Explanatory Artificial Intelligence, Results and Prospects

Alexey Averkin[1,2,3]

.

[1] Federal Research Center of Computer Sciences and Management of RAS, Vavilova, 40, Moscow, 117333, Russia
[2] Plekhanov Russian University of Economics, Stremyanny per. 36, Moscow 117997, Russia
[3] Dubna State University, Universitetskaya street, 19, Dubna, 141982, Moscow Region, Russia

Abstract

Describes the DARPA Explanatory Artificial Intelligence (XAI) program, which seeks to create artificial intelligence systems whose learning models and solutions can be understood and properly validated by end users. DARPA considers XAI as artificial intelligence systems AI that can explain their decision to a human user, characterize their strengths and weaknesses, and how they will behave in the future. To achieve this goal, methods have been developed for constructing explainable models of intelligent systems that are effective explanatory interfaces and psychological models of users for effective explanation. The XAI development teams are described that solve these three problems by creating and developing explainable machine learning (ML) technologies, developing principles, strategies, and methods of human-computer interaction for obtaining effective explanations and applying psychological explanatory theories to assess the quality of XAI systems.

Keywords

Explainable artificial intelligence, DARPA, Machine learning, Neural networks

1. Introduction

Major advances in artificial intelligence (AI), machine learning and especially deep learning neural networks have led to a new wave of AI systems (transport, security, medicine, mechanical engineering, defense). Often these systems offer solutions that are superior in quality to human ones but cannot explain their decisions and actions to users. This disadvantage is especially significant for military applications, which require the development of increasingly intelligent and autonomous. The DARPA Explanatory Artificial Intelligence (XAI) Program seeks to create artificial intelligence systems whose learning models and solutions can be understood and properly validated by end users. Achieving this goal requires building new generations of explainable models, developing effective explanatory interfaces, and building user models for more effective explanation. Explainable AI is needed for users to understand, trust, and effectively manage their smart partners. DARPA views XAI as AI systems that can explain their decision to a human user, characterize their strengths and weaknesses, and predict their future behavior. The goal of DARPA is to create more human-readable artificial intelligence systems using effective explanations. XAI development teams create and develop Explainable Machine Learning (ML) technologies, developing principles, strategies, and methods for human-computer interaction to generate effective explanations. The development teams are also evaluating how well XAI system explanations improve user experience, confidence, and productivity.

Russia also pays great attention to the direction of explainable artificial intelligence. So Nizhny Novgorod State University in 2020 became the winner in the competition of large scientific projects from the Ministry of Education and Science of the Russian Federation with the project "Reliable and

---

Russian Advances in Fuzzy Systems and Soft Computing: Selected Contributions to the 10th International Conference «Integrated Models and Soft Computing in Artificial Intelligence» (IMSC-2021), May 17–20, 2021, Kolomna, Russian Federation

EMAIL: averkin2003@inbox.ru.
ORCID: 0000-0003-1571-358

logically transparent artificial intelligence: technology, verification and application for socially significant and infectious diseases." The main result of the project should be the development of new methods and technologies that allow to overcome two main barriers of machine learning and artificial intelligence systems: the problem of errors and the problem of explicitly explaining solutions. Today these problems do not have a satisfactory solution and require new developments. The project manager, Professor Alexander Gorban, explained the main idea of the project: “These problems are closely related: without the possibility of logical reading, the errors of artificial intelligence will remain inexplicable. Additional training of the system within the framework of existing methods can damage existing skills and, on the other hand, can require huge resources, which is impractical in serious tasks. For example, the well-known IBM Watson cognitive computing system has failed in the personalized medicine market due to systematic errors in diagnosing and recommending cancer treatments that could not be found and eliminated".

2. Technologies and standards related to XAI

The main trend in Hype Cycle 2020 (Fig. 1) is a shift in focus from robotics and hardware to artificial intelligence, including explainable and interpretable artificial intelligence.

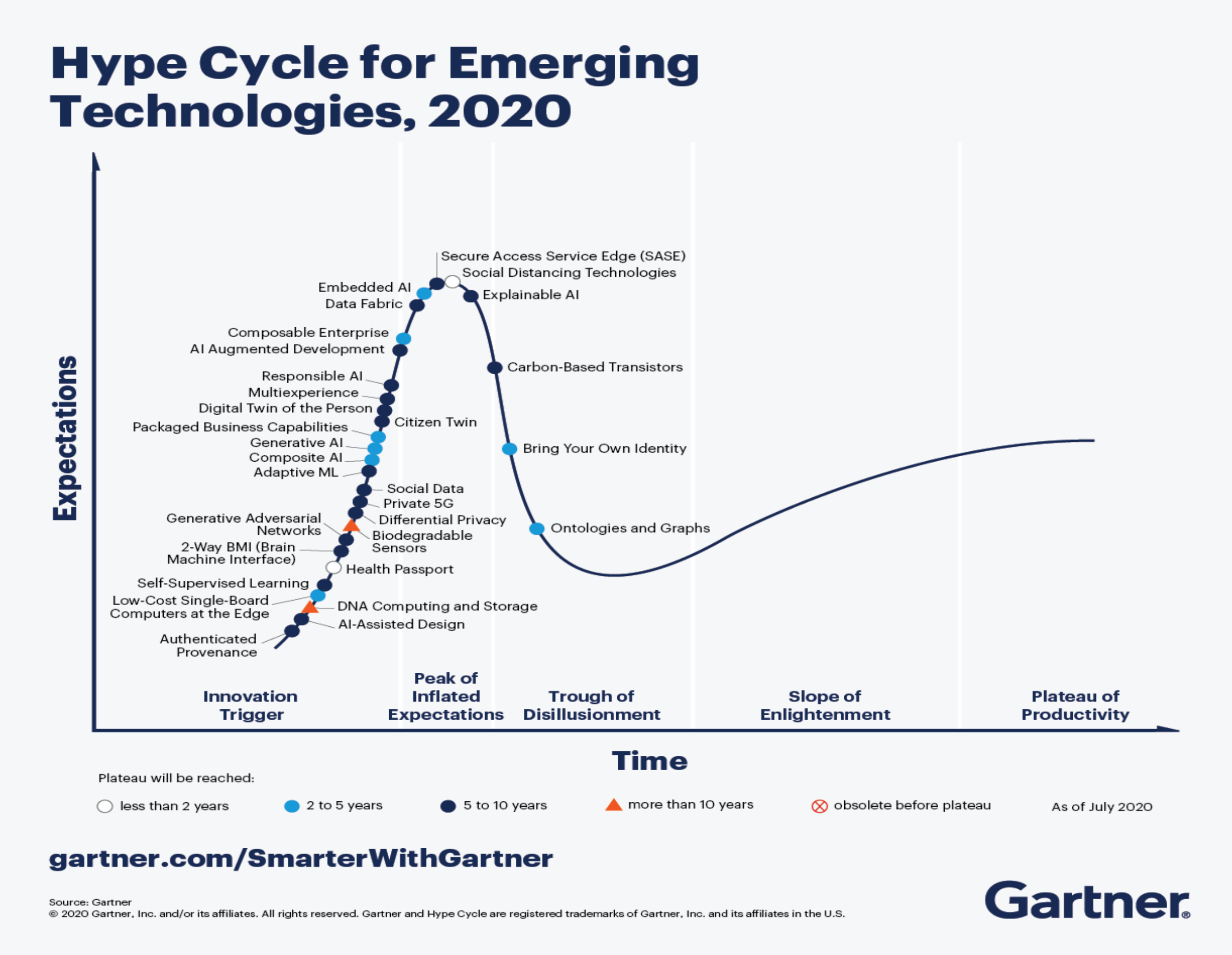


Figure 1: Gardner's Hypercycle for Emergent Technologies 2020.

On the 2020 hypercycle, the following Explainable AI technologies can be noted:

- Generative AI. Generative AI refers to programs that can use existing content, such as text, audio files, or images, to create new believable content. Various techniques exist for this, such as generative adversarial networks (GANS), transformers, and variational autoencoders.
- Adaptive ML. Adaptive machine learning - algorithms that are retrained as new data becomes available.

- Augmented Intelligence. Artificial intelligence that helps a person and does not replace him in decision-making processes. Contrasted with the general term "Artificial Intelligence" as a collective term for "Human Intelligence + AI".
- Transfer Learning. An approach in machine learning, when the accumulated experience in solving one problem is used to accelerate the learning of another similar problem (Wikipedia)
- Emotion AI AI that recognizes human emotions
- Responsible AI. Responsible AI is focused on ensuring the ethical, transparent, and accountable use of AI technologies in accordance with user expectations, organizational values, and social laws and regulations. Responsible AI ensures that automated decisions are justified and explainable and helps maintain user trust and personal privacy.
- Explainable AI. AI techniques and techniques that explain AI results to living experts. Contrasted with the concept of "AI as a black box", when it is impossible to understand the essence of the algorithms used and the relationships found.

The National Institute of Standards and Technology (NIST) has published a draft list of principles for Explainable Artificial Intelligence (XAI) [1]. AI systems with an emphasis on human-computer interaction. The document defines four principles underlying explainable AI:

1. Explanation. AI systems should provide reasons and circumstances based on which certain decisions were made. The principle of explanation obliges the AI system to provide an explanation in the form of "evidence or justification for each outcome."

2. Significance. Explainable AI systems must provide explanations that are understandable to individual users. The principle of significance states that the recipient of the explanation must be able to understand the explanation. The explanations should be tailored to the audience, both at the group and individual level.

3. Accuracy of explanation. The explanation must reliably reflect the nature of the processes that the AI system produces to generate the results. This principle is a detailed explanation of how the system generated the final result. The application of this principle also depends on the context and the end user. Thus, different measures of explanation accuracy will be presented for different types of groups and users.

4. Limits of knowledge. The system works only under the conditions for which it was designed, or when the system achieves adequate confidence in its results. The principle of knowledge limits requires the system to note any cases for which it was not designed.

These four principles show that AI-based solutions must have the necessary transparency and explainability in order to generate credibility in their functioning and confidence in the conclusions of the system.

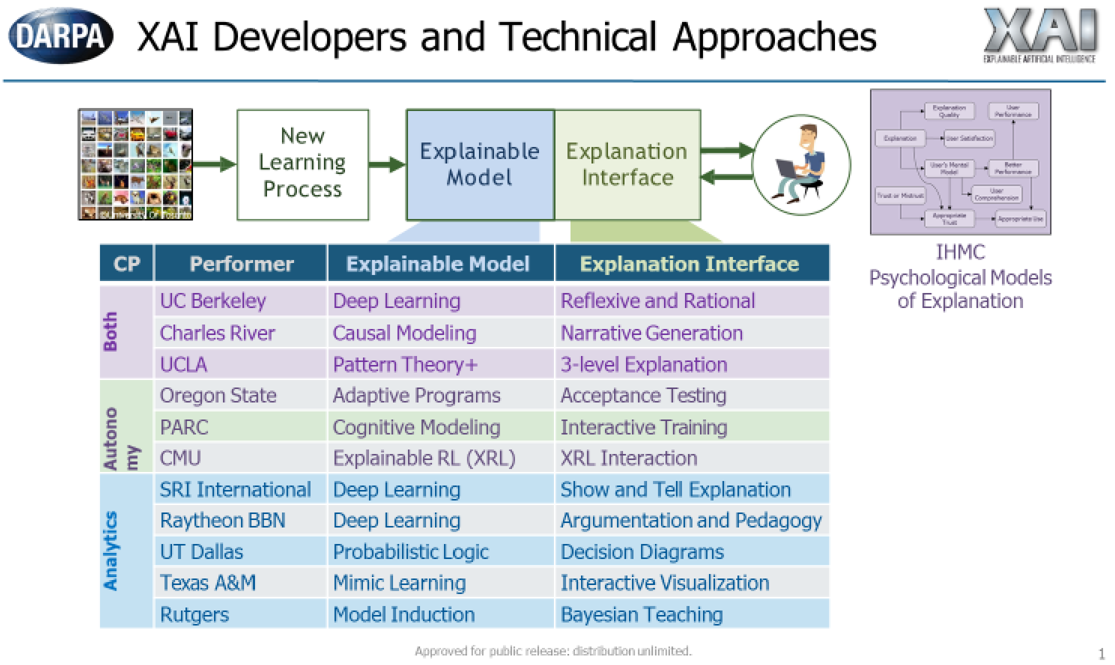
In 2021, the IEEE Society for Computational Intelligence and the Committee on Standards (CIS / SC) launched the project "Standard for XAI - Explainable Artificial Intelligence - to Achieve Clarity and Interoperability in the Design of Artificial Intelligence Systems" [2].

This standard defines mandatory and optional requirements and constraints that must be met for an AI method, algorithm, application, or system to be considered explainable. Both partially explainable and fully or strictly explainable methods, algorithms and systems are defined. There is no standard today that provides a single high-level methodology for classifying AI products as partially or fully explicable, but there is a great need for it (for example, in the Defense Advanced Research Projects Agency's DARPA program). Today, scientists and engineers developing artificial intelligence systems are limited by their specific products, customers, and conflicting interests. The problem becomes more acute when interoperability comes into play. A single standard allows you to optimize requirements and quality, increase productivity, improve the quality of the final product, and satisfy the needs of customers.

3. Development and progress of the XAI DARPA program

This section describes the DARPA XAI development teams that create and develop Explainable Machine Learning (ML) technologies, developing principles, strategies, and methods for human-computer interaction to generate effective explanations. Development teams also evaluate how well XAI system explanations improve user experience, confidence, and productivity.

Table 1 and Figure 2 show the 11 XAI technical domain teams and a team from the Florida Institute of Human and Machine Cognition (IHMC) that are developing a psychological model of explanation. Three teams are working on both research areas of concern (autonomy and data analysis), three are working only on the first, and five are working on only the second. They all explore different methods for developing explainable models and explanatory interfaces [3].



XAI Developers and Technical Approaches

CP	Performer	Explainable Model	Explanation Interface
Both	UC Berkeley	Deep Learning	Reflexive and Rational
	Charles River	Causal Modeling	Narrative Generation
	UCLA	Pattern Theory+	3-level Explanation
Autonomy	Oregon State	Adaptive Programs	Acceptance Testing
	PARC	Cognitive Modeling	Interactive Training
	CMU	Explainable RL (XRL)	XRL Interaction
Analytics	SRI International	Deep Learning	Show and Tell Explanation
	Raytheon BBN	Deep Learning	Argumentation and Pedagogy
	UT Dallas	Probabilistic Logic	Decision Diagrams
	Texas A&M	Mimic Learning	Interactive Visualization
	Rutgers	Model Induction	Bayesian Teaching

Approved for public release: distribution unlimited.

1

.

Figure 2: XAI developers and their technical approaches

Table 1.
XAI Program Development Teams. [3].

Team	Explained model	Explaining interface	Tasks to be solved
1.University of Berkeley (UCB)	Explaining ex post facto by training additional DL models. Explicit Introspective Explanations (NMN) Reinforcement Learning (Informative Deployments, Explicit Modular Agent)	Reflexive explanations (derived from the model). Rational explanations (based on reasoning about the user's beliefs)	Autonomy: vehicle control (BDD - X , CARLA), strategy games (StarCraft II) . Analytics: visual quality control and task filtering (VQA - X , ACT - X , ,xView, DiDeMo)
2. Charles River Analytics (CRA)	Experiment with a trained model to create an explainable causal programming model.	Interactive visualization based on the generation of temporal, spatial narratives from causal, probabilistic models	Autonomy: Atari , StarCraft II Analytics: Pedestrian Detection (INRIA), Activity Recognition (ActivityNet)
3.University of Los Angeles (UCLA)	Interpreted representations: STC-AOG (spatial, temporal, causal models), STC - ZP (interpretation and analytics of scenes and events)	Three-level explanation of concepts, causal and counterfactual reasoning, the	Autonomy: The robot performs daily tasks in a physically realistic virtual reality platform for

		explanation of the usefulness	autonomous driving (GTA5 game engine)
4. Oregon State University (OSU)	xDAP, a combination of adaptive programs, deep learning and explainability	It provides visual alternation and NL explained interface for acceptance test pilots - verifiers based on IFT	Autonomy: the same real-time strategies based on a specially designed game engine that supports explanations; Starcraft
5. PARC	Three-level architecture: learning level, cognitive level, explanation level	Interactive visualization of states, actions, strategies, quantities	Autonomy: shell MAVSim in the simulation environment ARduPilot
6. Carnegie Mellon University	New scientific discipline for XRL with work on new algorithms and representations	Interactive explanations of dynamical systems. Human-machine interaction for increased productivity	Autonomy: OpenAI Gym, grid autonomy, mobile service robots, self-improving educational software
7. SRI	Multiple DL machinery, mechanisms, based on the attention, composite NMN, GAN.	DNN visualization. A response to a query explaining DNN solutions. Generation of NL justifications.	Analytics: VQA Visual Gnome. Flick30), MovieQA
8. Raytheon BBN	DNN semantic markup. Create DNN audit trail	Gradient-weighted display of class activation	Analytics for images and videos
9. UTD	Controlled Probabilistic Logic Models (TPLM)	Allows users to explore and correct the baseline model, and add baseline knowledge	Analytics: explanation action system in a multi-modal data (video and text), biological data and data sets culinary scenes with text annotations
10. Texas A&M University (TAMU)	The simulation learning framework combines DL models for prediction and shallow models for explanations. Interpretable learning algorithms extract knowledge from DNN for appropriate explanations.	Interactive visualization of multiple news stories using heatmaps and topic modeling clusters to display predictive functions	Analytics: Multiple tasks using data from Twitter, Facebook, ImageNet, and news sites.
11. Rutgers	Selection of optimum examples to explain the model solutions based on Bayesian their reasoning	Explanation of the complete model based on examples; examples provided by the user	Analytics: image processing, text corpus, VQA, movie events

Below is a more detailed description of the activity of the commands.

3.1. Deep Explainable AI (DEXAI).

The University of California Berkeley (UCB) team (including researchers from Boston University, University of Amsterdam, and Kitware) is developing an artificial intelligence system that is understandable to humans through explicit structural interpretation [4] and introspective explanation

[5] that has predictable behavior and high confidence in the result [6]. The key challenges of Deep Explainable AI (DEXAI) are generating accurate explanations of the model's behavior and choosing those that are most useful to the user. UCB addresses the first problem by creating implicit or explicit explanation models: they can implicitly represent complex hidden representations in understandable ways, or they can build explicit structures. These DEXAI models create a set of possible explainable actions. For the second problem, UCB proposes rational explanations that use the user's belief model in decision making to select explanatory actions. UCB is also developing an explain interface based on the principles of iterative design. DEXAI's autonomy is demonstrated in vehicle handling (using the Berkeley Deep Drive dataset and CARLA simulator) [7] and in strategy game scenarios (StarCraft II). For analytics, DEXAI uses visual question answers (VQA) and filtering techniques, for example using large datasets such as VQA-X and ACT-X for VQA and activity recognition tasks [8].

3.2. Causal models to explain machine learning.

The goal of the Charles River Analytics (CRA) team (including researchers from the University of Massachusetts and Brown University) is to create and provide causal explanations for machine learning using causal models to explain the Learning Approach (CAMEL). CAMEL explanations are presented to the user as stories in an interactive, intuitive interface. CAMEL includes a framework for causal probabilistic programming that integrates concepts and teaching methods from causal modeling [9] with probabilistic programming languages [10]. Generative probabilistic models, presented in a probabilistic programming language, naturally express cause-and-effect relationships. CAMEL builds a causal model of their impact on the operation of a machine learning system by conducting experiments in which areas of agreement are systematically included or removed. After training, he uses causal models to derive explanations for the predictions or actions of the system. In the field of analytics and data analysis, CAMEL is solving the problem of pedestrian detection (using the INRIA pedestrian dataset) [11], and the CRA is working on the problems of activity recognition (using ActivityNet). CAMEL's autonomy is demonstrated in the Atari Amidar game, and CRA is working to demonstrate it in StarCraft II.

3.3. Learning and communicating explainable views for analytics and autonomy.

The University of California Los Angeles (UCLA) team (including researchers from Oregon State University and Michigan State University) develops interpretable models that combine representational paradigms, including interpreted DNNs, compositional graphical models such as AND / OR graphs, and models that produce explanations at three levels (i.e., compositionality, causality, and utility). The UCLA system includes an execution module, which performs tasks with multimodal inputs, and an explain module, which explains its perception, cognitive reasoning, and decisions to the user. The execution engine outputs interpreted representations in the form of a graph of spatial, temporal and causal analysis (STC-PG) for 3D scene perception (for analytics) and task scheduling (for autonomy). STC-PG are compositional, probabilistic, interpretable and based on DNN techniques and are used for image and video analysis. The explain module displays an explanatory syntactic graph during the dialogue [12], localizes the corresponding subgraph in the STC-PG and determines the user's intentions. For data analytic analysis, UCLA applied its system to a network of video cameras to understand the meaning of the scene and analyze the events. UCLA has demonstrated the autonomy of the system in scenarios using robots performing tasks on virtual reality platforms and in a game with driving an autonomous vehicle.

3.4. Acceptance testing of deep adaptive programs with sound information

Oregon State University (OSU) develops tools to explain the actions of trained agents that perform consistent decision making and identifies the best principles for developing user interfaces with explanations. The OSU Explainable Agent Model uses Explainable Deep Adaptive Programming (xDAP), which combines adaptive programming, deep reinforcement learning (RL), and explainability. With xDAP, programmers can create agents that represent solutions that are automatically optimized through deep RL when interacting with the simulator. For each selection point, deep RL connects a trained deep decision-making neural network (dNN), which can provide high performance, but is inherently inexplicable. After initial xDAP training, the xACT deep adaptive acceptance testing system trains an explanatory neural network [13] for each dNN. They provide a set of explain functions (x-functions) that encode properties of the dNN decision logic. Such x-functions, which are neural networks, are not originally interpretable by humans. To solve this problem, xACT allows domain experts to attach interpretable descriptions to x-functions, and xDAP programmers to annotate environment reward types and other concepts that are automatically embedded in dNNs as “annotation concepts” during training. OSU has demonstrated xACT in scripts using a custom-built real-time game engine. Pilot studies have provided information to explain user interface design by describing how users navigate in an AI game and explain game decisions.

3.5. General training and explanation.

A Palo Alto Research Center (PARC) team (including researchers from Carnegie Mellon University, the Army CyberInstitute, the University of Edinburgh, and the University of Michigan) is developing an interactive reasoning system that could explain the capabilities of the XAI system, which controls a simulated unmanned aerial system. Explanations of the XAI system should communicate what information it uses to make decisions, how the system itself works and its goals. To this end, the PARC (COGLE) general learning and explanation system and its users establish a common basis for defining which terms to use in explanations and their meanings. This is provided by the PARC introspective discourse model, which alternates between learning and explanation.

COGLE's layered architecture separates information processing into comprehension, cognitive modeling, and learning. The learning layer uses repetitive and hierarchical DNNs with limited bandwidth to create abstractions and compositions on the states and actions of unmanned aerial systems to support understanding of generalized patterns.

COGLE's two explanatory interfaces support performance analysis, risk assessment, and training. The first is a map that tracks the actions of unmanned aerial systems and divides actions or decisions in flight into explainable segments. Second interface tools allow users to explore and assess system competencies and make predictions about mission performance. COGLE is being demonstrated on the ArduPilot Software-in-the-Loop Simulator and on the discrete abstract simulation test bed. Its quality is evaluated by drone operators and analysts. Competency-based assessment will help PARC determine how best to develop suitable models that are understandable for the domain.

3.6. Explainable reinforcement learning.

Carnegie Mellon University is creating a new direction of explainable RL to enable dynamic human-machine interaction and adaptation. It has two goals: to develop new methods for learning explainable RL algorithms, and to develop strategies that can explain existing black box algorithms. To achieve the first goal, Carnegie Mellon is developing methods to improve model learning for RL agents to take advantage of model-based approaches while combining them with the benefits of model-free approaches. Methods are used that gradually add states and actions to models of the world after hidden information is discovered, study models through end-to-end training on complex optimal control algorithms, study general DL models that use rigid body physics [15] and predict states using iterative architectures [16]. Carnegie Mellon University is also developing methods that can explain the actions and plans of RL black box agents. This includes answering questions such as “Why did the agent choose a particular action?” or “What training data influenced this choice the most?” To this end, Carnegie Mellon University has developed methods that generate NL descriptions of agents from behavior logs and detect outliers or anomalies. Carnegie Mellon has demonstrated XRL in several scenarios including OpenAI Gym, Atari games, autonomous vehicle simulations, mobile service robots.

3.7. Explainable generative adversarial networks.

The SRI International team (including researchers from the University of Toronto, the University of Guelph, and the University of California, San Diego) is developing an explainable machine learning framework for multimodal data analysis that generates understandable explanations with rationale for decisions, accompanied by visualizations of input data used to generate inferences. The Deep Attention-Based Representation System for Explainable Generative Adversarial Networks (DARE / X-GANS) uses DNN architectures like models of attention in visual neuroscience. It identifies, extracts, and presents evidence to the user as part of the explanation. Attention mechanisms provide the user with the means to explore the system and work together. DARE / X-GANS uses generative adversarial networks (GANs) that learn to understand data by creating it while learning representations with explanatory power. GANs become explicable with interpreted decoders. This includes generating visual evidence for given text queries using chunked text generation [17], with chunks being interpreted features such as human poses or bounding boxes. The system presents explanations of its answers based on visual concepts extracted from multimodal input data and queries to the knowledge base. Asking explanatory questions, she provides the rationale and visual evidence used to make decisions and a visualization of the inner workings of the system. SRI focuses on the data analytics problem area and has demonstrated DARE / X-GAN work using VQA and multimodal QA tasks with image and video datasets.

3.8. A system of answers to explainable questions.

The Raytheon BBN Technologies team (including researchers from Georgia Institute of Technology, Massachusetts Institute of Technology, and the University of Texas at Austin) is developing a system that answers any natural language (NL) questions users ask about media and provides interactive possible explanations as to why he got this answer. Explainable Answer to Questions System (EQUAS) studies explainable DNN models in which internal structures (eg, individual neurons) are aligned with semantic concepts [18]. EQUAS also uses neural imaging techniques to highlight the input areas associated with neurons that most influenced its decisions. To express case-based explanations, EQUAS stores indices and extracts cases from its training data that support its selection. The four modes of explanation correspond to the key elements of argument building and interactive pedagogy: didactic statements, visualizations, cases, and rejection of alternatives. The EQUAS explain interface provides iterative and controlled collaboration, allowing users to dig deeper into corroborating evidence from each category of explanation. Raytheon BBN, demonstrated the initial capabilities of EQUAS for VQA tasks for image analysis, exploring how various explanation methods allow users to understand and predict the behavior of the underlying VQA system.

3.9. Controlled probabilistic logical models.

A team at the University of Texas at Dallas (UTD) (including researchers from UCLA, Texas A&M, and Indian Institute of Technology Delhi) is developing a unified approach to XAI using controlled probabilistic logic models (TPLM). TPLM is a family of representations that includes decision trees, binary decision diagrams, section networks, maximal decision diagrams, first-order arithmetic schemes, and controlled Markov logic [19]. UTD extends TPLM to generate explanations of query results. For scalable inference, the system uses new algorithms to answer complex explanatory queries using techniques such as generalized inference, variational inference, and combinations thereof. It uses discriminatory techniques to quickly and improve training accuracy, deriving algorithms that make up NN and support vector machines with TPLM. These approaches are then extended to handle real-world situations. The UTD explain interface displays interpreted views with multiple related explanations. Its interactive component allows users to debug the model and provide alternative explanations. UTD focuses on the analytics problem area and is demonstrating its human activity recognition system in multimodal data (video and text) such as a text annotated cookery scene dataset.

3.10. Interpretable Deep Learning.

A team at Texas A&M University (TAMU) (including researchers from Washington State University) is developing an interpretable DL framework that uses simulation learning to use explicable shallow models and facilitates domain interpretation with visualization and interaction. Simulation learning bridges the gap between deep and shallow models and provides interpretability. The system also extracts informative patterns from raw data to improve interpretability and learning efficiency. Interpretable system learning algorithms extract knowledge from DNN for appropriate explanations. Its DL module connects to the template generation module using the interpretability of shallow models. The TAMU system processes image data [20] and text [21] and it is applied in the XAI analytics problem domain. It provides efficient interpretation of detected inaccuracies from a variety of sources while maintaining competitive detection performance. The system has been deployed to address multiple challenges using data from Twitter, Facebook, ImageNet, CIFAR-10, online health forums and news websites.

3.11. Explaining the model using the optimal choice of training examples.

Rugers University is expanding the capabilities of Bayesian learning to enable automatic explanation by choosing the subset of the data that is most representative of the model's inference. Rugers' approach allows one to explain the conclusions of any probabilistic generative and discriminative model, as well as DL models [22]. Rudgers also develops a formal theory of human-machine interaction and supports interactive explanations of complex compositional models. Common among these is a basic approach based on human learning models that promote explainability and carefully controlled behavioral experiments to quantify explainability. Explaining with Bayesian Learning introduces a dataset, a probabilistic model, and an inference method, and returns a small subset of examples that best explain the inference of the model. Using composition and co-modification of machine learning models, Rudgers offers a general approach to understanding through guided exploration. Interaction occurs through an interface that exposes the structure of the model and explains each data component. It has been demonstrated that Ruggers' approach facilitates understanding of large corpora, as measured by a person's ability to accurately compose corpus summaries after short, guided explanations. Rudgers focuses on the data analysis problem area and has demonstrated his approach in images, text, combinations of both (such as VQA) and structured modeling using a temporal causal structure.

4. Methods for extracting rules from neural networks

The development of explicable artificial intelligence methods was largely preceded by methods for extracting rules from neural networks. In artificial intelligence, neural networks and rule-based learning methods are two approaches to solving classification problems. Both methods are known variants of learning models that predict classes for new data. For many tasks, rule-based neural network learning methods are very accurate. However, neural networks have one major drawback: the ability to understand the conceptual essence of trained models is weaker in the neural network than in the rule-based approaches. The concepts gained from training neural networks are difficult to understand because they are represented using a large set of parameters [23].

Increasing the transparency of neural networks by extracting rules from them has two main advantages. This gives the user some insight into how the neural network uses the input variables to decide allows the hidden features in the neural networks to be revealed when rules are used to explain individual neurons. Identifying critical attributes or identifying the causes of neural network errors can be part of the understanding. To make opaque neural networks more understandable, rule extraction techniques are bridging the gap between precision and clarity [23,24]. For a neural network, for example, to be used in mission-critical applications such as airplanes or power plants, a clearer form is required. In these cases, it is extremely important that the user of the system can check the output values of the artificial neural network under all possible input conditions [25].

To formalize the task of extracting rules from a neural network, you can use the following definition: "Given a trained neural network and the data on which it was trained, create a description of a network hypothesis that is understandable, but approximates the behavior of a given network." To distinguish between different approaches to rule extraction from neural networks, a multidimensional taxonomy was introduced in [25]. The first dimension it describes is the expressive power of the extracted rules (for example, IF-THEN rules or fuzzy production rules). The second dimension is called transparency and describes the strategy followed by the rule extraction algorithm. If a method uses a neural network only as a black box, regardless of the architecture of the neural network, we call it a pedagogical approach. If instead the algorithm considers the internal structure of the neural network, we call this approach decomposition. If an algorithm uses components of both pedagogical and decomposition methods, then this approach is called eclectic. The third dimension is the quality of the extracted rules. Since quality is a broad term, it is divided into several criteria, namely neatness, accuracy, consistency, and intelligibility. While accuracy measures the ability to correctly classify previously unseen examples, accuracy measures the degree to which rules can mimic neural network behavior well [24]. Precision can be thought of as precision in relation to the output of the neural network. Consistency can only be measured when the rule extraction algorithm involves training a neural network instead of processing an already trained neural network. The extracted rule set is considered consistent when the neural network generates rule sets that correctly classify test data for different training sessions. Comprehensibility is considered here as a measure of the size of the rules, that is, short rules are considered more comprehensible when there are fewer rules. An overview of many methods for extracting rules from neural networks based on this taxonomy is given in [26].

The most interesting from the point of view of this study is rule extraction using neuro-fuzzy models. Fuzzy rule-based systems (FRBS) developed using fuzzy logic have become a field of active research over the past few years. These algorithms have proven their strengths in tasks such as managing complex systems, creating fuzzy controls. The relationship between both worlds (ANN and FRBS) has been thoroughly studied and shown to be equivalent [27]. This provides important insights. First, we can apply what was found for one of the models to the other. Second, we can translate the knowledge embedded in the neural network into a more cognitively acceptable language - fuzzy rules. In other words, we get a semantic interpretation of neural networks [28,29].

5. Conclusion

Advances in machine learning and the rise in computing power have led to the development of intelligent systems that can be used to recommend a movie, diagnose cancer, make investment decisions, or drive a car without a driver. However, the effectiveness of these systems is limited by the inability to explain decisions and actions to the user. The XAI DARPA program develops and evaluates a wide range of new machine learning methods: modified DL methods that study explainable functions; methods that explore more structured, interpretable causal patterns; and model induction methods that derive an explainable model from any black box model. The technologies and results obtained show that these three broad strategies deserve further study and will provide future developers with design options that increase productivity and explainability. A very important special case of explainable artificial intelligence is rule extraction from neural networks. For this, experience, and knowledge in the field of fuzzy logic is well suited for modeling ambiguities in big data, modeling uncertainty in knowledge representation, and providing learning with non-inductive inference.

6. Acknowledgements

This work was carried out with the financial support of the Russian Foundation for Basic Research, project No. 20-07-00770

7. References

1. P. Jonathon Phillips et al. Four Principles of Explainable Artificial Intelligence, Draft NISTIR 8312, 2020. URL: https://doi.org/10.6028/NIST.IR.8312-draft.
2. P2976 - Standard for XAI – eXplainable Artificial Intelligence - for Achieving Clarity and Interoperability of AI Systems Design. URL: https://standards.ieee.org/project/2976.html
3. D. Gunning. DARPA's explainable artificial intelligence (XAI) program. IUI '19: Proceedings of the 24th International Conference on Intelligent User Interfaces, 2019. doi.org/10.1145/3301275.3308446.
4. D. H. Park, L. A. Hendricks, Z. Akata, A. Rohrbach, B. Schiele, T. Darrell, and M. Rohrbach, Multimodal Explanations: Justifying Decisions and Pointing to the Evidence, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, New York, 2018. doi.org/10.1109/CVPR. 2018.00915
5. V. Ramanishka, A. Das. J. Zhang, K. Saenko, Top- Down Visual Saliency Guided by Captions, in: Proceedings of the 30th IEEE Conference on Computer Vision and Pattern Recognition, 7206–15. New York, IEEE, 2017.
6. S. H. Huang, K. Bhatia, P. Abbeel, A.bDragan, Establishing Appropriate Trust via Critical States, 13th Annual ACM/IEEE International Conference on Human-Robot Interaction Workshop on Explainable Robot Behavior. Madrid, Spain; October 1–5, 2018. doi.org/10.1109/IROS.2018.8593649
7. J. Kim, J. Canny, Interpretable Learning for Self-Driving Cars by Visualizing Causal Attention, in: Proceedings of the International Conference on Computer Vision, 2942–50. New York, 2017. doi.org/10.1109/ICCV. 2017.320.
8. L. A. Hendricks, R. Hu, T. Darrell, Z. Akata, Grounding Visual Explanations. Presented at the European Conference of Computer Vision (ECCV). Munich, Germany; September 8–14, 2018. doi.org/10.1007/978-3-030-01216-8_17.
9. K. Marazopoulou, M. Maier, D. Jensen, Learning the Structure of Causal Models with Relational and Temporal Dependence, in: Proceedings of the Thirty-First Conference on Uncertainty in Artificial Intelligence, 572–81. Association for Uncertainty in Artificial Intelligence, 2015.
10. A. Pfeffer, Practical Probabilistic Programming. Greenwich, CT: Manning Publications, 2016.
11. M. Harradon, J. Druce, B. Ruttenberg, Causal Learning and Explanation of Deep Neural Networks via Autoencoded Activations. arXiv preprint. arXiv:1802.00541v1 [cs.AI]. Ithaca, NY: Cornell University Library, 2018.
12. L. She, J. Y. Chai, Interactive Learning for Acquisition of Grounded Verb Semantics towards Human-Robot Communication, in: Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics, vol. 1, pp.1634–44. Stroudsburg, 2017, PA: Association for Computation Linguistics. doi.org/10.18653/v1/P17-1150
13. Z. Qi, F. Li, Learning Explainable Embeddings for Deep Networks. Paper presented at the NIPS Workshop on Interpreting, Explaining and Visualizing Deep Learning. Long Beach, CA, December 9, 2017.
14. J. Dodge, S. Penney, C. Hilderbrand, A. Anderson, M. Burnett, How the Experts Do It: Assessing and Explaining Agent Behaviors in Real-Time Strategy Games, in: Proceedings of the 2018 CHI Conference on Human Factors in Computing Systems, New York: Association for Computing Machinery, 2018. doi.org/10.1145/3173574.3174136
15. F. Belbute-Peres, J. Z. Kolter, A Modular Differentiable Rigid Body Physics Engine. Paper presented at the Neural Information Processing Systems Deep Reinforcement Learning Symposium. Long Beach, CA, December 7, 2017.
16. A. Hefny, Z. Marinho, W. Sun, S. Srinivasa, G. Gordon, Recurrent Predictive State Policy Networks. In Proceedings of the 35th International Conference on Machine Learning, 1954–63. International Machine Learning Society, 2018.
17. P. Vicol, M. Tapaswi, L. Castrejon, S. Fidler, MovieGraphs: Towards Understanding Human-Centric Situations from Videos. In IEEE Conference on Computer Vision and Pattern Recognition. New York, 2018: IEEE. doi.org/10. 1109/CVPR.2018.00895

18. B. Zhou, A. Khosla, A. Lapedriza, A. Oliva, A. Torralba, Object Detectors Emerge in Deep Scene CNNs. Paper presented at the International Conference on Learning Representations. San Diego, CA, May 7–9, 2015.

19. V. Gogate, P. Domingos, Probabilistic Theorem Proving. Communications of the ACM 59(7), 2016: 107–15. doi.org/10. 1145/2936726

20. M. Du, N. Liu, Q. Song, X. Hu, Towards Explanation of DNN-Based Prediction and Guided Feature Inversion. In Proceedings of the 24th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, 1358–67. New York, 2018: Association for Computing Machinery. doi.org/ 10.1145/3219819.3220099.

21. J. Gao, N. Liu, M. Lawley, X. Hu, An Interpretable Classification Framework for Information Extraction from Online Healthcare Forums. Journal of Healthcare Engineering, 2017: 2460174. doi.org/10.1155/2017/ 2460174

22. S. C.-H. Yang, P. Shafto, Explainable Artificial Intelligence via Bayesian Teaching. Paper presented at the 31st Conference on Neural Information Processing Systems Workshop on Teaching Machines, Robots and Humans. Long Beach, CA, December 9, 2017.

23. M. Craven, J. Shavlik, Rule extraction: Where do we go from here. University of Wisconsin Machine Learning Research Group Working Paper, 1999, pp. 99–108.

24. U. Johansson, T. Lofstrom, R. Konig, C. Sonstrod, L. Niklasson, Rule extraction from opaque models – a slightly different perspective. In Machine Learning and Applications, ICMLA'06. 5th International Conference on Machine Learning and Applications, 2006, pp. 22–27.

25. R. Andrews, J. Diederich, A. B. Tickle, Survey and critique of techniques for extracting rules from trained artificial neural networks. Knowledge-based systems, 8(6), 1995, pp. 373–389.

26. A. N. Averkin, S. A. Yarushev, Hybrid intelligent system of rules extraction for decision making 2020 J. Phys.: Conf. Ser. 1703 012007.

27. R. Setiono, W. K. Leow, FERNN: An algorithm for fast extraction of rules from neural networks. Applied Intelligence, 12(1-2), 2000, pp. 15–25.

28. A. Averkin, S. Yarushev, Hybrid Neural Networks and Time Series Forecasting. Artificial Intelligence. Communication in Computer and Information Sciences 934 – Springer, 2018, pp.230-239.

29. A. N. Averkin, G. Pilato, S. Yarushev, An Approach for Prediction of User Emotions Based on ANFIS in Social Networks, Second International Scientific and Practical Conference Fuzzy Technologies in the Industry, FTI 2018– CEUR Workshop Proceedings, 2018, pp. 126-134.