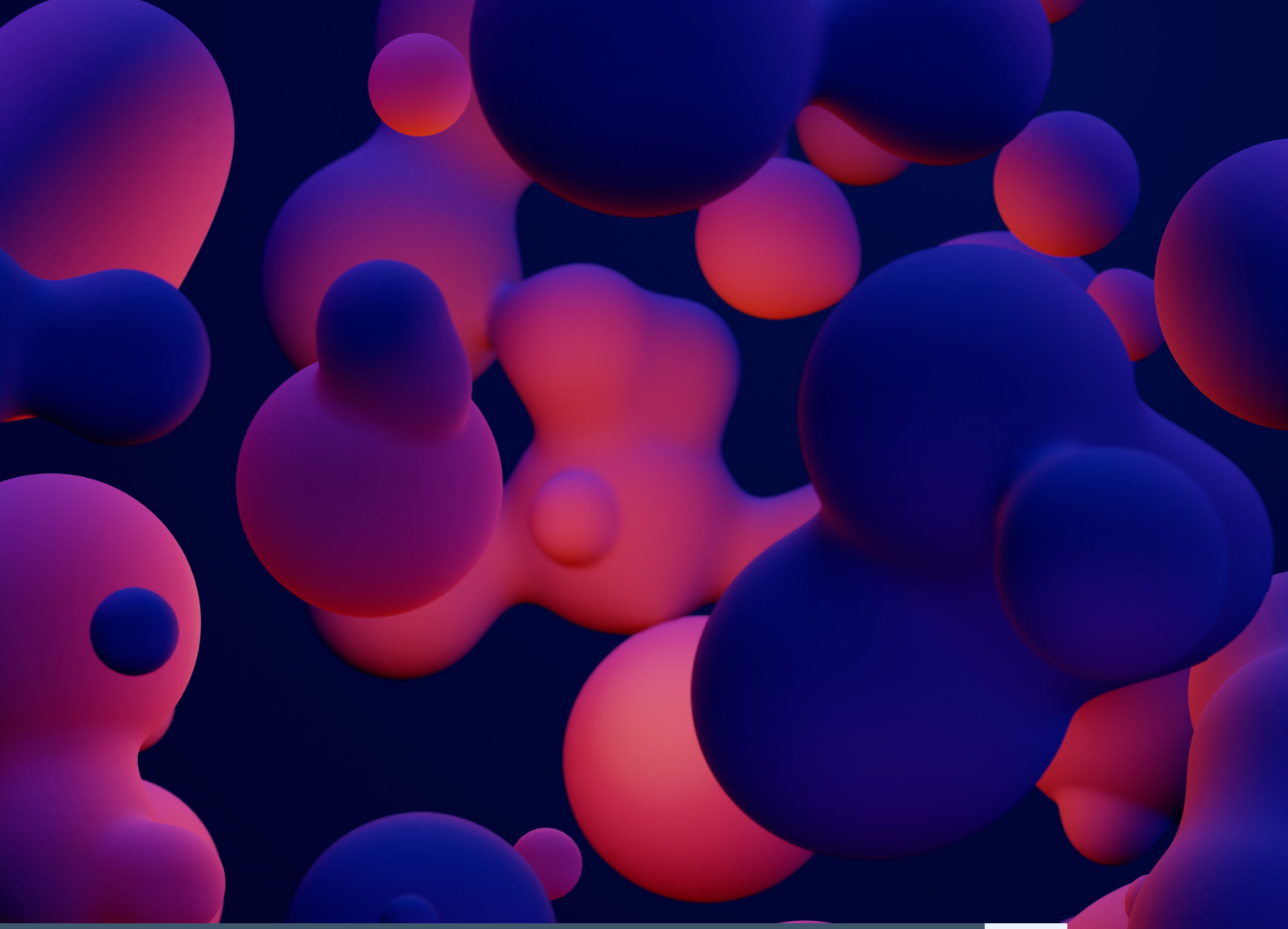


Transparency and explainability in AI

WHITE PAPER

Trustworthy artificial intelligence

3 October 2023

Peter Müllner*, Emanuel Lacic*, Simone Kopeinik*, Sebastian Scher*,
Tomislav Nad**, and Dominik Kowald*

* Know-Center GmbH

** SGS Digital Trusts Services GmbH

Partners of SGS

KNOW Center

SGS

Transparency and explainability in AI

Contents:

Introduction

In recent years, AI-based systems have been used for a broad body of tasks. These tasks rely on AI-made decisions which often are of a highly sensitive nature, possibly violating the well-being of individuals or social groups, e.g. the approval of loans, hiring processes, or health-related decisions. To oversee and appropriately use AI not only, but particularly, in such sensitive tasks and decisions, stakeholders require AI-made decisions to be understandable and reasonable for human beings. Therefore, explainability and transparency are essential for AI systems to be trustworthy.

Transparency can be defined as the understandability of a specific AI system – i.e. how well we know what happens in which part. This can be a mechanism that facilitates accountability (Lepri et al. 2018). Explainability is a closely related concept (Lepri et al. 2018, Larsson and Heintz, 2020) and provides information in a reverse manner on the logic, process, factors, or reasoning upon which the AI system's actions are based. Explainable AI (XAI) can be achieved via various means, for example, adapting existing AI systems or developing AI systems that are explainable by design. Commonly, these methods are referred to as "XAI methods".

According to Meske et al. (2022), transparency and explainability in AI pertain to five stakeholder groups:

1. **AI regulators,** which need explanations to test and certify the system
2. **AI managers,** who need explanations to supervise and control the algorithm and its usage, and to ensure the algorithm's compliance
3. **AI developers,** who use explanations to improve the algorithm's performance as well as for debugging and verification. This helps to pursue a structured engineering approach based on cause analysis instead of trial and error
4. **AI users,** who are interested in understanding and comparing the reasoning of the algorithm with his or her own way of thinking, to assess validity and reliability
5. **Individuals affected by AI decisions,** who are interested in explainability to evaluate the fairness of a given AI-based decision

Motivated by the importance of the explainability of AI systems for many sensitive real-world tasks, this white paper (i) provides a high-level overview of the taxonomy of XAI methods, (ii) reviews existing XAI methods, and (iii) thoroughly discusses possible challenges and future directions.

Definition of terms

To help readers grasp the different families and kinds of XAI methods, a high-level categorization of these methods is briefly presented (Ding and Abdel-Basset et al. 2022). Overall, there is an abundance of XAI methods that differ in multiple things, for example:

- The task at hand which is solved by the underlying AI system (e.g. classification) (Ribeiro et al. 2016)
- Whether XAI methods are incorporated directly into the AI system or are applied on top of the AI system (i.e. to the AI-made decisions). (Holzinger et al. 2019, Weidele et al. 2020)
- The data on which the XAI methods can be applied to, e.g. image data (Chattopadhay et al. 2018), textual data (Ribeiro et al. 2016), or graph data (Pope et al. 2019)
- What kind of explanations are generated, e.g. text-based explanations

These are examples of the various things in which XAI methods can differ. However, which XAI method should be applied strongly depends on the stakeholder group and on the application and domain in which the AI system is used.

Transparency can be defined as the understandability of a specific **AI** system.

Relevance and potential risks

Transparency in AI systems, especially in sensitive domains such as healthcare, finance, or human resources, can be achieved by applying XAI methods. This helps to better understand the system's decisions (Ding and Abdel-Basset et al. 2022). For example:

- **Healthcare,** explainable AI systems are used in the following: identifying prostate cancer, forecasting the effects of pneumonia treatment, forecasting deaths in hospitals, classifying autism spectrum disorders, or explaining the survival rate of breast cancer patients. In all these cases, explanations help patients and clinicians to understand how the AI system came to its conclusion, and thus generate trust in the AI system
- **Finance,** applications of explainable AI systems include predicting corporate financial distress, counterfeit detection of banknotes, predicting and explaining unusual and suspicious employee expenses, and mortgage lending and credit approval
- **Human resources,** explainable AI methods explain why a candidate was selected by the AI system. With this, recruiters can make informed decisions about which people to employ

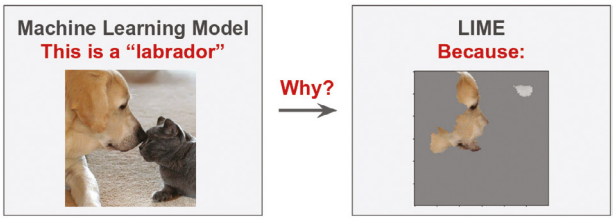
For non-sensitive domains, XAI methods also have clear benefits. For example, if an AI system classifies fabricated parts as faulty through a visual inspection, explanations can inform human inspectors why this part was classified as faulty. However, using XAI methods also generates specific potential additional risks that do not occur when using "normal" unexplainable AI methods. For example, explanations could leak crucial information about the AI system, which is a serious risk, especially in sensitive domains. Also, XAI methods may generate explanations that are not comprehensive and understandable for humans and therefore can be misunderstood. This impacts human decision-makers and harms the well-being of individuals, e.g. by denying a bank loan. Furthermore, it should be noted that XAI methods should increase the trust in AI systems, as explanations help stakeholders better understand decisions. However, without certifications and regulations, it is not guaranteed that an XAI method generates reliable and truthful explanations.

2 Methods for addressing existing issues

In this section, an overview of how the explainability of AI systems can be ensured is provided. Additionally, ways to measure and evaluate explainability are reviewed.

Examples of XAI methods

XAI methods generate explanations for AI systems which humans can understand. For example, Local Interpretable Model-agnostic Explanations (LIME) (Ribeiro et al. 2016) generates explanations for image classification tasks. Intuitively, it tries to understand why an AI system predicts that an instance belongs to a certain class. For example, the given picture is classified as a Labrador due to the snout and eyes of the dog (Russakovsky, 2015. Image from Arteaga, C.[1]).



Besides model-agnostic LIME, an example of a model-specific XAI method is provided, i.e. for Generative Adversarial Networks (GANs) (Bau et al. 2019, Bau et al. 2020). For this type of AI system, the inside of the system must be observed to generate explanations. Specifically, the internal representation of the GAN is matched with semantic concepts to obtain explanations for the AI system's decisions. In the example below (image from Bau et al.[2]), the AI system identifies the concepts of sky, tree, church, grass and road.

Many tools and frameworks for explainable AI exist. For example, AIX360, Skater, TF-explain, EthicalML-XAI, Captum, DALEX, Alibi and InterpretML. These include a variety of XAI methods, similar to the examples explained above.

Metrics and methods for evaluating explainability

To this date, no generally accepted metric exists for evaluating the quality of generated explanations. This is particularly because of the absence of ground truth data. In general, XAI evaluation can be of the following types (Adadi and Berrada, 2018, Dingand Abdel-Basset et al. 2022):

- **Application-grounded,** which uses humans to evaluate every result of the XAI method within real applications and decides if the explainable AI system works well or not
- **Human-grounded,** which is similar to the application-grounded approach, but which uses non-experts to evaluate tasks, rather than domain experts
- **Function-grounded,** where humans are not involved in the evaluation of XAI methods. The goal is to utilize structured proxy concerns to access explainability

These types of XAI evaluation lead to different evaluation dimensions that must be considered when comparing XAI methods (Hedström et al. 2023):

- **Faithfulness,** quantifies how precisely the explanations follow the predictive behavior of the AI system. Specifically, important features play a larger role in the system's outcomes
- **Robustness,** measures to what extent explanations are stable when subject to slight perturbations of the input, assuming that the system's output approximately stayed the same
- **Localization,** tests if the explainable evidence is centered around a region of interest (RoI), which may be defined around an object by a bounding box, a segmentation mask or a cell within a grid
- **Complexity,** captures to what extent explanations are concise, i.e. that few features are used to explain an AI system's prediction
- **Randomization,** tests to what extent explanations deteriorate as inputs to the evaluation problem, e.g. the AI system's parameters are increasingly randomized
- **Axiomatic,** means that explanations fulfil certain axiomatic properties

[1] *https://colab.research.google.com/github/flecue/xai-aaai2022/blob/main/XAI_LIME_Image.ipynb* (Last accessed: June 14, 2023).

[2] *https://colab.research.google.com/gist/flecue/2d7ade58aa3292733974a72df5363362/gandissect_exercise.ipynb* (Last accessed: June 14, 2023).


XAI methods promote
trust in AI systems,
as explanations
help stakeholders
to understand
decisions better

Open issues and challenges

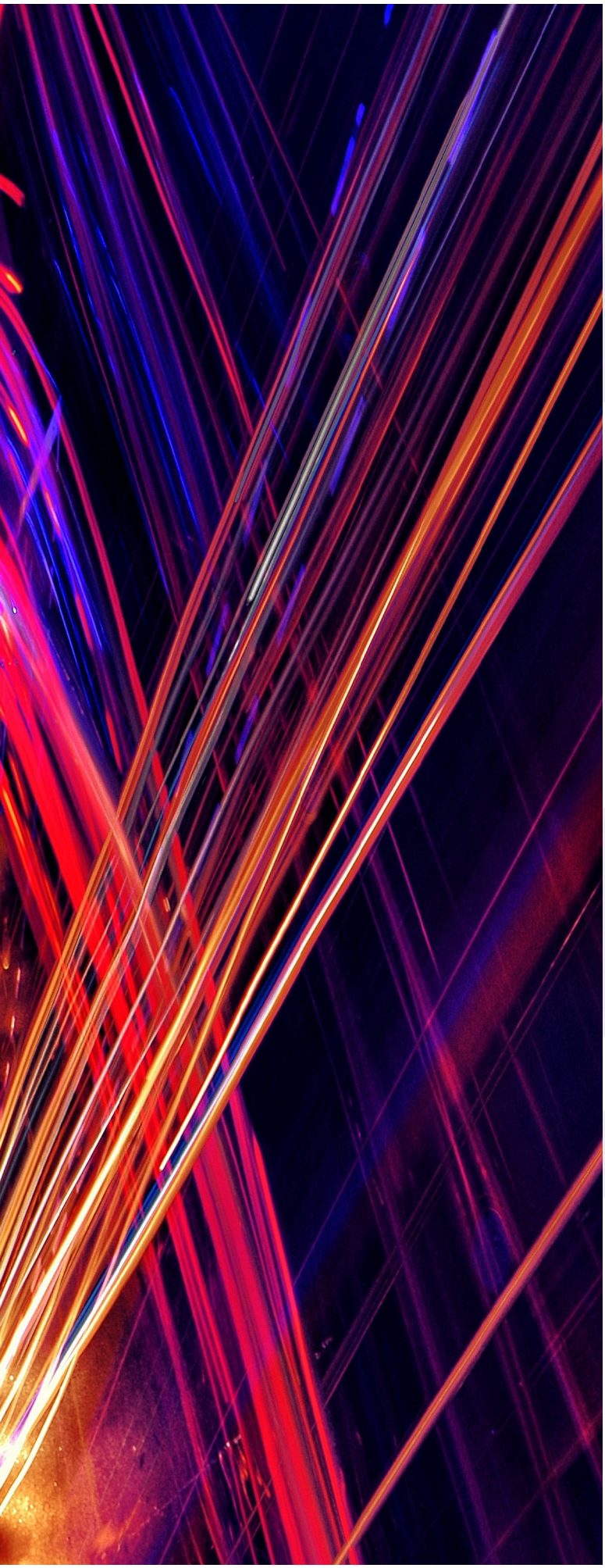
Overall, XAI methods are already applied in many real-world applications and are critical tools for sensitive domains. However, there are many aspects of explainable AI systems that – to this day – remain understudied and not well understood:

- **Trade-off between the secrecy of the AI system and its explainability,** explaining details of an AI system can leak too many details about the architecture of the system and, in this way, compromise its secrecy. This is especially problematic in corporate environments when AI systems are monetized, and when disclosing the architecture to third parties is not tolerated
- **Quantification of the comprehensibility of explanations for humans,** the explanations may be understandable for one stakeholder (e.g. a technical expert), but not for the other stakeholder (e.g. end-users of the system). Therefore, measuring how easy it is for humans to understand a given explanation is important
- **Regulation and certification of AI systems,** stakeholders need to be sure that the AI system generates reliable, trustworthy, and true predictions and decisions, especially in sensitive domains. For this, it is crucial that AI systems (and XAI methods) are regulated and certified appropriately

Also, using an XAI method is only one part of incorporating XAI into an AI system. Incorporating XAI into a fully functional user interface (UI) with the right user experience is crucial as it also has a final impact on the user. As such, while designing an explainable UI for an explainable AI system, the following principles should be considered (Chromik and Butz, 2021):

- **Complementary naturalness,** which complements implicit explanations with rationales in natural language
- **Responsiveness through progressive disclosure,** which offers hierarchical or iterative functionalities that allow follow-ups to initial explanations
- **Flexibility through multiple ways to explain,** which offers multiple explanation methods and modalities to enable insights to be triangulated
- **Sensitivity to the mind and context,** which offers functionalities to adjust explanations to the stakeholders' mental models and contexts

Overall, many facets of explainable AI systems are poorly researched and remain open challenges. Therefore, these are key topics for researching future explainable AI systems.



Summary

Today, AI systems are often used for very sensitive tasks, and can possibly violate the well-being of individuals or social groups. Therefore, stakeholders require AI-made decisions to be understandable and reasonable for human beings. This can be achieved with XAI methods, which are tools that generate explanations of the AI system's decisions. There is an increasing number of different XAI methods that can be applied to a variety of tasks. However, they heavily rely on the underlying data structure, the utilized AI system, the target stakeholders, and the application domain. All this needs to be considered when choosing an XAI method for a specific task. Additionally, more research is needed in the direction of evaluation metrics for XAI methods since, to this day, there is no generally accepted metric that quantifies how comprehensive and understandable explanations are for humans. Finally, there is a lack of regulation on how to certify XAI methods. More work is needed to understand which type of explanation – and thus which type of XAI method – is the best one for the given task and the stakeholders.

Acknowledgements

Know-Center is a leading European research center for big data, artificial intelligence (AI) and data-driven business models. Know-Center is a COMET Centre within COMET – Competence Centers for Excellent Technologies. This program is funded by the Austrian Federal Ministries for Climate Policy, Environment, Energy, Mobility, Innovation and Technology (BMK), and for Labor and Economy (BMAW), represented by Österreichische Forschungsförderungsgesellschaft mbH (FFG), Steirische Wirtschaftsförderungsgesellschaft mbH (SFG) and the Province of Styria, Wirtschaftsagentur Vienna and Standortagentur Tyrol GmbH.

References

Adadi, A and Berrada, M. (2018). Peeking Inside the Black-Box: A Survey on Explainable Artificial Intelligence (XAI). IEEE Access, 6, 52138–52160. https://doi.org/10.1109/ACCESS.2018.2870052

Arteaga, C., https://nbviewer.org/url/arteagac.github.io/blog/lime_image.ipynb, Last Accessed: 7 June 2023.

Bau D., Zhu JY., Strobelt, H., Zhou B.,Tenenbaum J.B., Freeman, W.T. and Torralba A. GAN Dissection: Visualizing and Understanding Generative Adversarial Networks, Proceedings of the International Conference on Learning Representations (ICLR), 2019.

Bau D., Zhu JY., Strobelt H., Lapedriza A., Zhou B. and Torralba A. Understanding the role of individual units in a deep neural network. Proceedings of the National Academy of Sciences (2020).

Chattopadhay, A., Sarkar, A., Howlader, P. and Balasubramanian, V. N. (2018). Grad-CAM++: Generalized Gradient-Based Visual Explanations for Deep Convolutional Networks. 2018 IEEE Winter Conference on Applications of Computer Vision (WACV), 839–847. https://doi.org/10.1109/WACV.2018.00097

Chromik, M., & Butz, A. (2021). Human-XAI Interaction: A Review and Design Principles for Explanation User Interfaces. In Ardito C., Lanzilotti R., Malizia A., Petrie H., Piccinno A., Desolda G. and Inkpen K. (Eds.), Human-Computer Interaction – INTERACT 2021 (Vol. 12933, pp. 619–640). Springer International Publishing. https://doi.org/10.1007/978-3-030-85616-8_36

Ding, W., Abdel-Basset, M., Hawash, H. and Ali A. M. (2022). Explainability of artificial intelligence methods, applications and challenges: A comprehensive survey. Information Sciences.

Hedström, A., Weber, L., Bareeva, D., Krakowczyk, D., Motzkus, F., Samek, W., Lapuschkin S. and Höhne, M. M. C. (2023). Quantus: An Explainable AI Toolkit for Responsible Evaluation of Neural Network Explanations and Beyond. Journal of Machine Learning Research, 24(34), 1-11.

Holzinger, A., Langs, G., Denk, H., Zatloukal, K. and Müller, H. (2019). Causability and explainability of artificial intelligence in medicine. WIREs Data Mining and Knowledge Discovery, 9(4). https://doi.org/10.1002/widm.1312

Larsson, S. and Heintz, F. (2020). Transparency in artificial intelligence. Internet Policy Review, 9(2). https://doi.org/10.14763/2020.2.1469

Lepri, B., Oliver, N., Letouzé, E., Pentland, A. and Vinck, P. (2018). Fair, transparent, and accountable algorithmic decision-making processes: The premise, the proposed solutions, and the open challenges. Philosophy & Technology, 31, 611-627.

Meske, C., Bunde, E., Schneider, J. and Gersch, M. (2022). Explainable Artificial Intelligence: Objectives, Stakeholders, and Future Research Opportunities. Information Systems Management, 39(1), 53–63. https://doi.org/10.1080/10580530.2020.1849465

Russakovsky O., Deng J., Su H., Krause J., Satheesh S., Ma S., Huang Z., Karpathy A., Khosla A., Bernstein M., Berg A.C. and Li Fei-Fei. ImageNet Large Scale Visual Recognition Challenge. IJCV, 2015.

Pope, P. E., Kolouri, S., Rostami, M., Martin, C. E., and Hoffmann, H. (2019). Explainability Methods for Graph Convolutional Neural Networks. 2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 10764–10773. https://doi.org/10.1109/CVPR.2019.01103

Ribeiro, M. T., Singh, S., and Guestrin, C. (2016, August). " 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 (pp. 1135-1144).

Weidele, D. K. I., Weisz, J. D., Oduor, E., Muller, M., Andres, J., Gray, A. and Wang, D. (2020). AutoAIViz: Opening the blackbox of automated artificial intelligence with conditional parallel coordinates. Proceedings of the 25th International Conference on Intelligent User Interfaces, 308–312. https://doi.org/10.1145/3377325.3377538

Transparency and Explainability in AI

SGS.COM/DIGITAL

CONTACT US

SGS

Emerging Technology

✉ Enquiry.Emerging-Technology@sgs.com

KNOW CENTER

Leading Research and Innovation Center for Trustworthy AI

https://know-center.at/

✉ info@know-center.at

WHEN YOU NEED TO BE SURE