Chapter

Explainable Artificial Intelligence (xAI) Approaches and Deep Meta-Learning Models

Evren Dağlarli

Abstract

The explainable artificial intelligence (xAI) is one of the interesting issues that has emerged recently. Many researchers are trying to deal with the subject with different dimensions and interesting results that have come out. However, we are still at the beginning of the way to understand these types of models. The forthcoming years are expected to be years in which the openness of deep learning models is discussed. In classical artificial intelligence approaches, we frequently encounter deep learning methods available today. These deep learning methods can yield highly effective results according to the data set size, data set quality, the methods used in feature extraction, the hyper parameter set used in deep learning models, the activation functions, and the optimization algorithms. However, there are important shortcomings that current deep learning models are currently inadequate. These artificial neural network-based models are black box models that generalize the data transmitted to it and learn from the data. Therefore, the relational link between input and output is not observable. This is an important open point in artificial neural networks and deep learning models. For these reasons, it is necessary to make serious efforts on the explainability and interpretability of black box models.

Keywords: explainable artificial intelligence (xAI), meta-learning, deep learning

1. Introduction

Explainable artificial intelligence (xAI) is one of the research topics that has been intriguing in recent years. Today, even if we are at the beginning of understanding this type of models, the studies that show interesting results about this issue are getting more and more intensive. In the near future, it is predicted that there will be years when the interpretability of artificial intelligence and deep meta-learning models is frequently explored [1]. It is thought to be a solution to overcome constraints in classical deep learning methods.

In classical artificial intelligence approaches, we frequently encounter deep learning methods available today. Currently, in classical deep learning methods, input data and target (class) information can be trained with high performance and tested with new data input [2]. These deep learning methods can yield highly effective results according to the data set size, data set quality, the methods used in feature extraction, the hyper parameter set used in deep learning models, the activation functions, and the optimization algorithms [3]. Many layers in a deep

network allow it to recognize things at different levels of abstraction. For example, in a structure designed to recognize dogs, the lower layers recognize simple things such as outlines or color; the upper layers recognize more complex things like fur or eyes, and the upper layers define them all as a dog. Presumably speaking, the same approach can be applied to other inputs that lead a machine to teach itself. For example, it can be easily applied to the sounds that make up the words in the speech, the letters and words that form the sentences in the text, or the steering movements required to drive.

However, there are important shortcomings that current deep learning models are currently inadequate [4]. For deep learning, huge data sets are needed to train on, and these data sets must be inclusive/unbiased, and of good quality [5]. In addition, traditional deep learning requires a lot of time to train models for satisfying their purpose with an admissible amount of accuracy and relevancy [6]. Although deep learning is autonomous, it is highly susceptible to errors. Assume that an algorithm is trained with data sets small enough to not be inclusive [4]. The models trained by this way cause to irrelevant responses (biased predictions coming from a biased training set) being displayed to users [7]. One of the most important problems in artificial learning models is transparency and interpretability [8]. These artificial neural network-based models are black box models that generalize the data transmitted to it and learn from the data. Therefore, the relational link between input and output is not observable [9]. In other words, when you receive an output data against the input data, the deep learning model cannot provide the information for which reason the output is generated. The user cannot fully grasp the internal functions of these models and cannot find answers to question why and how the answers the models produce [10]. This situation creates difficulties in the application areas of these models in many aspects. For example, you stopped a taxi and got on it. The driver is such a driver that when he takes you to your destination, he turns right, turns left, and tries to get you on a strange route than you expect, but when you ask why he did so, he cannot give you a satisfactory answer. Would you be nervous? If there is no problem for you, you can ride an autonomous vehicle without a driver. As another example, when you go to the doctor, the doctor you send your complaint asks for tests and when you have those tests and send it to the doctor, the doctor tells you what your illness is. Even though he says his treatment, he does not give explanatory information about the cause of your illness. In this case, questions remain about what caused the disease and you would not be satisfied with the doctor. This is an important open point in artificial neural networks and deep learning models.

The explainable artificial intelligence (xAI) approach can be considered as an area at the intersection of several areas. One of these areas is the end user explanation section that includes social sciences. This area provides artificial intelligence to gain cognitive abilities. Another area is the human machine interface, where it can demonstrate the ability to explain; because explainable artificial intelligence needs a very high-level interaction with the user. And finally, deep learning models are an important part of an explicable artificial intelligence approach (**Figure 1**).

In this new approach, it is aimed to provide the user with the ability to explain the output data produced as well as being trained at high performance with the input data and target (class) information and tested with the new data input as in the classical machine learning models. This will create a new generation artificial intelligence approach that can establish a cause and effect relationship between input and output. It will also be the mechanism of monitoring the reliability of artificial intelligence from the user point of view. While a classic deep learning model can answer "what" or "who" questions, learning models in explainable artificial intelligence approaches can also answer "why," "how," "where," and "when" questions [10] (**Figure 2**).

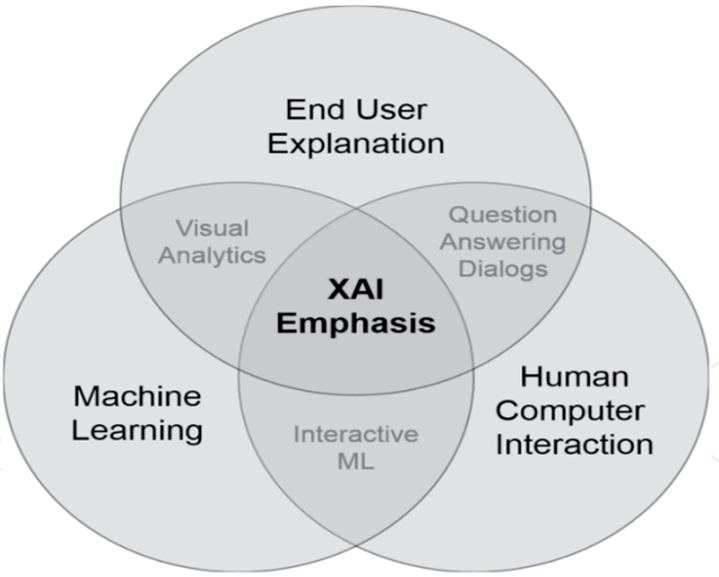

Figure 1.
Explainable artificial intelligence (xAI) [8].

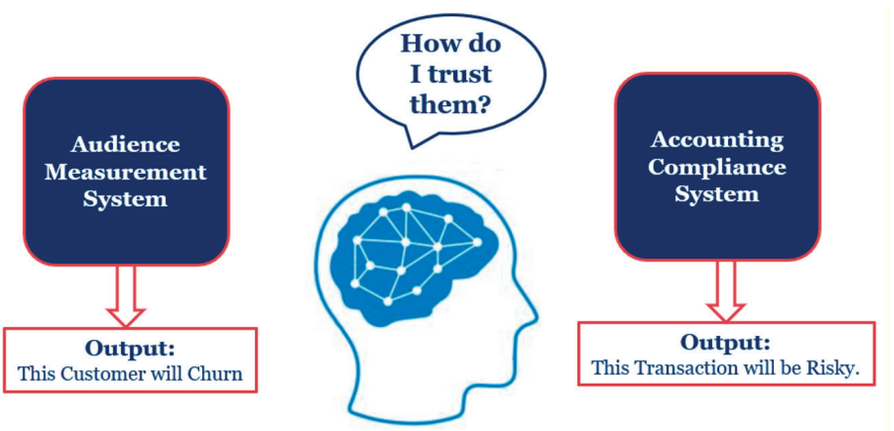

Figure 2.
How can explainable artificial intelligence (xAI) be reliable [11]?

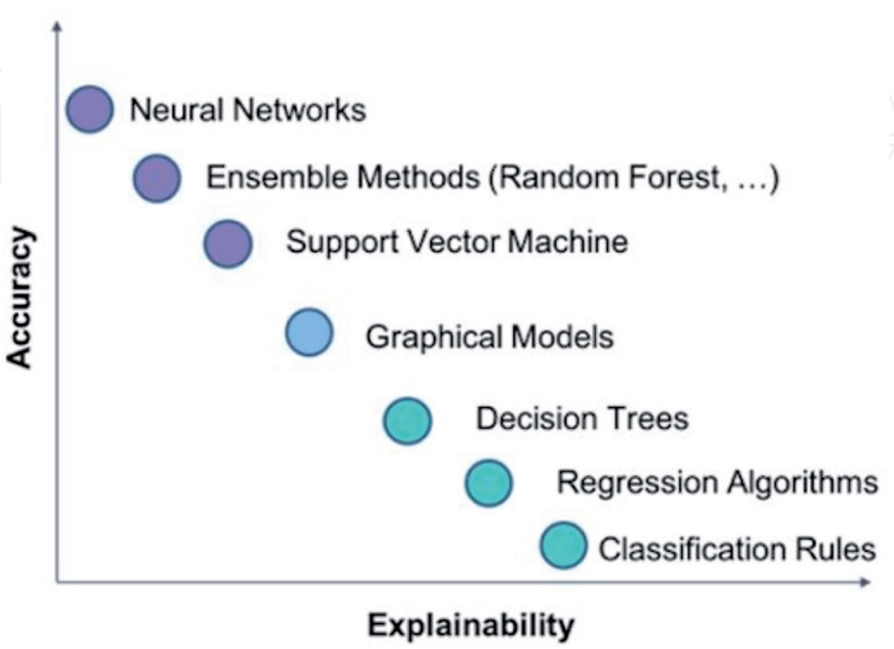

Figure 3.
Machine learning models with respect to accuracy-explainability domain [12].

Explainability and accuracy are two separate domains. In general, models that are advantageous in terms of accuracy and performance are not very successful in terms of explainability. Likewise, methods with high explainability are also disadvantageous in terms of accuracy. When methods such as classical deep learning models, artificial neural networks support vector machines are utilized, they do not give reasons why, and how their outputs created in terms of explainability. On the other hand, they are very successful in accuracy and performance. Rule-based structures, decision trees, regression algorithms, and graphical methods are good explainability but not advantageous in terms of performance and accuracy. At this point, explanatory artificial intelligence (xAI), which is targeted to be at the highest level of both explainability and accuracy and performance, reveals its importance at this point (**Figure 3**).

2. Related works

There is a transformation of machine learning that has been going on since the 1950s, sometimes faster and sometimes slower. The most studied and remarkable area in the recent past is artificial learning, which aims to model the live decision system, behavior, and responses. Successful results in the field of artificial learning led to the rapid increase of AI applications. Further studies promise to be autonomous systems capable of self-perception, learning, decision-making, and action [13].

Especially after the 1990s, although deep learning concept and foundations go back to the past, the accompanying recurrent neural networks, convolutional neural networks, deep reinforcement learning, and adversarial generative networks have achieved remarkable successes. Although successful results are obtained, these systems are insufficient in terms of explaining the decisions and actions to human users and there are limits.

The U.S. Department of Defense (DoD) explains that it is facing the challenges posed by autonomous and symbiotic systems, which are becoming smarter with each passing day. Explaining artificial intelligence or especially explanatory machine learning is important in terms of being a preview that users will encounter machines with human-like artificial intelligence in the future [14, 15]. Explained artificial intelligence is one of the Defense Advanced Research Projects Agency (DARPA) programs aimed at the development of a new generation of artificial intelligence systems, where they understand the context and environment in which machines operate and build descriptive models that enable them to characterize the real world phenomenon over time. For this purpose, DARPA recently issued a call letter for the Explainable Artificial Intelligence (XAI)—Explanatory Artificial Intelligence project [15]. Within the scope of the project, it is aimed to develop a system of machine learning techniques that focus on machine learning and human-machine interaction, and produce explanatory models that will enable end users to understand, trust, and manage emerging artificial intelligence systems. According to the researchers from DARPA, the striking successes in machine learning have led to a huge explosion in new AI capabilities that enable the production of autonomous systems that perceive, learn, decide, and act on their own. Although these systems provide tremendous benefits, their effectiveness is limited due to the inability to explain machine decisions and actions to human users.

The Explanatory Artificial Intelligence project aims to develop the machine learning and computer-human interaction tools to ensure that the end user, who depends on decisions, recommendations, or actions produced by the artificial intelligence system, understands the reason behind the system's decisions [1]. For example, an intelligence analyst who gets recommendations from big data

analytics algorithms may need to understand why the algorithm advises to examine a particular activity further. Similarly, the operator, who tests a newly developed autonomous system, has to understand how he makes his own decisions to determine how the system will use it in future tasks.

The xAI tools will provide end users with explanations of individual decisions, which will enable them to understand the strengths and weaknesses of the system in general, give an idea of how the system will behave in the future, and perhaps teach how to correct the system's mistakes. The XAI project addresses three research and development challenges: how to build more models, how to design an explanation interface, and how to understand psychological requirements for effective explanations [2].

For the first problem, the xAI project aims to develop machine learning techniques to be able to manufacture explanatory models. To solve the second challenge, the program envisions integrating state-of-the-art human-machine interaction techniques with new principles, strategies, and techniques to produce effective explanations. To solve the third problem, the xAI project plans to summarize, disseminate, and apply existing psychological theory explanations. There are two technical areas in the program: the first is to develop an explanatory learning system with an explanatory model and an explanation interface; and the second technical area covers psychological theories of explanation [8].

In 2016, a self-driving car was launched on quiet roads in Monmouth County, New Jersey. This experimental tool developed by researchers at chip maker Nvidia did not look different from other autonomous cars; however, Google was different from what Tesla or General Motors introduced and showed the rising power of artificial intelligence. The car had not even followed a single instruction provided by an engineer or a programmer. Instead, it relied entirely on an algorithm that allowed him to learn to drive by watching a person driving [3]. It was an impressive success to have a car self-driving in this way. But it was also somewhat upsetting as it was not entirely clear how the car made its own decisions. The information from the vehicle's sensors went directly to a huge artificial neural network that processes the data and then delivers the commands needed to operate the steering wheel, brakes, and other structures. The results seem to match the reactions you can expect from a human driver. But what if one day something unexpected happens; hits a tree or stops at the green light? According to the current situation, it may be difficult to find the cause. The system is so complex that even the engineers who designed it can find it difficult to pinpoint the cause of any action. Moreover, you cannot ask this; there is no obvious way to design such a system that can always explain why it does what it does. The mysterious mind of this vehicle points to a vague-looking issue of artificial intelligence. Artificial intelligence technology, which is located at the base of the car and known as deep learning, has proven to be very strong in problem-solving in recent years, and this technology has been widely applied in works such as image content estimation, voice recognition, and language translation. Now the same methods can be used to diagnose lethal diseases, make million-dollar business decisions, etc. to change all industries.

Currently, the mathematical models are used to help determine who will be on parole, who will be approved to borrow money, and who will be hired. If you can access these mathematical models, it is possible to understand their reasoning. But banks, the military, employers, and others are now turning their attention to more complex machine learning approaches. These approaches can make automated decision-making completely incomprehensible. The most common of these approaches represents deep learning, a fundamentally different way of programming computers. Whether it is an investment decision or a medical decision, or a military decision, you do not want to rely solely on a "black box" method [1]. There is

already a debate that it is a fundamental legal right to question a system of artificial intelligence about how it arrived at its conclusions. Starting in the summer of 2018, the European Union may require companies to provide users with an explanation of the decisions made by automated systems. This may be impossible even for systems that look comparatively simple on the surface, such as applications and Websites that use deep learning to offer advertising or song suggestions. Computers performing these services have programmed themselves and have done so in ways we cannot understand. Even the engineers who build these applications cannot fully explain their behavior.

As technology advances, we can go beyond some thresholds where using artificial intelligence in recent times requires a leap of faith. The mankind, of course, are not always able to fully explain our thought processes; but we find a variety of methods to intuitively trust people and measure them. Will this be possible for machines that think and make decisions differently than a person does? We have never built machines that operate in ways that their manufacturers do not understand. How long can we hope to communicate and deal with intelligent machines that can be unpredictable or incomprehensible? These questions take a journey toward new technology research on artificial intelligence algorithms, from Google to Apple and many other places between them, including a conversation with one of the greatest thinkers of our time.

3. Explainable artificial intelligence (xAI)

You cannot see how the deep neural network works just by looking inside. The reasoning of a network is embedded in the behavior of thousands of nerves, which are stacked and tied to tens or even hundreds of layers, mixed together. Each of the nerves in the first layer receives an input, such as the voltage of a pixel in an image, and then performs a calculation before sending a new signal as an output. This output is sent to the next layer in a complex network, and this process continues until a general output is produced. There is also a process known as back propagation that modifies the calculations of individual nerves so that a network learns to produce a desired output. Because deep learning is inherently a dark black box by nature, artificial learning models designed with millions of artificial nerve cells with hundreds of layers like traditional deep learning models are not infallible [1]. Their reliability is questioned when simple pixel changes can be seriously misleaded by causing significant deviations in the weight values in all layers of the neural network, especially in an example such as a one-pixel attack [16]. So, it becomes inevitable to ask the question of how it can succeed or fail. With the success of this type of advanced applications, its complexity also increases and its understanding/clarity becomes difficult.

It is aimed to have the ability to explain the reasons of new artificial learning systems, identify their strengths and weaknesses, and understand how they will behave in the future. For an ideal artificial intelligence system, the best accuracy and best performance, as well as the best explainability and the best interpretability are required within the cause-effect relationship. The strategy developed to achieve this goal is to develop new or modified artificial learning techniques that will produce more explicable models. These models are aimed to be combined with state-of-the-art human-computer interactive interface techniques that can be translated into understandable and useful explanation dialogs for the end user (**Figure 4**).

In this structure, unlike the classical deep learning approaches, two different elements draw attention as well as a new machine learning process. One of these

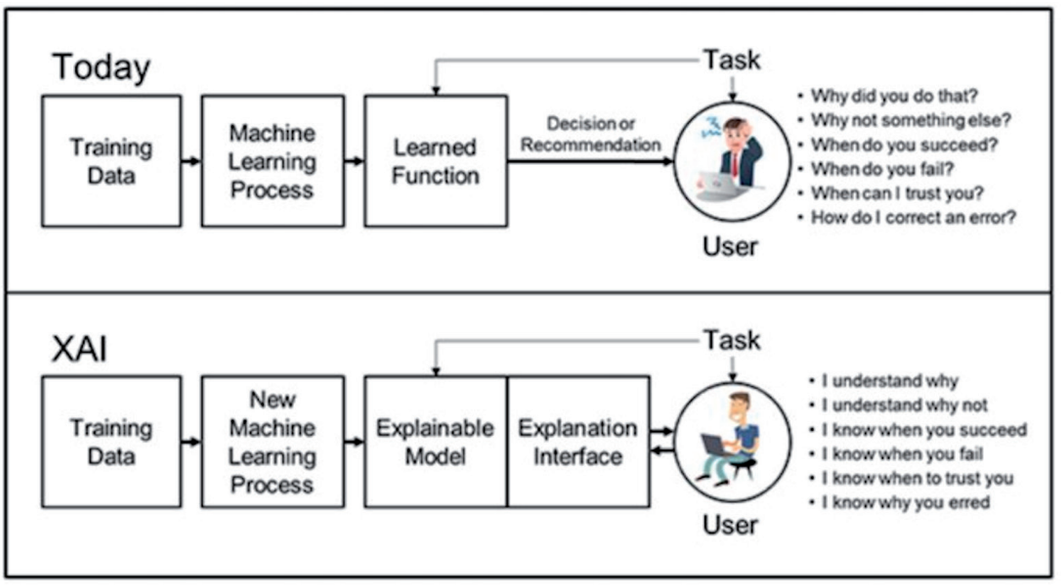

Figure 4.
Explainable artificial intelligence (xAI) project proposed by DARPA [14, 15].

is the explanatory model and the other is the explanation interface. The process of deep neural network-based machine learning is explained at the core of the artificial intelligence approach. Among the known deep learning models, autoencoder, convolutional, recurrent (LSTM), deep belief network, or deep reinforcement learning can be preferred. However, it is also possible to use a hybrid structure where several deep learning approaches are used together. Autoencoder-type model of deep neural networks are multilayered perceptron structure. In convolution neural network-type models, layers consist of convolutional layer, ReLU activation function, and max pool layer. A conventional component of the LSTM is composed of a memory cell including input, output, and forget gates. For training, the back-propagation through time algorithm can be preferred. Although the most common form of deep reinforcement learning models is deep Q network (DQN), many different variations of this model can be addressed. Many different algorithms are used as optimization algorithm. Gradient-based algorithms are the most common form of these algorithms (**Figure 5**).

Explainable model is an adaptive rule-based reasoning system. It is a structure that reveals the cause-effect relations between input data and the results obtained from the machine learning process. This causal structure learns the rules with its own internal deep learning method. In this way, the explanatory artificial intelligence model allows it to explore the causes and develop new strategies against different situations [20].

The explanation interface is a part of the user interaction. It is similar to the question-answer interface in voice digital assistants. This interface consists of a decoder that evaluates the demands of the user and an encoder unit that enables the responses from the explanatory model, which constitutes the causal mechanism of the explainable artificial intelligence, to the user (**Figure 6**).

In fact, the large networks of semantic technologies (entities) and relationships associated with Knowledge Graphs (KGs) provide a useful solution for the issue of understandability, several reasoning mechanisms, ranging from consistency checking to causal inference [21]. The ontologies realizing these reasoning procedures provide a formal representation of semantic entities and relationships relevant to a particular sphere of knowledge [21]. The input data, hidden layers, encoded features, and predicted output of deep learning models are passed into knowledge graphs (KGs) or concepts and relationships of ontologies (knowledge

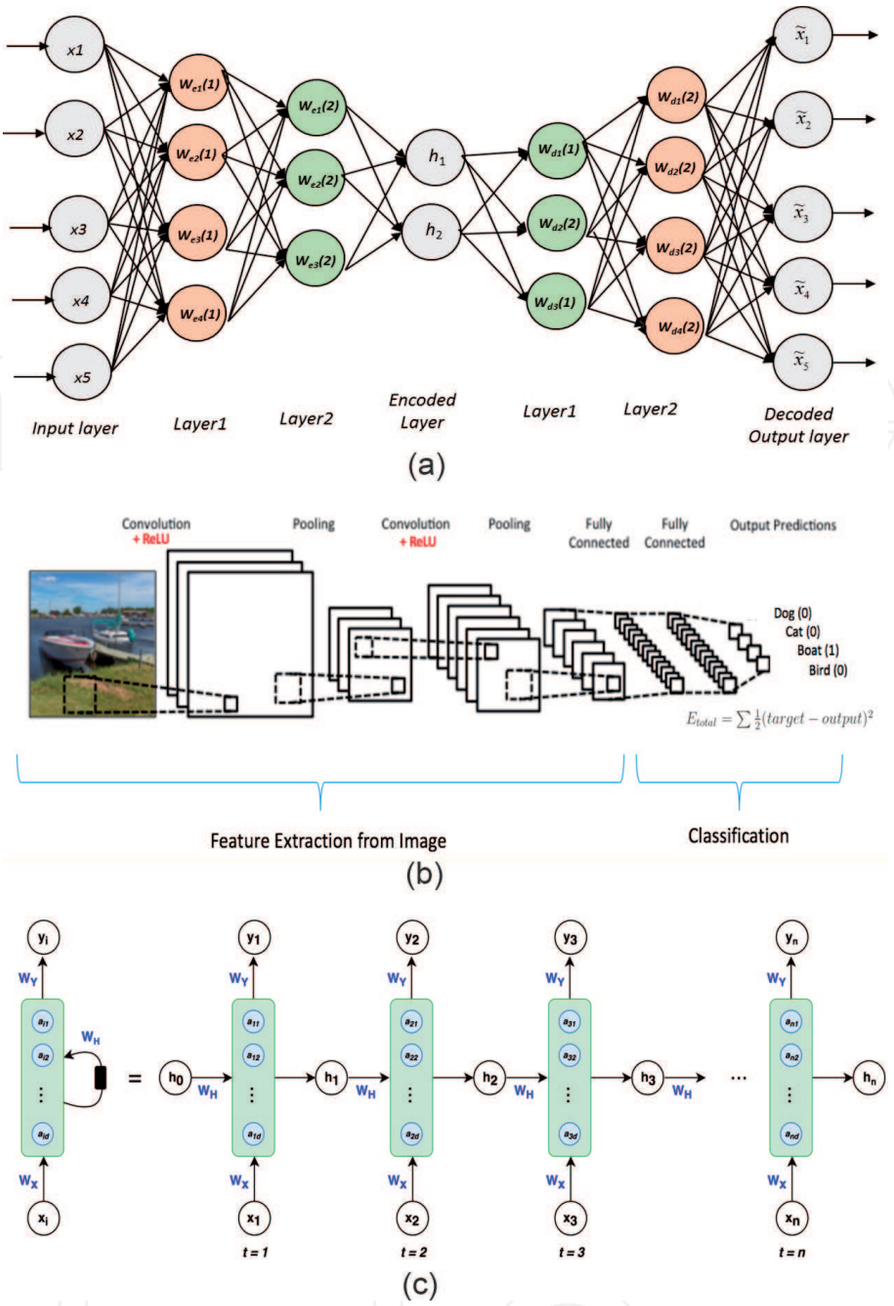

Figure 5.
Deep learning models: (a) autoencoder [17], (b) convolutional neural network [18], and (c) recurrent (LSTM) neural network [19].

matching) [21]. Generally, the internal functioning of algorithms to be more transparent and comprehensible can be realized by knowledge matching of deep learning components, including input features, hidden unit and layers, and output predictions with KGs and ontology components [21]. Besides that, the conditions for advanced explanations, cross-disciplinary and interactive explanations are enabled by query and reasoning mechanisms of KGs and ontologies [21].

Although explanatory artificial intelligence forms are of very different structures, all modules such as this explanation interface, explanatory model, and deep learning work in coordination with each other. For example, while a deep learning process estimates classes, such as the explanatory artificial intelligence model (xAI tool) developed by IBM, the concept features data obtained from this process, and another deep learning process using the same input data set produces an explanatory output for the predicted class label output [22] (**Figure 7**).

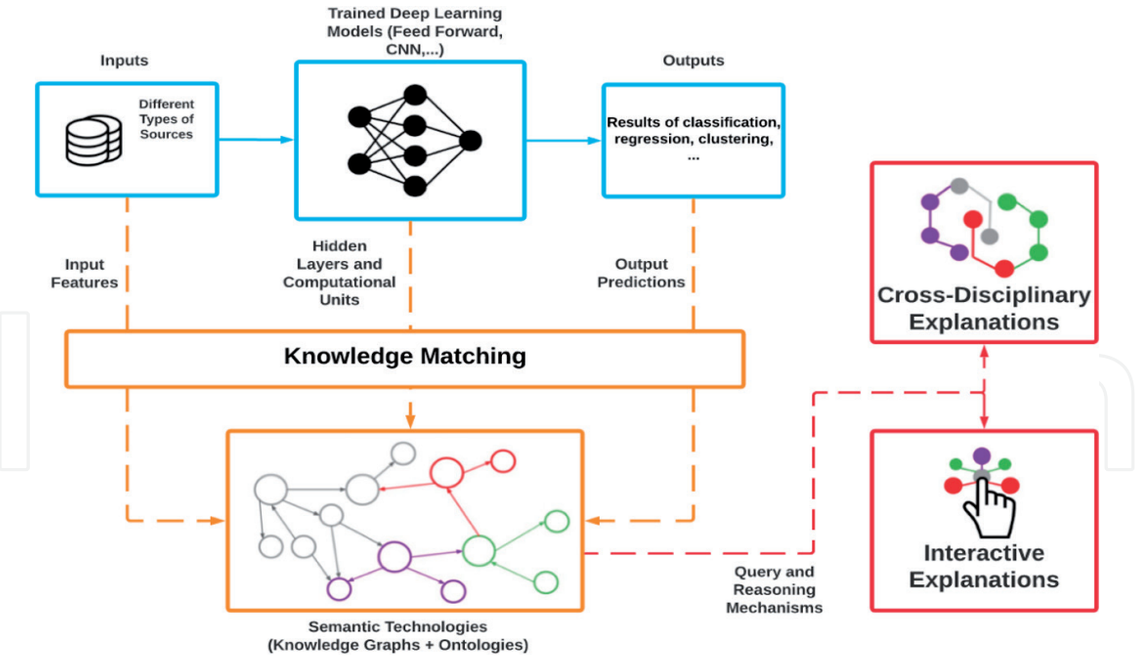

Figure 6.
Semantic knowledge matching for explainable artificial intelligence model [21].

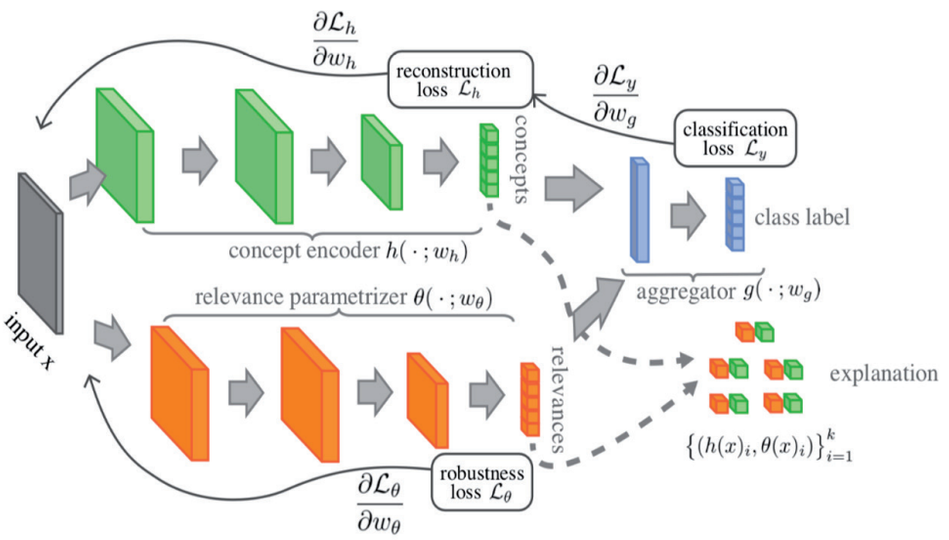

Figure 7.
Explainable artificial intelligence (xAI) tool developed by IBM [22].

At this point, the explainable artificial intelligence (xAI) tool developed by IBM is referred as a self-explaining neural network (SENN) which can be trained end-to-end with back-propagation in case of that g depends on its arguments in a continuous way [18]. The input is transformed into a small set of interpretable basis features by a concept encoder [22]. The relevance scores are produced by an input-dependent parametrizer. A prediction to be generated is merged by an aggregation function. The full model to behave locally as a linear function on h(x) with parameters θ(x), producing interpretation of both concepts and relevances, is induced by the robustness loss on the parametrizer [22]. θ(x) modeling capacity is important so that the model richness realizing higher-capacity architectures is sustained although the concepts are chosen to be raw inputs (i.e., h is the identity).

4. Meta-learning

As research and technology on machine learning progresses, artificial intelligence agents consistently display impressive learning performances that meet and exceed the cognitive skills of people in different fields. However, most AI programs are based on computing technology and even reinforcement learning (RL) models that try to regularly improve their knowledge to match human performance. By contrast, people can quickly learn new skills of new skills, simply by having a new skill [23]. The learning of the human brain so efficiently has surprised neuroscientists for years.

In traditional deep learning approaches, the system develops a data-specific model that is transmitted to it by learning from the data. The learning system will perform a certain task only for a certain environment. In the case of another environment, when a very different data is transmitted to it, this deep learning model will be insufficient to perform the task [24]. This issue reveals hard constraints in utilizing machine learning or data mining methods, since the relationship between the learning problem and the effectiveness of different learning algorithms is not yet understood. Under ideal conditions, a system should be designed in which the quality of the data given to the system differs and it can easily adapt to changes in different environments [25]. The deep learning methods used in the current situation are not successful in these situations. At this point, meta-learning, which learns to learn, is an integrated and hierarchical learning model over several different environmental models [26, 27]. As a subfield of machine learning, meta-learning learning algorithms are applied on metadata about machine learning experiments. Instead of classical machine learning approaches that only learn a specific task with single massive dataset, meta-learning is a high-level machine learning approach that learns other tasks together. Therefore, this approach requires a hierarchical structure that learns to learn a new task with distributed hierarchically structured metadata. It is generally applied for hyper parameter adjustment; recent applications have started to focus on a small number of learning. For example, if the system has already learned a few different models or tasks, meta-learning can generalize them and learn how to learn more efficiently. In this way, it can learn new tasks efficiently and create a structure that can easily adapt to changes in multiple tasks in different environments.

People are good at figuring out the meaning of a word after seeing it used only in a few sentences. Similarly, we want our ML algorithms to be generalized to new tasks, without the need for a large data set each time, and to change behavior after a few samples. In typical learning (on a single dataset), each sample targets pair functions as a training point. However, in a small number of learning situations, each "new" sample area is actually another task in itself. In other words, understanding the way that you use unique words in a particular social environment becomes a new task for your language-understanding model, and when you enter a different social environment, it means that the system can adapt to a different language-understanding model than before since it requires to dominate the words that are specific to that social environment. To make sure an ML framework can behave similarly, we have to train it on multiple tasks on its own, so we make each data set a new example of training [28] (**Figure 8**).

An alternative is to handle the task consecutively as a sequential input array and create a repetitive model that can create a representation of this array for a new task. Typically, in this case, we have a single training process with a memory or attention repetitive network [30]. This approach also gives good results, especially when the installations are properly designed for the task. The calculation performed by the

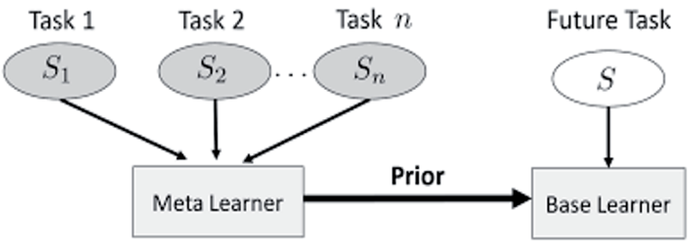

Figure 8.
Meta-learning approach [29].

optimizer during the meta-forward transition is very similar to the calculation of a repetitive network [31]. It repeatedly applies the same parameters over a series of inputs (consecutive weights and gradients of the model during learning). In practice, this means that we meet a common problem with repetitive networks. Since the models are not trained to get rid of training errors, they have trouble returning to a safe path when they make mistakes, and the models have difficulty generalizing longer sequences than those used in the order in which they were used. In order to overcome these problems, if the model learns an action policy related to the current educational situation, reinforcement learning approaches can be preferred [32] (**Figure 9**).

Formal reinforcement learning algorithm learns a policy for only single task.

$$\theta^* = \text{argmax}_\theta E_{\pi\theta(\tau)}(R(\tau)) \tag{1}$$

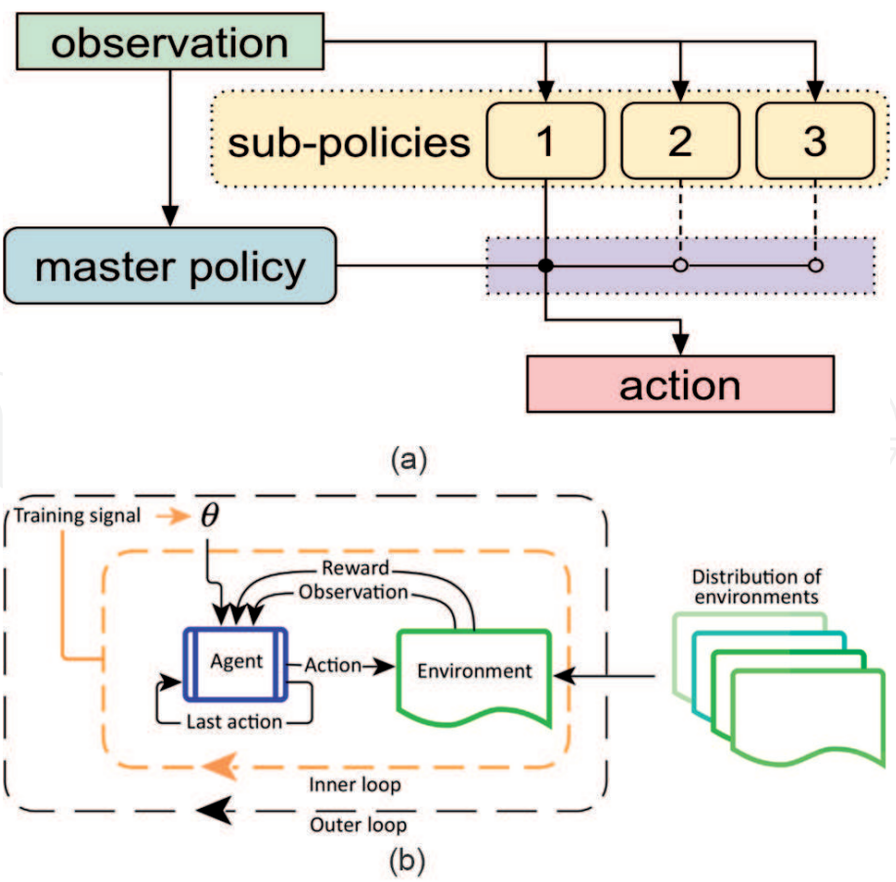

Figure 9.
(a) Meta-reinforcement learning (stack of sub-policies representation) [33] and (b) meta-reinforcement learning (inner-outer loop representation) [34].

In meta-reinforcement learning, there are two distinct processes. One of them is adaptation (inner-loop) behaving ordinary RL policy learning to produce sub-policy where $\phi_i = f_\theta(\mathcal{M}_i)$ for each environment (task) $\mathcal{M}_i$.

$$\theta^* = \text{argmax}_\theta \sum_{i=1}^{n} E_{\pi\phi_i(\tau)}[R(\tau)] \quad (2)$$

Another process is meta-training (outer-loop), which is described as meta-policy learning from all sub-policies in the adaptation process (inner-loop).

One of the main differentiers between the human brain and artificial intelligence structures such as deep neural networks, is the brain that utilizes different chemicals known as neurotransmitters to perform different cognitive functions. A new study by DeepMind believes that one of these neurotransmitters plays an important role in the brain's ability to quickly learn new topics. Dopamine acts as a reward system that strengthens connections between neurons in the brain.

The DeepMind team has used different meta-reinforcement learning techniques that simulate the role of dopamine in the learning process. Meta-learning trained a repetitive neural network (representing the prefrontal cortex) using standard deep reinforcement learning techniques (representing the role of dopamine) and then compared the activity dynamics of the repetitive network with actual data from previous findings in neuroscience experiments [27]. Recurrent networks are a good example of meta-learning because they can internalize past actions and observations and then use these experiences while training on various tasks.

The meta-learning model recreated the Harlow experiment by saying a virtual computer screen and randomly selected images, and the experiment showed that the "meta-RL agent" was learned in a similar way to the animals found in the Harlow Experiment, even when presented with the Harlow Experiment. All new images were never seen before. The meta-learning agent quickly adapted to different tasks with different rules and structures.

5. Explainable meta-reinforcement learning (xMRL)

In this section, we will discuss the development of deep reinforcement learning models with an explicable approach to artificial intelligence. Deep reinforcement learning models are machine learning models that learn what action to take according to status and reward information by maximizing reward [27]. Generally, it is widely preferred in robotic, autonomous driverless vehicles, unmanned aerial vehicles, and games. Explanatory artificial intelligence, on the other hand, provides the knowledge of why action should be taken against the situation and reward for deep reinforcement learning models. In this way, it will be possible to gain the causal decision-making ability of the model by revealing the relational links between the input and output of the developed agent (**Figure 10**).

In addition, it is possible to learn the reward derivation mechanism by using the inverse reinforcement learning model [36, 37]. In this case, unlike the previous approach, a meta-cognitive artificial intelligence model that can adapt to other environments instead of just one environment is developed [38, 39]. Taken together with the explainable artificial intelligence approach, it will be possible for the developed agent to develop his own strategy by establishing a cause-effect relationship. For example, the explainable meta-reinforcement learning agent to be developed means that in terms of meta-learning, it can learn to play Go, chess, checkers, and even learn and adapt when it is encountering a new game, and in terms of explainable artificial intelligence, it means that being aware of why it is doing any specific action against a move made by the opponent, it can explain this.

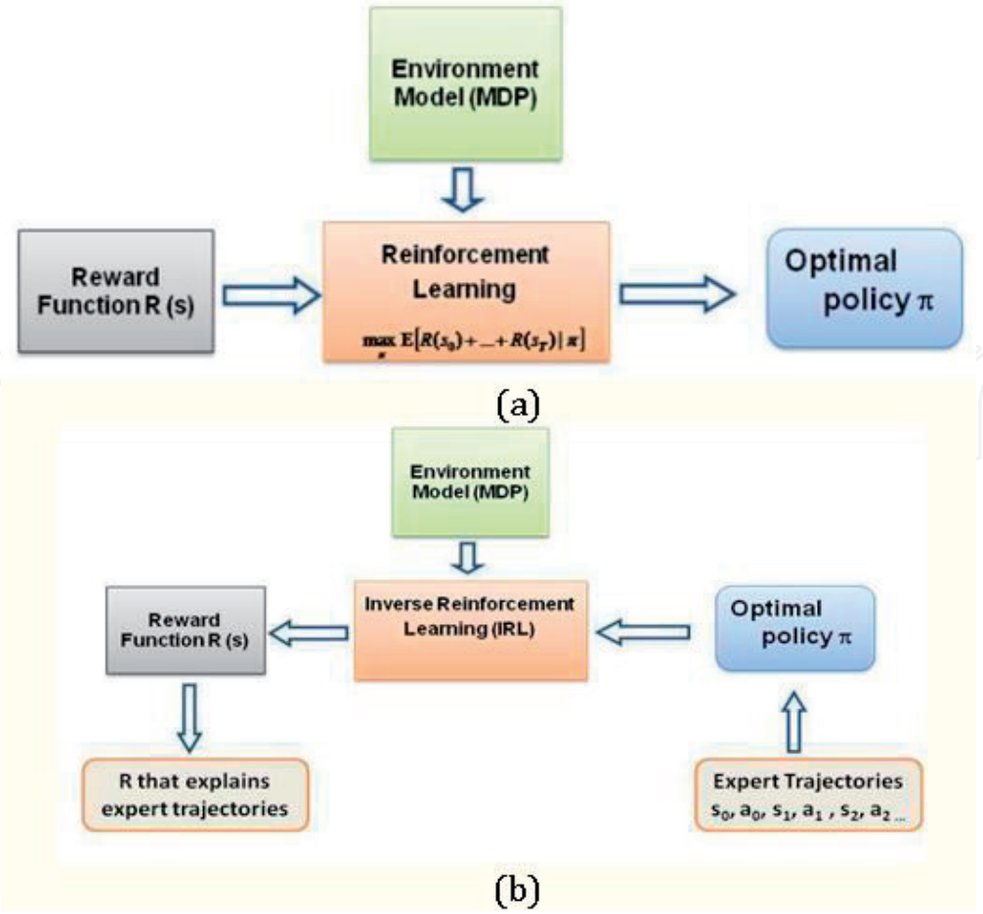

Figure 10.
(a) Reinforcement learning and (b) inverse reinforcement learning [35].

6. Discussion and concluding remarks

Next generation artificial intelligence structures are expected to have a hierarchical meta-learning ability that can adapt to many different environments, besides being a causal and explanatory power by establishing a cause-effect relationship. For this, serious effort is still needed to create flexible and interpretable models that can hold opinions from many different disciplines together and work in harmony.

We cannot ignore the advantages this will give us. For example, if we start with a medical application, after the patient data is examined, both the physician must understand and explain to the patient why he/she suggested that the explanatory decision support system suggested to the related patient that there was a "risk of heart attack." At the same time, as a meta-learning agent of this system, it has the same ability against all other diseases and it will be possible to develop appropriate treatment strategies.

While coming to this stage, what data is evaluated first is another important criterion. It is also necessary to explain what data is needed and why, and what is needed for proper evaluation. In the future, next generation deep learning and artificial intelligence forms are expected to reach the level of intelligence (singularity), which has higher performance and ability than human level. Artificial intelligence and deep learning structures mentioned in this section are thought to shed light on reaching these levels. In particular, it can be said that meta-learning approaches are capable of supporting the formation of structures that learn and adapt to multiple tasks and are also called general artificial intelligence (AGI). In the same way, it can be stated that artificial intelligence structures will help the formation of self-awareness and artificial consciousness structures based on content and causality.

Conflict of interest

The authors declare no conflict of interest.

Author details

Evren Dağlarli
Faculty of Computer and Informatics Engineering, Istanbul Technical University, Istanbul, Turkey

*Address all correspondence to: evren.daglarli@itu.edu.tr

References

[1] Adadi A, Berrada M. Peeking inside the black-box: A survey on explainable artificial intelligence (XAI). IEEE Access. 2018;**6**:52138-52160

[2] Došilović FK, Brčić M, Hlupić N. Explainable artificial intelligence: A survey. In: 2018 41st International convention on information and communication technology, electronics and microelectronics (MIPRO). IEEE; 2018. pp. 0210-0215

[3] Core MG, Lane HC, Van Lent M, Gomboc D, Solomon S, Rosenberg M. Building explainable artificial intelligence systems. AAAI; 2006. pp. 1766-1773

[4] Schnack H. Bias, noise, and interpretability in machine learning: From measurements to features. In: Machine Learning. Academic Press; 2020. pp. 307-328

[5] Huang F, Zhang J, Zhou C, Wang Y, Huang J, Zhu L. A deep learning algorithm using a fully connected sparse autoencoder neural network for landslide susceptibility prediction. Landslides. 2020;**17**(1):217-229

[6] Malik MM. A Hierarchy of Limitations in Machine Learning. 2020. arXiv preprint arXiv:2002.05193.

[7] Ahuja R, Chug A, Gupta S, Ahuja P, Kohli S. Classification and clustering algorithms of machine learning with their applications. In: Nature-Inspired Computation in Data Mining and Machine Learning. Cham: Springer; 2020. pp. 225-248

[8] Miller T. Explanation in artificial intelligence: Insights from the social sciences. Artificial Intelligence. 2019;**267**:1-38

[9] Samek W, Wiegand T, Müller KR. Explainable Artificial Intelligence: Understanding, Visualizing and Interpreting Deep Learning Models. 2017. arXiv preprint arXiv:1708.08296

[10] Fernandez A, Herrera F, Cordon O, del Jesus MJ, Marcelloni F. Evolutionary fuzzy systems for explainable artificial intelligence: why, when, what for, and where to? IEEE Computational Intelligence Magazine. 2019;**14**(1):69-81

[11] Kaushik S. Enterprise Explainable AI. 2018. Available from : https://www.kdnuggets.com/2018/10/enterprise-explainable-ai.html

[12] Dam HK, Tran T, Ghose A. Explainable software analytics. In: Proceedings of the 40th International Conference on Software Engineering: New Ideas and Emerging Results; 2018. pp. 53-56

[13] Ha T, Lee S, Kim S. Designing explainability of an artificial intelligence system. In: Proceedings of the Technology, Mind, and Society; 2018. pp. 1-1

[14] Gunning D. Explainable Artificial Intelligence (XAI). Defense Advanced Research Projects Agency (DARPA), and Web, 2; 2017

[15] Gunning D, Aha DW. DARPA's explainable artificial intelligence program. AI Magazine. 2019;**40**(2):44-58

[16] Su J, Vargas DV, Sakurai K. One pixel attack for fooling deep neural networks. IEEE Transactions on Evolutionary Computation. 2019;**23**(5):828-841

[17] Prakash PKS, Rao ASK. R Deep Learning Cookbook. Packt Publishing Ltd.; 2017

[18] Karn U. Intuitive Explanation Convolutional-Neural Networks. 2016.

Available from: https://www.kdnuggets.com/2016/11/intuitive-explanation-convolutional-neural-networks.html/3

[19] Schmidt RM. Recurrent Neural Networks (RNNs): A gentle Introduction and Overview. 2019. arXiv preprint arXiv:1912.05911.

[20] Keneni BM, Kaur D, Al Bataineh A, Devabhaktuni VK, Javaid AY, Zaientz JD, et al. Evolving rule-based explainable artificial intelligence for unmanned aerial vehicles. IEEE Access. 2019;**7**:17001-17016

[21] Futia G, Vetrò A. On the integration of knowledge graphs into deep learning models for a more comprehensible AI—Three Challenges for Future Research. Information. 2020;**11**(2):122

[22] Melis, D. A., Jaakkola, T. Towards robust interpretability with self-explaining neural networks. In: Advances in Neural Information Processing Systems; 2018. pp. 7775-7784

[23] Vilalta R, Drissi Y. A perspective view and survey of meta-learning. Artificial Intelligence Review. 2002;**18**(2):77-95

[24] Pfahringer B, Bensusan H, Giraud-Carrier CG. Meta-learning by landmarking various learning algorithms. In: ICML; 2000. pp. 743-750

[25] Finn C, Abbeel P, Levine S. Model-agnostic meta-learning for fast adaptation of deep networks. In: Proceedings of the 34th International Conference on Machine Learning-Volume 70. JMLR. org; 2017. pp. 1126-1135

[26] Chan PK, Stolfo SJ. Experiments on multistrategy learning by meta-learning. In: Proceedings of the Second International Conference on Information and Knowledge Management; 1993. pp. 314-323

[27] Schweighofer N, Doya K. Meta-learning in reinforcement learning. Neural Networks. 2003;**16**(1):5-9

[28] Santoro A, Bartunov S, Botvinick M, Wierstra D, Lillicrap T. Meta-learning with memory-augmented neural networks. In: International Conference on Machine Learning; 2016. pp. 1842-1850

[29] Amit R, Meir R. Meta-Learning by Adjusting Priors Based on Extended Pac-Bayes Theory. 2017. arXiv preprint arXiv:1711.01244

[30] Chan PK, Stolfo SJ. Toward parallel and distributed learning by meta-learning. In: AAAI workshop in Knowledge Discovery in Databases; 1993. pp. 227-240

[31] Vanschoren J. Meta-learning. In: Automated Machine Learning. Cham: Springer; 2019. pp. 35-61

[32] Khodak M, Balcan MFF, Talwalkar AS. Adaptive gradient-based meta-learning methods. In: Advances in Neural Information Processing Systems; 2019. pp. 5915-5926

[33] Frans K, Ho J, Chen X, Abbeel P, Schulman, J. Meta Learning Shared Hierarchies. 2017. arXiv preprint arXiv:1710.09767

[34] Botvinick M, Ritter S, Wang JX, Kurth-Nelson Z, Blundell C, Hassabis D. Reinforcement learning, fast and slow. Trends in Cognitive Sciences. 2019

[35] Jangir R. Apprenticeship Learning Using Inverse Reinforcement Learning. 2016. Available from: https://jangirrishabh.github.io/2016/07/09/virtual-car-IRL/

[36] Rakelly K, Zhou A, Quillen D, Finn C, Levine S. Efficient Off-Policy Meta-Reinforcement Learning via Probabilistic Context Variables. 2019. arXiv preprint arXiv:1903.08254

[37] Sugiyama T, Schweighofer N, Izawa J. Reinforcement Meta-Learning Optimizes Visuomotor Learning. 2020. bioRxiv

[38] Parisotto E, Ghosh S, Yalamanchi SB, Chinnaobireddy V, Wu Y, Salakhutdinov R. Concurrent Meta Reinforcement Learning. 2019. arXiv preprint arXiv:1903.02710

[39] Jabri A, Hsu K, Gupta A, Eysenbach B, Levine S, Finn C. Unsupervised curricula for visual meta-reinforcement Learning. In: Advances in Neural Information Processing Systems; 2019. pp. 10519-10530