Explainable AI: A Brief Survey on History, Research Areas, Approaches and Challenges

Feiyu Xu[1], Hans Uszkoreit[2], Yangzhou Du[1], Wei Fan[1], Dongyan Zhao[3], and Jun Zhu[4]

[1] AI Lab, Lenovo Research, Lenovo Group, China
{fxu,duyz1,fanwei2}@lenovo.com
[2] DFKI GmbH, Germany and Giance Technologies
uszkoreit@dfki.de
[3] Institute of Computer Science and Technology, Peking University, China
zhaody@pku.edu.cn
[4] Department of Computer Science and Technology, Tsinghua University, China
dcszj@mail.tsinghua.edu.cn

Abstract. Deep learning has made significant contribution to the recent progress in artificial intelligence. In comparison to traditional machine learning methods such as decision trees and support vector machines, deep learning methods have achieved substantial improvement in various prediction tasks. However, deep neural networks (DNNs) are comparably weak in explaining their inference processes and final results, and they are typically treated as a black-box by both developers and users. Some people even consider DNNs (deep neural networks) in the current stage rather as *alchemy*, than as real science. In many real-world applications such as business decision, process optimization, medical diagnosis and investment recommendation, explainability and transparency of our AI systems become particularly essential for their users, for the people who are affected by AI decisions, and furthermore, for the researchers and developers who create the AI solutions. In recent years, the explainability and explainable AI have received increasing attention by both research community and industry. This paper first introduces the history of Explainable AI, starting from expert systems and traditional machine learning approaches to the latest progress in the context of modern deep learning, and then describes the major research areas and the state-of-art approaches in recent years. The paper ends with a discussion on the challenges and future directions.

Keywords: Explainable artificial intelligence · intelligible machine learning · explainable interfaces · XAI · interpretability

1 A Brief History of Explainable AI

In wiktionary, the word "explain" means for humans "to make plain, manifest, or intelligible; to clear of obscurity; to illustrate the meaning of" [23]. In scientific research, a scientific explanation is supposed to cover at least two parts: 1) the

object to be explained (the "Explanandum" in Latin), and 2) the content of explanation (the "Explanans" in Latin).

Explainable AI is not a new topic. The earliest work on Explainable AI could be found in the literature published forty years ago [16] [19], where some expert systems explained their results via the applied rules. Since AI research began, scientists have argued that intelligent systems should explain the AI results, mostly when it comes to decisions. If a rule-based expert system rejects a credit card payment, it should explain the reasons for the negative decision. Since the rules and the knowledge in the expert systems are defined and formulated by human experts, these rules and knowledge are easy for humans to understand and interpret. Decision tree is a typical method designed with explainable structure. As illustrated in Fig. 1, starting at the top and going down, the solution path in the decision tree presents the reasoning of a final decision.

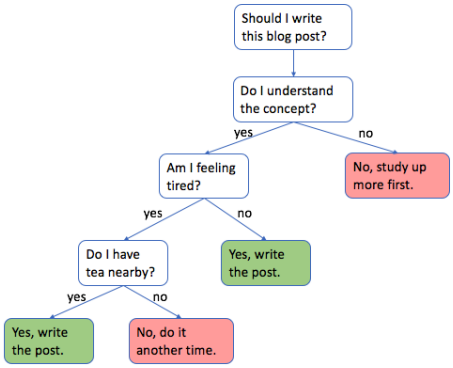

Fig. 1. An example of decision tree, used by starting at the top and going down, level by level, according to the defined logic. (Image courtesy of J. Jordan [10])

However, Explainable AI has become a new research topic in the context of modern deep learning. Without completely new explanatory mechanisms, the output of today's Deep Neural Networks (DNNs) cannot be explained, neither by the neural network itself, nor by an external explanatory component, and not even by the developer of the system. We know that there are different architectures of DNNs designed for different problem classes and input data, such as CNN, RNN, LSTM, shown in Fig. 2. All of them have to be considered as black boxes - whose internal inference processes are neither known to the observer nor interpretable by humans [7].

Explainability of a machine learning model is usually inverse to its prediction accuracy - the higher the prediction accuracy, the lower the model explainability. The DARPA Explainable AI (XAI) program presents a nice chart to illustrate this interesting phenomena, as shown in Fig. 3, where decision trees have an excellent degree of explainability but exhibit worst prediction accuracy among the listed learning techniques. In the other extreme, Deep Learning methods are

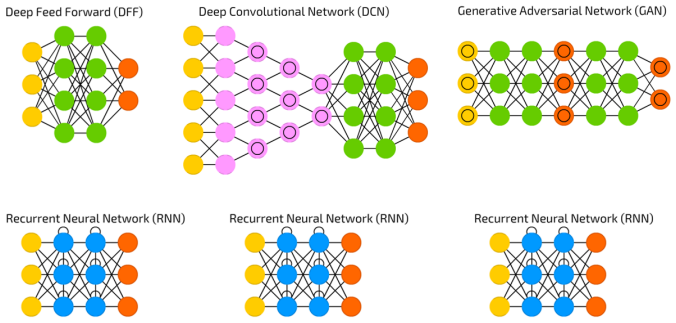

Fig. 2. A chart of several typical Deep Neural Networks (DNNs). (Image courtesy of Fjodor Van Veen [22])

better in predictive capacity than any other learning methods but they are least likely to be explicable.

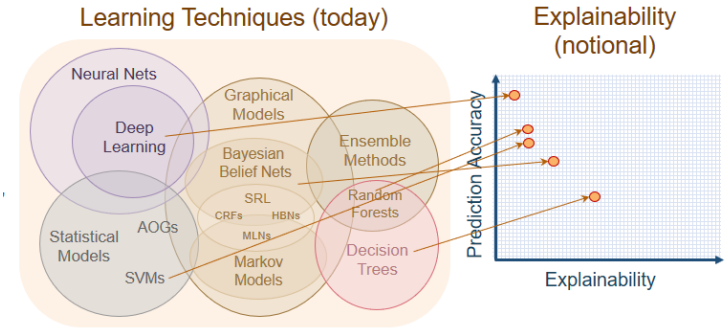

Fig. 3. Explainability of machine learning models appear inverse to their prediction accuracy. (Image courtesy of DARPA [21])

In recent years, AI researchers aim to open the black-box of neural networks and turn it into a transparent system. As shown in Fig. 4, there are two main strands of work in Explainable AI - transparency design and post-hoc explanation. The transparency design reveals how a model functions, in the view of developers. It tries to (a) understand model structure, e.g., the construction of a decision tree; (b) understand single components, e.g., a parameter in logistic regression; (c) understand training algorithms, e.g., solution seeking in a convex optimization. The post-hoc explanation explains why a result is inferred, in the view of users. It tries to (d) give analytic statements, e.g. why a goods is recommended in a shopping website; (e) give visualizations, e.g. saliency map is used to show pixel importance in a result of object classification; (f) give ex-

planations by example, e.g. K-nearest-neighbors in historical dataset are used to support current results. A thorough description of the categorization of explanation methods is found in Lipton et al. [13]. A comprehensive survey on recent development of Explainable AI is provided in [5].

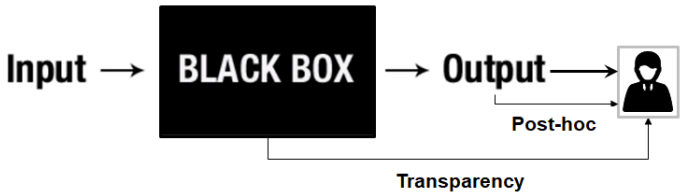

Fig. 4. Two categories of Explainable AI work: transparency design and post-hoc explanation.

2 Relevance of Explainable AI

Increasing attention has recently been paid to Explainable AI across the world both in research and in industry. In April 2017, DARPA funded the "Explainable AI (XAI) program", aimed at improving explainability of AI decision [21]. In July 2017, the Chinese government released "The Development Plan for New Generation of Artificial Intelligence" to encourage high-explainability AI and strong-extensibility AI [18]. In May 2018, the "General Data Protection Regulation" (GDPR) was published, in which the European Union grants their citizens a "right to explanation" if they are affected by algorithmic decision-making [6]. Explainable AI will become increasingly important to all groups of stakeholders, including the users, the affected people, and the developers of AI systems.

Explainable AI is important to the users who utilize the AI system. When the AI recommends a decision, the decision makers would need to understand the underlying reason. For example, medical doctor needs to understand what pathological features in the input data were were guiding the algorithm before accepting auto-generated diagnosis reports. A maintenance engineer needs to understand which abnormal phenomena were captured by the inference algorithm before following the repair recommendations. A financial investor wants to understand what influencing factors were regarded as the critical ones by the system algorithm before making the final investment decision. We have to verify that the AI inference works as expected, because wrong decisions can be costly and dangerous. Caruana et al. [3] presented a famous example "Pneumonia - Asthma" to illustrate this point. An AI system which had been trained to predict the pneumonia risk of a person arrived at totally wrong conclusions. From real data the model had learned that asthmatic patients with heart problems have a much lower risk of dying of pneumonia than healthy persons. This cannot be true since asthma is a factor that negatively affects the recovery. The training

data were systematically biased, because in contrast to healthy persons, the majority of these asthma patients were under strict medical supervision. Hence this group had a significant lower risk of dying of pneumonia. It should be noted, though, that both the learning and the inference algorithms probably worked correctly and also that the training data represented real cases. The insight that the selection of the training data was not appropriate for predictions affecting other populations may remain undiscovered if we have a black-box AI system.

Explainable AI is important to the people who are affected by AI decision. If the AI makes its own decisions, e.g., braking of the car, shutting down a plant, selling shares, assessing a job, issuing a traffic punishment order, the affected people must be able to understand the reason. There are already legal regulations that codify this demand [6]. Houston schools were using an AI algorithm, called Educational Value-Added Assessment System (EVAAS), to evaluate the performance of teachers. However, this AI system was successfully contested by teachers in court, because negative reviews of teachers could not be explained by the AI system [2].

Explainable AI could help developers to improve AI algorithm, by detecting data bias, discovering mistakes in the models, and remedying the weakness. Lapuschkin et al. [11] presented an impressive example. As shown in Fig. 5, they observed that the Fisher Vector method usually shows lower accuracy than Deep Neural Networks in the task of object recognition. However, two methods reach almost equal accuracy of recognition rate in the category "horse", which is unexpected. A saliency map method called "Layer-wise Relevance Propagation" [12] was then employed to analyze which pixel areas exactly make the models arrive at their predictions. The authors observed that the two models use different strategies to classify images of that category. The Deep Neural Network looked at the contour of the actual horse, whereas the Fisher Vector model mostly relied on a certain copyright tag, that happens to be present on many horse images. Removing the copyright tag in the test images would consequently significantly decrease the accuracy of the Fisher Vector model.

	aeroplane	bicycle	bird	boat	bottle	bus	car
Fisher	79.08%	66.44%	45.90%	70.88%	27.64%	69.67%	80.96%
DeepNet	88.08%	79.69%	80.77%	77.20%	35.48%	72.71%	86.30%
	cat	**chair**	**cow**	**diningtable**	**dog**	**horse**	**motorbike**
Fisher	59.92%	51.92%	47.60%	58.06%	42.28%	80.45%	69.34%
DeepNet	81.10%	51.04%	61.10%	64.62%	76.17%	81.60%	79.33%
	person	**pottedplant**	**sheep**	**sofa**	**train**	**tvmonitor**	**mAP**
Fisher	85.10%	28.62%	49.58%	49.31%	82.71%	54.33%	59.99%
DeepNet	92.43%	49.99%	74.04%	49.48%	87.07%	67.08%	72.12%

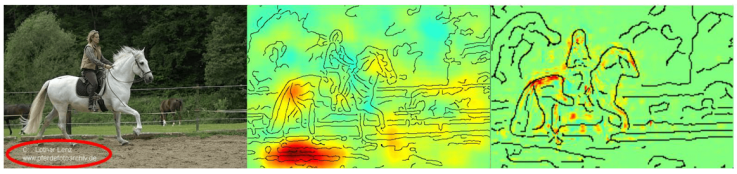

Fig. 5. Upper: the prediction accuracy of Fisher Vector and Deep Neural Network in tasks of object recognition; Lower: model diagnosis using saliency map method. (Image courtesy of Lapuschkin et al. [11])

3 Relevant Explainable AI Problems and Current Approaches

As shown in Fig. 6, there are three typical approaches to understand the behavior of a Deep Neural Network: (a) making the parts of the network transparent - the color of the neuron indicates its activation status; (b) learning semantics of the network components - a neuron could have a meaning if it is often activated by a certain part of the object; (c) generation of explanations - a human-readable textual explanation tells the underlying reason to support current decision.

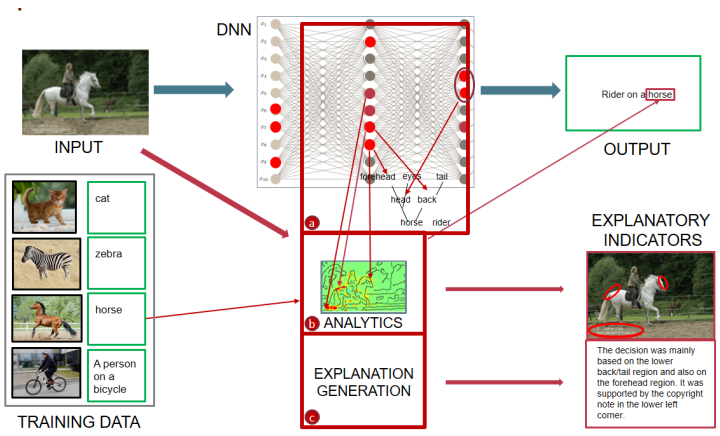

Fig. 6. Three approaches for understanding a neural network, indicated by red-boxes (a), (b) and (c)

3.1 Making the parts in DNN transparency

This section introduces two popular techniques, namely sensitivity analysis (SA) [15] [1] and layer-wise relevance propagation (LRP) [17], for explaining prediction of deep learning models.

SA explains a prediction based on the model's locally evaluated gradient.

$$R_i = \| \frac{\partial}{\partial x_i} f(x) \|. \tag{1}$$

It assumes that the most relevant input features are the most sensitive for the output. SA doesn't explain the function value $f(x)$, but rather quantifies the importance of each input variable x_i.

In contrast to SA, LRP explains predictions relative to the state of maximum uncertainty. It redistributes the prediction $f(x)$ backwards using local redistribution rules until it assigns a relevance score R_i to each input variable. The relevance score R_i of each input variable determines the variable's importance

to the prediction.

$$\sum_i R_i = \sum_j R_j = ... = \sum_k R_k = ... = f(x). \tag{2}$$

The Relevance conservation is the key property of the redistribution process. This property ensures that no relevance is artificially added or removed during redistribution. Thus, LRP truly decomposes the function values $f(x)$ in contrast to SA.

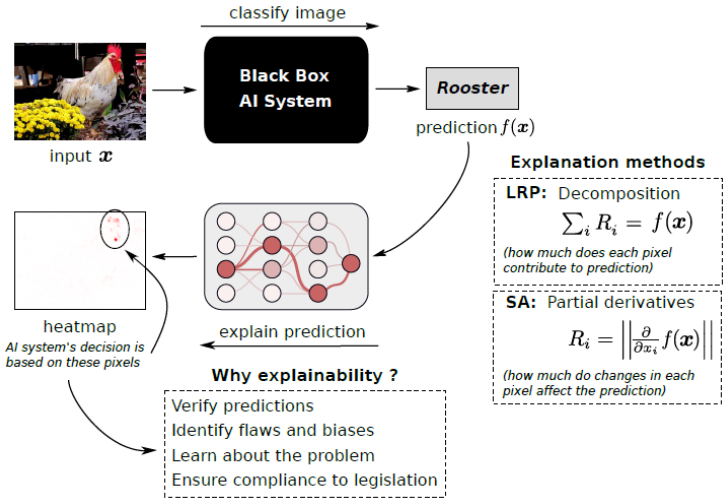

Fig. 7. Explaining predictions of an AI system using SA and LRP. (Image courtesy of W. Samek [15])

Fig. 7 summarizes the process of explanation. The AI system correctly classifies the input image as "rooster". SA indicates yellow flowers which occlude part of the rooster need to be changed to make the image look more like the predicted. However, such result would not indicate which pixels are actually pivotal for the prediction "rooster". In contrast to SA, the heatmap computed with LRP identifies pixels which are pivotal for the prediction "rooster".

Additionally, SA and LRP are evaluated on three different classification tasks, namely the annotation of images, the classification of text documents and the recognition of human actions in videos. Fig. 8(A) shows two images from the ILSVRC2012 [4] dataset, which have been correctly classified as "volcano" and "coffee cup", respectively. From the figure, we can see that SA heatmaps are much noisier than the ones computed with LRP. SA doesn't indicate how much every pixel contributes to the prediction. LRP produces better explanations than SA. Fig. 8(B) shows SA and LRP heatmaps overlaid on top of a document from the 20Newsgroup dataset. In contrast to LPR, SA methods don't distinguish between positive and negative evidence. Similarly, Fig. 8(C) shows LRP heatmaps

not only visualizes the relevant locations of the action within a video frame, but also identifies the most relevant time points within a video sequence.

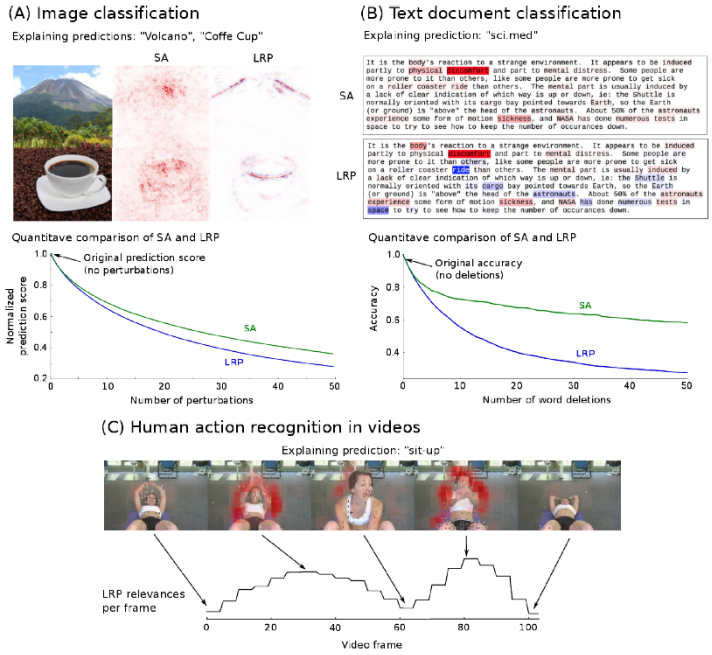

Fig. 8. Explaining prediction of three different problems using SA and LRP. (Image courtesy of W. Samek [15])

3.2 Learning semantic graphs from existing DNNs

Zhang et al. [24] proposes a method that learns a graphical model, called "explanatory graph", which reveals the knowledge hierarchy hidden inside a pre-trained Convolutional Neural Network (CNN), as shown in Fig. 9. The graph consists of multiple layers, each of them corresponds to a convolutional layer in the CNN. Each node in the graph represents a specific part of the detected object, as shown in the right side of the figure. These nodes are derived from responses of CNN filters with a disentangle algorithm. The edge connecting nodes indicates their co-activation relationship in filter response and the spatial relationship in parts location. The layer shows different granularity of the part of objects - larger parts appear in higher layers while smaller parts appear in lower layers. This work, however, adopts an explanatory graph as a bridge to understand the ordinary CNN. In later work [25], the authors introduce additional losses to force each convolutional filter in CNN to represent a specific object part directly, and produce an interpretable CNN.

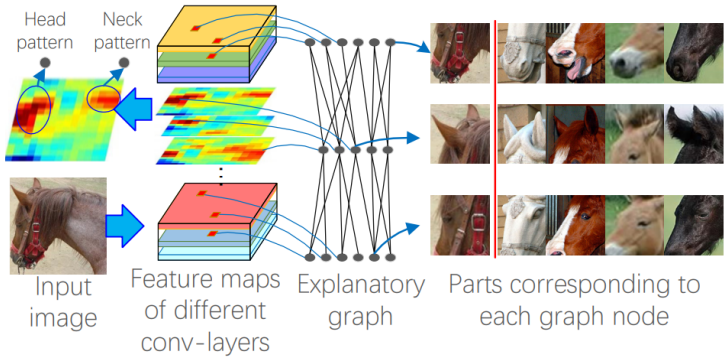

Fig. 9. An explanatory graph represents the knowledge hierarchy hidden in convolutional layers of a CNN. (Image courtesy of Zhang et al. [24])

3.3 Generation of Explanations

This section introduces a novel framework which provides visual explanations of a visual classifier [8]. Visual explanations are both image relevant and class relevant. From Fig. 10 we can find image descriptions provides a sentence based on visual information but not necessarily class relevant, while class definitions are class relevant but not necessarily image relevant. In contrast, Visual explanation such as "This is a western grebe because this bird has a long white neck, pointy yellow beak, and a red eye." includes the "red eye" property which is important to distinguish between "western grebe" and "laysan albatross". Therefore, Visual explanations are both image relevant and class relevant. It explains why the predicted category is the most appropriate for the image.

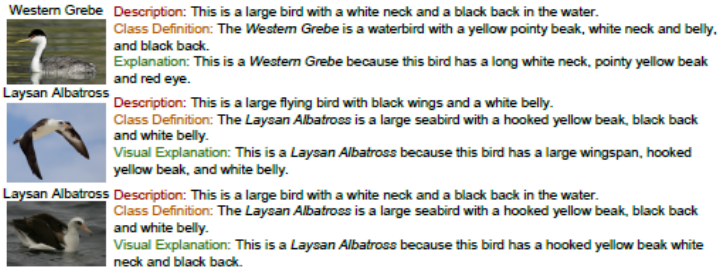

Fig. 10. Visual explanations are both image relevant and class relevant. (Image courtesy of L. A. Hendricks [9])

Fig. 11 shows the generation of explanatory text on both an image and a predicted class label. The input is run through a deep fine-grained recognition pipeline to pick out nuanced details of the image and classifying it. The features

and the label are then forwarded to the LSTM stack to produce a sequence of words.

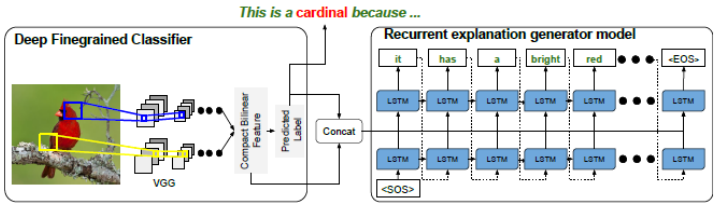

Fig. 11. Generation of explanatory text with joint classification and language model. (Image courtesy of L. A. Hendricks [9])

4 Challenges and Future Directions

The development of Explainable AI is facing both scientific and social demands. We expect AI systems could help humans make decisions in mission-critical tasks. Therefore, We need a more trustworthy and transparent AI, instead of alchemy AI [20]. Ali Rahimi, the winner of the test-of-time award in NeurIPS 2017, expressed his expectations concerning AI solutions as follows: "We are building systems that govern healthcare and mediate our civic dialogue. We would influence elections. I would like to live in a society whose systems are built on top of verifiable, rigorous, thorough knowledge, and not on alchemy. Let's take machine learning from alchemy to electricity" [14]. The term "electricity" in his speech could be replaced by "chemistry" from our perspective, meaning that AI Deep Learning AI should become part of science.

DARPA has invested 50 Million USD and launched a 5-year research program on Explainable AI (XAI) [21], aiming to produce "glass-box" models that are explainable to a "human-in-the-loop", without greatly sacrificing AI performance, as shown in Fig. 12. Human users should be able to understand the AI's cognition both in real-time and after the results achieved, and furthermore might be able to determine when to trust the AI and when the AI should be distrusted. In Phase 1, it is planned to achieve initial implementations of their explainable learning systems. In Phase 2, it is to build a toolkit library consisting of machine learning and human-computer interface software modules that could be utilized for developing future explainable AI systems.

It is known that humans can acquire and use both explicit knowledge and implicit knowledge. Moreover, humans can combine the two forms of knowledge to a certain degree. For humans, understanding and explaining require explicit knowledge. However, DNNs acquire and use implicit knowledge in the form of probabilistic models. As they stand, they cannot understand anything. Other AI methods model explicit knowledge, such as Knowledge Graphs. Today the two worlds in AI technology are still largely separated. Researchers are now

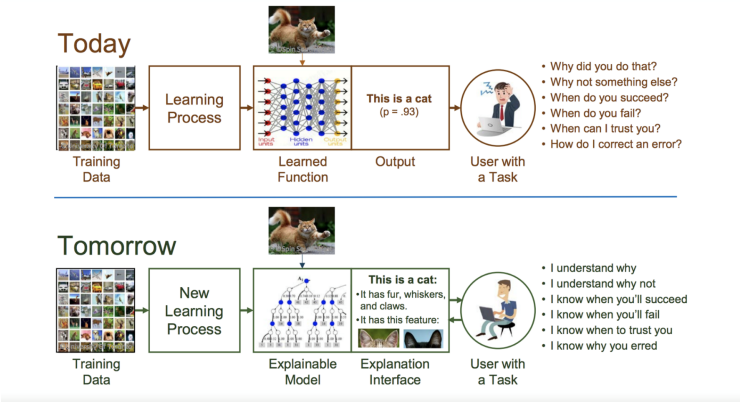

Fig. 12. Explainable AI (XAI) Concept presented by DARPA. (Image courtesy of DARPA XAI Program [21])

strengthening their efforts to bring the two worlds together. The need-driven research on Explainable AI is a source and a catalyst for the work dedicated to this grand challenge.

References

1. Baehrens, D., Schroeter, T., Harmeling, S., Kawanabe, M., Hansen, K., Müller, K.: How to explain individual classification decisions. J. Mach. Learn. Res. **11**, 1803–1831 (2010), http://portal.acm.org/citation.cfm?id=1859912
2. Cameron, L.: Houston Schools Must Face Teacher Evaluation Lawsuit (2017), https://www.courthousenews.com/houston-schools-must-face-teacher-evaluation-lawsuit/
3. Caruana, R., Lou, Y., Gehrke, J., Koch, P., Sturm, M., Elhadad, N.: Intelligible models for healthcare: Predicting pneumonia risk and hospital 30-day readmission. In: Proceedings of the 21th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. pp. 1721–1730. ACM (2015)
4. Deng, J., Dong, W., Socher, R., Li, L., Li, K., Li, F.: Imagenet: A large-scale hierarchical image database. In: 2009 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR 2009), 20-25 June 2009, Miami, Florida, USA. pp. 248–255 (2009). https://doi.org/10.1109/CVPRW.2009.5206848, https://doi.org/10.1109/CVPRW.2009.5206848
5. Doilovi, F.K., Bri, M., Hlupi, N.: Explainable artificial intelligence: A survey. In: 2018 41st International Convention on Information and Communication Technology, Electronics and Microelectronics (MIPRO). pp. 0210–0215 (May 2018). https://doi.org/10.23919/MIPRO.2018.8400040
6. Goodman, B., Flaxman, S.: European union regulations on algorithmic decision-making and a "right to explanation". AI magazine **38**(3), 50–57 (2017)
7. Guidotti, R., Monreale, A., Ruggieri, S., Turini, F., Giannotti, F., Pedreschi, D.: A survey of methods for explaining black box models. ACM Computing Surveys (CSUR) **51** (2019)

8. Hendricks, L.A., Akata, Z., Rohrbach, M., Donahue, J., Schiele, B., Darrell, T.: Generating visual explanations. In: European Conference on Computer Vision. pp. 3–19. Springer (2016)
9. Hendricks, L.A., Akata, Z., Rohrbach, M., Donahue, J., Schiele, B., Darrell, T.: Generating visual explanations. In: European Conference on Computer Vision. pp. 3–19. Springer (2016)
10. Jeremy, J.: Decision trees (2017), https://www.jeremyjordan.me/decision-trees/
11. Lapuschkin, S., Binder, A., Montavon, G., Muller, K.R., Samek, W.: Analyzing classifiers: Fisher vectors and deep neural networks. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. pp. 2912–2920 (2016)
12. Lapuschkin, S., Binder, A., Montavon, G., Müller, K.R., Samek, W.: The lrp toolbox for artificial neural networks. Journal of Machine Learning Research **17**(114), 1–5 (2016), http://jmlr.org/papers/v17/15-618.html
13. Lipton, Z.C.: The mythos of model interpretability. ACM Queue - Machine Learning **16** (2018)
14. Rahimi, A.: NIPS 2017 Test-of-Time Award presentation (2017), https://www.youtube.com/watch?v=ORHFOnaEzPc
15. Samek, W., Wiegand, T., Müller, K.R.: Explainable artificial intelligence: Understanding, visualizing and interpreting deep learning models. arXiv preprint arXiv:1708.08296 (2017)
16. Scott, A.C., Clancey, W.J., Davis, R., Shortliffe, E.H.: Explanation capabilities of production-based consultation systems. American Journal of Computational Linguistics **62** (1977)
17. Simonyan, K., Vedaldi, A., Zisserman, A.: Deep inside convolutional networks: Visualising image classification models and saliency maps. In: 2nd International Conference on Learning Representations, ICLR 2014, Banff, AB, Canada, April 14-16, 2014, Workshop Track Proceedings (2014), http://arxiv.org/abs/1312.6034
18. State Council Chinese Government: Development Plan for New Generation Artificial Intelligence (2017), http://www.gov.cn/zhengce/content/2017-07/20/content_5211996.htm
19. Swartout, W.R.: Explaining and justifying expert consulting programs. In: Proceedings of the 7th International Joint Conference on Artificial Intelligence (1981)
20. Tony, P.: LeCun vs Rahimi: Has Machine Learning Become Alchemy? (2017), https://medium.com/@Synced/lecun-vs-rahimi-has-machine-learning-become-alchemy-21cb1557920d
21. Turek, M.: DARPA - Explainable Artificial Intelligence (XAI) Program (2017), https://www.darpa.mil/program/explainable-artificial-intelligence
22. Van Veen, F.: The Neural Network Zoo (2016), http://www.asimovinstitute.org/neural-network-zoo/
23. Wiktionary: Explain (2019), https://en.wiktionary.org/wiki/explain
24. Zhang, Q., Cao, R., Shi, F., Wu, Y.N., Zhu, S.C.: Interpreting cnn knowledge via an explanatory graph. In: Thirty-Second AAAI Conference on Artificial Intelligence (2018)
25. Zhang, Q., Wu, Y.N., Zhu, S.C.: Interpretable convolutional neural networks. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. pp. 8827–8836 (2018)