On Explanations for Hybrid Artificial Intelligence

Lars Nolle[1,2](✉), Frederic Stahl[2], and Tarek El-Mihoub[2]

[1] Jade University of Applied Sciences, Wilhelmshaven, Germany
lars.nolle@jade-hs.de
[2] German Research Center for Artificial Intelligence, Oldenburg, Germany
{frederic_theodor.stahl,tarek.elmihoub}@dfki.de

Abstract. The recent developments of machine learning (ML) approaches within artificial intelligence (AI) systems often require explainability of ML models. In order to establish trust in these systems, for example in safety critical applications, a number of different explainable artificial intelligence (XAI) methods have been proposed, either post-hoc or intrinsic models. These can help to understand why a ML model has made a particular decision. The authors of this paper point out that the abbreviation XAI is commonly used in the literature referring to explainable ML models, although the term AI encompasses many more topics than ML. To improve efficiency and effectiveness of AI, two or more AI subsystems are often combined to solve a common problem. In this case, an overall explanation has to be derived from the subsystems' explanations. In this paper we define the term hybrid AI. This is followed by reviewing the current state of XAI before proposing the use of blackboard systems (BBS) to not only share results but also to integrate and to exchange explanations of different XAI models as well, in order to derive an overall explanation for hybrid AI systems.

Keywords: Hybrid Artificial Intelligence · Trust · Blackboard Systems

1 Introduction

In recent years, artificial intelligence (AI) has found its way from the research laboratories into many modern-day applications, ranging from personal assistants [1] to autonomous vehicles [2]. To keep up with the technological and social-economic challenges and developments of AI, the European Commission has developed an AI strategy, aiming at boosting excellence in AI and developing trustworthy AI in Europe [3]. State-of-the-art machine learning (ML) models [4], like deep neural networks (DNNs) [5], are often based on extremely complex non-linear functions, which makes it difficult to understand the inner workings of the trained models for humans. Consequently, the outputs of ML models, such as DNNs, are non-interpretable and non-transparent, which limits the trust in the overall system. Especially safety-critical systems [6] and safety critical applications [7, 8] require transparency to be considered reliable and trustworthy. Furthermore, a lack of transparency can have severe consequences in high-stakes domains,

M. Bramer and F. Stahl (Eds.): SGAI 2023, LNAI 14381, pp. 3–15, 2023.
https://doi.org/10.1007/978-3-031-47994-6_1

like medical diagnosis or financial decision-making [9]. This, for example, prompted the European Union Aviation Safety Agency (EASA) to lay out its AI Roadmap in 2020, to ensure that future ML-based systems can be safely integrated into the aviation domain [10].

Despite lack of interpretability, multiple AI methodologies are often combined to form hybrid AI systems for solving a mutual problem more effectively and efficiently. This aims to bring together different and currently separated AI techniques, including low-level perception and high-level reasoning [11]. Many real-world scientific and industrial applications require the results and recommendations derived by AI systems to be trustworthy and explainable. In hybrid AI systems, different types of AI methods collaborate on a mutual problem to arrive at decisions or recommendations for actions.

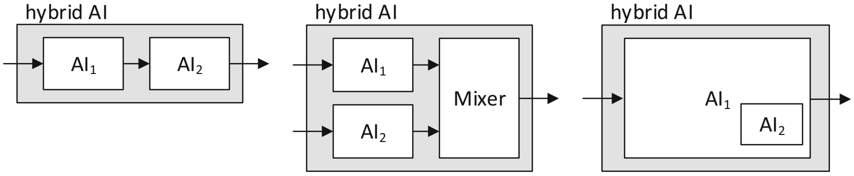
Blackboard systems (BBS) can be used to integrate and make different types of AI interact and use each other's results [12]. This paper proposes the use of BBS as a possible architecture for hybrid AI Systems where different AI models can exchange/access each other's explanations to derive a global solution.

Section 2 defines and explains the term hybrid AI for the context of this paper and Sect. 3 distinguishes different approaches to explainable and interpretable AI systems. Section 4 poses the research question how to combine explanations produced by different XAI methods in various stages of the hybrid AI system. Section 5 presents the proposed architecture before Sect. 6 summarises the presented work and discusses future work.

2 Hybrid Artificial Intelligence

For increased effectiveness and efficiency, two or more AI methods are often combined to solve a common problem. For example, Bielecki and Wojcik [13] recently used such a hybrid AI system based on ART neural networks and Gaussian mixture models for the monitoring of wind turbines. Tachmazidis et al. [14] used a hybrid approach, consisting of a ML model and a knowledge model, which captures the expertise of medical experts through knowledge engineering. The authors in [15] combined artificial neural networks, particle swarm optimisation and K-harmonic means clustering for colour design. Zheng et al. [16] proposed a hybrid AI model for COVID-19 prediction.

In this context, we define hybrid AI as a combination of two or more AI subsystems. There are, in principle, three ways of combining two AIs, in sequence, in parallel, or embedded (Fig. 1). When arranged in sequence, the output of the first AI is used as an input into the second AI, which produces the overall solution.



a) sequential hybrid AI b) parallel hybrid AI c) embedded hybrid AI

Fig. 1. Three different ways of combining two AI subsystems, AI$_1$ and AI$_2$.

When the AIs are arranged in parallel, they work independently of each other, either on the same data or on different data, and their output needs to be combined in a subsequent mixer stage. For example, a voting system can be used in this stage to produce the final output. Finally, an AI can be embedded into another AI in order to enhance the problem-solving potential of this AI, which would produce the overall output, respectively the solution to the problem [17]. For more than two AI subsystems, any combination of the above are possible.

There is also a need for trust in hybrid AI systems, hence researchers have recently begun to work on making hybrid AI systems explainable. For example, Li, et al. [6], proposed a vision-based object detection and recognition framework for autonomous driving. Here they used an optimized YOLOv4 model for object detection together with CNN models for recognition tasks. For the generation of explanations for the classification results, they used saliency maps-based algorithms. Another example can be found in [18]. Here, a hybrid conceptual/ML model for streamflow predictions was developed, and two model-agnostic techniques were subsequently applied. However, in both examples, the models have been made explainable only partially. Developing a holistic model-agnostic approach for generic hybrid AI models is still an open research question [19].

3 Explainable Artificial Intelligence

In the XAI literature, the terms explainability and interpretability are often used interchangeably [20, 21], although explainability has a wider meaning than interpretability [22]. Interpretability is often associated to answering the question of *"why?"*, related to a specific phenomenon and based on a specific opinion. Meanwhile, explainability is the ability to provide a set of related inference rules to answer the questions of *"how?"* and *"why?"* [22]. An explanation relies on facts, which can be described by words or formulas. Explanations can reveal the facts and the rules that are governing the behaviour of a phenomenon. According to [21], an explanation in AI has a different meaning from its traditional meaning and does not require interpretability. They also view causal explanations as the strictest form of scientific explanation. Kim et al. [21] also provided practical guidance for developing XAIs by defining fundamental requirements for such a system. An explanation, according to [23], has a flexible philosophical concept of "satisfying the subjective curiosity for causal information". Explainability in the context of XAI is a concept that enables understanding the overall strengths and weaknesses of AI models, predicting their behaviours and taking corrective actions [24]. However, XAI often shares a common aim of making AI understandable for people. This paper adopts the pragmatic definition of XAI stated in [25], where XAI is considered as broadly encompassing all techniques that service making AI understandable, such as direct interpretability, generating an explanation or justification, providing transparency information, etc.

The main aims of XAI are to establish trust with the stakeholders and to confirm compliance with ethics and regulations. XAI can help in deep understanding of AI models' behaviours and the problems they solve [26]. It is our position that, to achieve these aims, the framework for developing AI models should be adapted so that an AI model produces explanations of the solution in addition to the solution itself. Explaining the solutions introduces changes in the AI model's representation and also adds an explanation interface to the XAI model [24]. Interpretable models and deep explanation approaches can be followed to incorporate explainability within AI models. Interpretable models, also called glass-box or intrinsic models, seek to combine the clarity of the internal behaviour of an AI model with high quality performance. DARPA claims that there is a trade-off between model accuracy and explainability [27], which is a widely accepted view. However, Rudin et al. [20] are of the opinion that there is no such trade-off. Instead, it is possible to have an explainable model with high accuracy. Explainable Boosting Machines (EBMs) [28], for example, support this claim. This technique is a generalised and more efficient version of the Generalised Additive Model (GAM) [29] that produces high-quality explainable models. TabNet [30] combines sequential attention with decision tree-based learning for interpretable and more efficient learning. Deep explanation approaches aim to benefit from the success of deep learning in solving complex problems to resolve the explanation problem. Deep explanations hybridise different deep learning models to produce richer representations of the features utilised during learning to enable extraction of underlying semantic information [31]. For example, a special prototype layer was added to a CNN to utilise case-based reasoning in explaining its predictions [32]. Lei et al. [33] have used extractive reasoning to incorporate interpretation in the framework of a neural network. Generating accurate and suitable explanations of the model behaviour to a user is the main challenge of deep explanation models [24].

However, due to the urgent need for building trust and compliance with regulations and ethics in already existing AI models, induction or post-hoc approaches have been proposed. Figure 2 shows how, in principle, a trained ML (AI) model is post-hoc analysed by a model-agnostic method, which manipulates the input data and measures the changes in output in order to generate an explanation. Commonly, three different types of post-hoc explanations are used: alternative advice, prediction confidence scores, and prediction rationale [34].

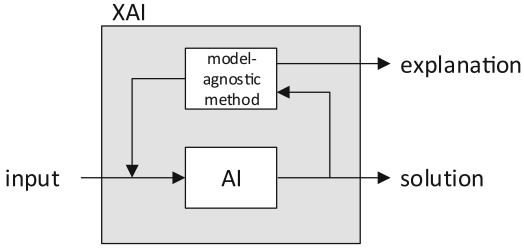


Fig. 2. Model-agnostic (post-hoc) explanation method for trained ML model AI.

Jiang et al. [34] have also shown that epistemic uncertainty is most important to users of post-hoc explanations. Meanwhile, the challenges associated with explaining black-box models, i.e. observable models with unknown transfer functions, have motivated researchers to develop different post-hoc interpretation and explanation facilities. Furthermore, the process of incorporating explanations within the framework of AI models sometimes can be more difficult than building a post-hoc tool [21]. Without providing explanations of its explanations, a post-hoc tool can be viewed as a black-box that explains another black-box [20]. It is also possible to generate two conflicting explanations for the same AI model's behaviour using two different post-hoc tools [32].

Various post-hoc tools have been developed. Some of these tools aim to interpret the general behaviour of an AI model, referred to as global post-hoc tools; others focus on a specific behaviour of the model with a specific input or set of inputs, referred to as local post-hoc tools. The surrogate model approach has been used to develop a new simple model that mimics the behaviour of a black-box AI at a global or local level. The new model should be interpretable or explainable. TREPAN [35] and Rule Extraction From Neural network Ensemble (REFNE) [36] are examples of global post-hoc tools that follow this approach. Knowledge distillation [37] can be viewed as a unified method for model extraction. Local Interpretable Model-agnostic Explanations (LIME) [38] builds a linear classifier to mimic the local behaviour of the black-box AI. Another post-hoc approach is to estimate the features' impact on the behaviour of the black-box model. Estimation of the features' importance can be done at a global or a local level. Such estimation can help in ensuring that worthy features are controlling the behaviour of the model. A feature's importance can be presented as a score according to its impact on the model prediction; for example, by generating saliency maps [39] and SHapley Additive exPlanations (SHAP) [40]. The features' importance can be presented as a relation between each feature and the model's global prediction, such as Partial Dependence Profiles (PDP) [41] and Accumulated Local Effects (ALE) [42]. Individual Conditional Expectations (ICE) are used to present such kind of relation at local level [43]. Counterfactual examples [44] can also be used to explain the local behaviour of a black-box model. These examples are used to show how an input can be modified to change the model's prediction. Diverse Counterfactual Explanations (DiCE) [45] uses a set of diverse counterfactual examples to inspect the behaviour of AI models. A generative counterfactual introspection has been used to produce inherently interpretable counterfactual visual explanations in the form of prototypes and criticisms [46]. In addition to model-agnostic post-hoc explanations, there are explanations that make use of some knowledge of a black-box AI model to provide explanations. For example, Grad-CAM [47], uses the inputs and the gradients of a deep neural network to determine the salient pixels to the model prediction. Some of these methods can be applied to AI models with specific properties. For example, Integrated Gradient (IG) [48] can be applied to any differentiable model for different types of input such as images, text, or structured data.

Landscape analysis tools [49] are commonly used to explain the behaviour of population-based metaheuristics, such as evolutionary algorithms. These tools can also help in understanding complex optimisation problems. Furthermore, visualizing the trajectories followed by these algorithms can enhance researchers and developers' comprehension of the behaviours of different search algorithms [50]. Dimension reduction techniques are typically employed to simplify the visualisation of these trajectories [51]. A data-driven, graph-based model, Search Trajectory Network (STNs), has been utilised to illustrate the changes in the algorithm's behaviour throughout the search process [52].

4 Explainability for Hybrid Artificial Intelligence

The open research question, which is addressed in this work, is to ascertain how to combine explanations produced by different XAI methods in various stages of the hybrid AI system, so that it provides meaningful explanations to the end-user.

Our approach is, in the case of sequential XAI methods, each of the methods is producing an explanation, which is fed forward to the Explanation Mixer. This produces the overall explanation for the solution for a given input (Fig. 3).

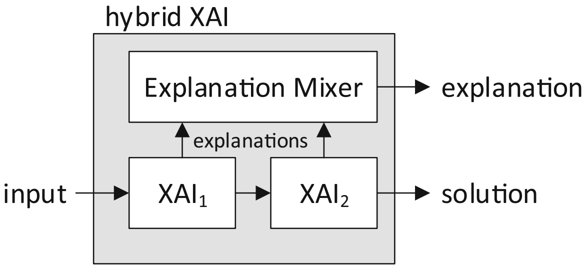


Fig. 3. Hybrid XAI consisting of two sequential XAIs and the Explanation Mixer.

In the case of parallel XAI subsystems, each of the methods is producing an explanation, which is forwarded to the Explanation Mixer (Fig. 4). The part solutions generated by each AI must be combined in a subsequent Solution Mixer stage, which produces the overall solution. Likewise, the Explanation Mixer generates the overall explanation for the overall solution.

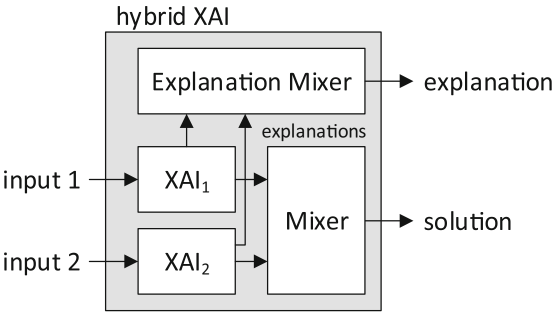


Fig. 4. Hybrid XAI consisting of two parallel XAI subsystems, an Explanation Mixer, and a Solution Mixer.

In the case of embedded XAI subsystems, the embedded XAI may be triggered multiple times during the execution of the master AI (XAI_1). The embedded AI (XAI_2) provides explanations and solutions for specific tasks to the hybrid XAI (Fig. 5). At this stage, it is not clear yet where these explanations can be incorporated in the master AI's explanation. This remains an open research question.

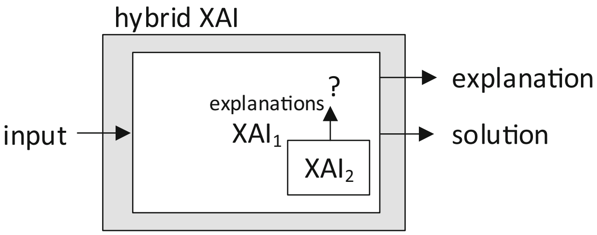


Fig. 5. Hybrid XAI consisting of two XAIs one embedded in a master XAI (XAI_1).

In the next section, an architecture is proposed for realizing these three types of hybrid AI systems.

5 Proposed Architecture

In order to allow for flexible and adaptive systems, the use of the BBS model is envisioned. BBSs facilitate the principles of a group of human experts, solving a common problem [53]. The experts group together around a blackboard. Each of the experts observes the blackboard constantly. If granted access by a moderator, they may add information to or remove information from the blackboard as a reaction to contents changes of the blackboard. By doing so, they contribute towards the global solution, which will evolve eventually on the blackboard. This approach has been proven very successful and is often facilitated in group decision-making processes.

In the BBS model, the human experts are replaced with so-called knowledge sources, i.e. data/information sources and algorithms, the latter often from the field of AI. The analogue to the blackboard is a common database system, and the analogue to the moderator is a scheduler. In such a BBS, the knowledge of the problem domain is distributed over several specialised knowledge sources, also known as agents [54]. The agents are autonomous and communicate with each other only by reading information from and writing information to the common database. Each AI method used in a particular application is implemented as an autonomous knowledge source.

BBSs were successfully employed to a wide range of different problems, ranging from improving classification accuracy in ML [55] or the control of a complex autonomous spacecraft [56], to the automated generation of poetry [57]. He et al. [58] used a BBS for controlling an Earth observation satellite. Stewart et al. [59] used an agent-based BBS for reactor design. Xu and Smith [60] achieved massive data sharing in distributed and heterogeneous environments using a BBS to reduce data sharing delay. However, there are still open research questions related to BBSs. For example, how to allow access to the common data repository [12] or how to maintain the blackboard over a long period of time [61]. There are different types of blackboard architectures available. A distinction can be made between the original monolithic architecture [62], distributed BBSs [12, 63] and fractal BBSs [64]. It is important to choose the appropriate architecture for a problem at hand. However, the BBS model is very flexible, i.e., it can be used to implement both, the sequential hybrid system, and the parallel hybrid system [65]. It is also possible, to change the configuration dynamically during runtime.

Figure 6 shows the proposed architecture based on the BBS design. It consists of application specific data sources and (X)AI modules as well as the generic BBS to produce an overall output (solution) to the problem (input).

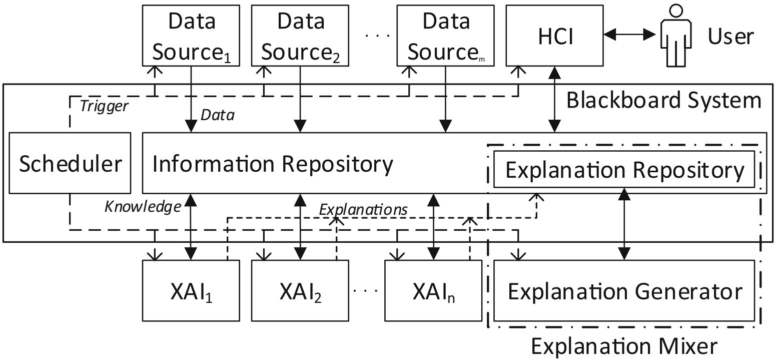


Fig. 6. Proposed architecture based on the Blackboard System design.

Here, *m* data sources receive input data from the environment and put it on the information repository, the shared data base of the Blackboard System. A scheduler is used to synchronise access to the information repository via trigger signals. At the same time, *n* XAI subsystems observe the data on the information repository in order to generate new knowledge, which is subsequently placed on the information repository. In addition, the XAI subsystems are writing their individual explanations on to the explanation repository, a specialized partition of the blackboard. These explanations are used by the application specific explanation generator to derive the overall explanation for the solutions. The solutions to the input data together with the overall explanations can be accessed via a human computer interface (HCI).

The data sources might supply multimodal data, e.g. images, text and sensor readings. This data might be unstructured, inconsistent, unreliable, and biased. Therefore, different AI algorithms must process the data to enable the detection of an event of interest and for deriving a recommendation for action. For example, there might be an AI algorithm for the identification of event related artefacts in pictures, like harmful algae bloom, or contaminants in bio-waste [66]. Another algorithm has to cluster the data, so that a record is associated with an individual event. Finally, a dedicated AI algorithm must make an expert decision about the positive identification of an event of interest.

If all these different AI algorithms also produce explanations, these explanations must be fused into an overall explanation, suitable for a human user. In order to be able to exchange knowledge in hybrid systems, domain-specific ontologies are often required. In computer science, an ontology is an explicit, formal specification of a shared conceptualisation [67]. An explicit formal representation facilitates sharing of knowledge and human-machine interaction. Utilising the concepts of ontology enables reusing and analysing domain knowledge. Formalising these concepts through logic languages ensures consistency of a domain knowledge, enabling extracting relations and reasoning. For example, a medical-ontology has been used in Doctor XAI [68] to build a model-agnostic explanation to deal with multi-labelled, sequential, ontology-linked data. Doctor XAI shows that utilising medical knowledge can produce a better approximation of the local behaviour of a black-box model. In [69] an explanation ontology is proposed to support user-centred AI system design. When applications span over different domains, their associated ontologies have to be aligned.

6 Summary and Future Work

In this work, the term hybrid AI was defined and examples of current applications of such hybrid systems were introduced. A need for trust in hybrid AI systems was identified. Subsequently, a survey of current XAI methods was provided. We presented our proposed architecture for hybrid XAI, which is based on the blackboard architecture. Here, a specialised partition of the information repository is used to collect the individual explanations from the knowledge sources, i.e. the XAI subsystems. In order to derive an overall explanation, an application specific explanation generator was proposed. An application specific ontology has to be followed to facilitate exchanging and sharing knowledge and explanations.

The proposed architecture is currently under development and will be used in subsequent research. For this, a number of research questions are still open: (i) How can multimodal explanations be formulated using an application specific ontology? (ii) How to combine such explanations in order to generate a meaningful explanation to the user? (iii) How to combine explanations in embedded hybrid AI systems? To find answers to these questions, the DFKI is currently conducting a 1.7M€ research project, which builds upon this proposed architecture, and aims at the automated scheduling of weed harvesting campaigns on lakes.

Acknowledgements. Partly funded by the Federal Ministry of Education and Research, Germany, grant number 01IW23003, and the Ministry for Science and Culture of Lower Saxony, grant number ZN3683.

References

1. Maedche, A., et al.: AI-based digital assistants: opportunities, threats, and research perspectives. Bus. Inf. Syst. Eng. **61**, 535–544 (2019). https://doi.org/10.1007/s12599-019-00600-8
2. Gao, X., Bian, X.: Autonomous driving of vehicles based on artificial intelligence. J. Intell. Fuzzy Syst. **41**, 1–10 (2021). https://doi.org/10.3233/JIFS-189982
3. EC. Artificial Intelligence for Europe, European Commission, COM (2018) 237. European Commission (2018)
4. Jordan, M.I., Mitchell, T.M.: Machine learning: trends, perspectives, and prospects. Science **349**, 255–260 (2015). https://doi.org/10.1126/science.aaa8415
5. Schmidhuber, J.: Deep learning in neural networks: an overview. Neural Net. **61**, 85–117 (2014). https://doi.org/10.1016/j.neunet.2014.09.003
6. Li, Y., et al.: A deep learning-based hybrid framework for object detection and recognition in autonomous driving. IEEE Access **8**, 194228–194239 (2020). https://doi.org/10.1109/ACCESS.2020.3033289

7. Hernandez, C.S., Ayo, S., Panagiotakopoulos, D.: An explainable artificial intelligence (xAI) framework for improving trust in automated ATM tools. In: 2021 IEEE/AIAA 40th Digital Avionics Systems Conference (DASC), pp. 1–10 (2021). https://doi.org/10.1109/DASC52595.2021.9594341
8. Wang, Y., Chung, S.: Artificial intelligence in safety-critical systems: a systematic review. Ind. Manag. Data Syst. **122**(2), 442–470 (2021). https://doi.org/10.1108/IMDS-07-2021-0419
9. Zhou, J., Gandomi, A.H., Chen, F., Holzinger, A.: Evaluating the quality of machine learning explanations: a survey on methods and metrics. Electronics **10**, 593 (2021). https://doi.org/10.3390/electronics10050593
10. EASA. Artificial intelligence roadmap: a human-centric approach to AI aviation. European Union Aviation Safety Agency (2020)
11. Kersting, K.: Rethinking computer science through AI. KI - Künstliche Intelligenz **34**(4), 435–437 (2020). https://doi.org/10.1007/s13218-020-00692-5
12. Nolle, L., Wong, K.C.P., Hopgood, A.A.: DARBS: a distributed blackboard system. In: Bramer, M., Coenen, F., Preece, A. (eds.) Research and Development in Intelligent Systems XVIII, pp. 161–170. Springer, London (2002). https://doi.org/10.1007/978-1-4471-0119-2_13
13. Bielecki, A., Wójcik, M.: Hybrid AI system based on ART neural network and Mixture of Gaussians modules with application to intelligent monitoring of the wind turbine. Appl. Soft Comput. **108**, 107400 (2021). https://doi.org/10.1016/j.asoc.2021.107400
14. Tachmazidis, I., Chen, T., Adamou, M., Antoniou, G.: A hybrid AI approach for supporting clinical diagnosis of attention deficit hyperactivity disorder (ADHD) in adults. Health Inf. Sci. Syst. **9**, 1 (2021). https://doi.org/10.1007/s13755-020-00123-7
15. Li, M., et al.: A decision support system using hybrid AI based on multi-image quality model and its application in color design. Future Gener. Comput. Syst. **113**, 70–77 (2020). https://doi.org/10.1016/j.future.2020.06.034
16. Zheng, N., et al.: Predicting COVID-19 in China using hybrid AI model. IEEE Trans. Cybern. **50**, 2891–2904 (2020). https://doi.org/10.1109/TCYB.2020.2990162
17. El-Mihoub, T., Hopgood, A.A., Nolle, L., Battersby, A.: Hybrid genetic algorithms – a review. Eng. Lett. **13**(2), 124–137 (2006). ISSN: 1816-093X
18. Althoff, D., Bazame, H.C., Nascimento, J.G.: Untangling hybrid hydrological models with explainable artificial intelligence. H2Open J. **4**, 13–28 (2021). https://doi.org/10.2166/h2oj.2021.066
19. Akata, Z., et al.: A research agenda for hybrid intelligence: augmenting human intellect with collaborative, adaptive, responsible, and explainable artificial intelligence. Computer **53**, 18–28 (2020). https://doi.org/10.1109/MC.2020.2996587
20. Rudin, C., Chen, C., Chen, Z., Huang, H., Semenova, L., Zhong, C.: Interpretable machine learning: fundamental principles and 10 grand challenges. arXiv (2021). https://doi.org/10.48550/ARXIV.2103.11251
21. Kim, M.-Y., et al.: A multi-component framework for the analysis and design of explainable artificial intelligence. Mach. Learn. Knowl. Extract. **3**, 900–921 (2021). https://doi.org/10.3390/make3040045
22. Buhrmester, V., Münch, D., Arens, M.: Analysis of explainers of black box deep neural networks for computer vision: a survey. Mach. Learn. Knowl. Extract. **3**, 966–989 (2021)
23. Li, X.-H., et al.: A survey of data-driven and knowledge-aware eXplainable AI. IEEE Trans. Knowl. Data Eng. **34**(1), 29–49 (2020)
24. Gunning, D., Vorm, E., Wang, J.Y., Turek, M.: DARPA's explainable AI (XAI) program: a retrospective. Appl. AI Lett. (2021)
25. Liao, Q.V., Varshney, K.R.: Human-centered explainable AI (XAI): from algorithms to user experiences, CoRR, Bd. abs/2110.10790 (2021)

26. El-Mihoub, T.A., Nolle, L., Stahl, F.: Explainable boosting machines for network intrusion detection with features reduction. In: Bramer, M., Stahl, F. (eds.) Artificial Intelligence XXXIX: 42nd SGAI International Conference on Artificial Intelligence, AI 2022, Cambridge, UK, December 13–15, 2022, Proceedings, pp. 280–294. Springer, Cham (2022). https://doi.org/10.1007/978-3-031-21441-7_20
27. Gunning, D., Aha, D.: DARPA's explainable artificial intelligence (XAI) program. AI Mag. **40**(2), 44–58 (2019)
28. Nori, H., Jenkins, S., Koch, P., Caruana, R.: InterpretML: a unified framework for machine learning interpretability. arXiv (2019)
29. Hastie, T., Tibshirani, R.: Generalized additive models: some applications. J. Am. Stat. Assoc. **82**(398), 371–386 (1987)
30. Arik, S.O., Pfister, T.: TabNet: attentive interpretable tabular learning. In: Proceedings of the AAAI Conference on Artificial Intelligence, vol. 35, pp. 6679–6687 (2021). https://ojs.aaai.org/index.php/AAAI/article/view/16826
31. Park, D.H., et al.: Multimodal explanations: justifying decisions and pointing to the evidence. In: 2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition (2018)
32. Li, O., Liu, H., Chen, C., Rudin, C.: Deep learning for case-based reasoning through prototypes: a neural network that explains its predictions. In: The Thirty-Second AAAI Conference, pp. 3530–3537 (2018)
33. Lei, T., Barzilay, R., Jaakkola, T.: Rationalizing neural predictions. arXiv (2016). https://doi.org/10.48550/ARXIV.1606.04155
34. Jiang, J., Kahai, S., Yang, M.: Who needs explanation and when? Juggling explainable AI and user epistemic uncertainty. Int. J. Hum. Comput. Stud. **165**, 102839 (2022)
35. Craven, M.W., Shavlik, J.W.: Extracting tree-structured representations of trained networks. In: Proceedings of the 8th International Conference on Neural Information Processing Systems, Denver, Colorado, pp. 24–30 (1995)
36. Zhou, Z.-H., Jiang, Y., Chen, S.-F.: Extracting symbolic rules from trained neural network ensembles. AI Commun. **16**(1), 3–15 (2003)
37. Hinton, G., Vinyals, O., Dean, J.: Distilling the knowledge in a neural network. arXiv (2015). https://doi.org/10.48550/ARXIV.1503.02531
38. Ribeiro, M.T., Singh, S., Guestrin, C.: Why should i trust you?: Explaining the predictions of any classifier. In: The 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, San Francisco, California, USA (2016)
39. Fong, R.C., Vedaldi, A.: Interpretable explanations of black boxes by meaningful perturbation. In: 2017 IEEE International Conference on Computer Vision (ICCV), pp. 3449–3457 (2017). https://doi.org/10.1109/ICCV.2017.371
40. Lundberg, S.M., Lee, S.-I.: A unified approach to interpreting model predictions. In: The 31st International Conference on Neural Information Processing Systems, Long Beach, California, USA, pp. 4768–4777 (2017)
41. Friedman, J.F.: Greedy function approximation: a gradient boosting machine. Ann. Stat. **29**, 1189–1232 (2001)
42. Apley, D.W., Zhu, J.: Visualizing the effects of predictor variables in black box supervised learning models. arXiv (2016). https://doi.org/10.48550/ARXIV.1612.08468
43. Goldstein, A., Kapelner, A., Bleich, J., Pitkin, E.: Peeking inside the black box: visualizing statistical learning with plots of individual conditional expectation. J. Comput. Graph. Stat. **24**(1), 44–65 (2015)
44. Karimi, A.-H., Barthe, G., Balle, B., Valera, I.: Model-agnostic counterfactual explanations for consequential decisions. In: Chiappa, S., Calandra, R. (ed.) Proceedings of the Twenty Third International Conference on Artificial Intelligence and Statistics, vol. 108, pp. 895–905. PMLR (2020). https://proceedings.mlr.press/v108/karimi20a.html

45. Mothilal, R.K., Sharma, A., Tan, C.: Explaining machine learning classifiers through diverse counterfactual explanations. In: FAT* 2020, Barcelona, Spain (2020)
46. Liu, S., Kailkhura, B., Loveland, D., Han, Y.: Generative counterfactual introspection for explainable deep learning. In: 2019 IEEE Global Conference on Signal and Information Processing (GlobalSIP) (2019)
47. Selvaraju, R.R., Cogswell, M., Das, A., Vedantam, R., Parikh, D., Batra, D.: Grad-CAM: visual explanations from deep networks via gradient-based localization. In: 2017 IEEE International Conference on Computer Vision (ICCV), pp. 618–626 (2017). https://doi.org/10.1109/ICCV.2017.74
48. Sundararajan, M., Taly, A., Yan, Q.: Axiomatic attribution for deep networks. In: Proceedings of the 34th International Conference on Machine Learning, Sydney, vol. 70, pp. 3319–3328. JMLR.org (2017)
49. Malan, K.M.: A survey of advances in landscape analysis for optimisation. Algorithms **14**(2), 40 (2021)
50. Michalak, K.: Low-dimensional euclidean embedding for visualization of search spaces in combinatorial optimization. In: Proceedings of the Genetic and Evolutionary Computation Conference Companion, New York, NY, USA (2019)
51. De Lorenzo, A., Medvet, E., Tušar, T., Bartoli, A.: An analysis of dimensionality reduction techniques for visualizing evolution. In: Proceedings of the Genetic and Evolutionary Computation Conference Companion, New York, NY, USA (2019)
52. Ochoa, G., Malan, K.M., Blum, C.: Search trajectory networks: a tool for analysing and visualising the behaviour of metaheuristics. Appl. Soft Comput. **109**, 107492 (2021)
53. Serafini, L., et al.: On some foundational aspects of human-centered artificial intelligence. arXiv preprint arXiv:2112.14480 (2021)
54. Weitz, K., Schiller, D., Schlagowski, R., Huber, T., André, E.: “Let me explain!”: exploring the potential of virtual agents in explainable AI interaction design. J. Multimodal User Interfaces **15**(2), 87–98 (2021). https://doi.org/10.1007/s12193-020-00332-0
55. Kokorakis, V.M., Petridis, M., Kapetanakis, S.: A blackboard based hybrid multi-agent system for improving classification accuracy using reinforcement learning techniques. In: Bramer, M., Petridis, M. (eds.) SGAI 2017. LNCS (LNAI), vol. 10630, pp. 47–57. Springer, Cham (2017). https://doi.org/10.1007/978-3-319-71078-5_4
56. Golding, D., Chesnokov, A.M.: Features of informational control complex of autonomous spacecraft. In: IFAC Workshop Aerospace Guidance, Navigation and Flight Control Systems. International Federation of Automatic Control, Laxenburg (2011)
57. Misztal-Radecka, J., Indurkhya, B.: A blackboard system for generating poetry. Comput. Sci. **17**(2), 265–294 (2016)
58. He, L., Li, G., Xing, L., Chen, Y.: An autonomous multi-sensor satellite system based on multi-agent blackboard model. Maintenance Reliab. **19**(3), 447–458 (2017)
59. Stewart, R., Palmer, T.S., Bays, S.: Toward an agent-based blackboard system for reactor design optimization. Nucl. Technol. **208**(5), 822–842 (2021). https://doi.org/10.1080/00295450.2021.1960783
60. Xu, J.S., Smith, T.J.: Massive data storage and sharing algorithm in distributed heterogeneous environment. J. Intell. Fuzzy Syst. **35**(4), 4017–4026 (2018)
61. Straub, J.: Automating maintenance for a one-way transmitting blackboard system used for autonomous multi-tier control. Expert. Syst. **33**(6), 518–530 (2016)
62. Engelmore, R.S., Morgan, A.J.: Blackboard Systems. Addison-Wesley (1988)
63. McManus, J.W.: A concurrent distributed system for aircraft tactical decision generation. In: IEEE/AtAA/NASA 9th Digital Avionics Systems Conference, New York, USA, pp. 161–170 (1990)
64. Naaman, M., Zaks, A.: Fractal blackboard systems. In: Proceedings of the 8th Israeli Conference on Computer-Based Systems and Software Engineering, pp 23–29 (1997)

65. Stahl, F., Bramer, M.: Computationally efficient induction of classification rules with the PMCRI and J-PMCRI frameworks. Knowl.-Based Syst. **35**, 49–63 (2012)
66. Stahl, F., Ferdinand, O., Nolle, L., Pehlken, A., Zielinski, O.: AI enabled bio waste contamination-scanner. In: Bramer, M., Ellis, R. (eds.) Artificial Intelligence XXXVIII: 41st SGAI International Conference on Artificial Intelligence, AI 2021, Cambridge, UK, December 14–16, 2021, Proceedings, pp. 357–363. Springer, Cham (2021). https://doi.org/10.1007/978-3-030-91100-3_28
67. Gruber, T.R.: A translation approach to portable ontology specifications. Knowl. Acquis. **5**, 199–220 (1993)
68. Panigutti, C., Perotti, A., Pedreschi, D.: Doctor XAI: an ontology-based approach to black-box sequential data classification explanations, pp. 629–639. Association for Computing Machinery, New York (2020)
69. Chari, S., Seneviratne, O., Gruen, D.M., Foreman, M.A., Das, A.K., McGuinness, D.L.: Explanation ontology: a model of explanations for user-centered AI. In: Pan, J.Z., et al. (eds.) ISWC 2020. LNCS, vol. 12507, pp. 228–243. Springer, Cham (2020). https://doi.org/10.1007/978-3-030-62466-8_15