Recent Trends in XAI: A Broad Overview on current Approaches, Methodologies and Interactions

Jakob M. Schoenborn[1,2] and Klaus-Dieter Althoff[1,2]

[1] University of Hildesheim
Samelsonplatz 1
31141 Hildesheim
schoenborn@uni-hildesheim.de
[2] German Research Center for Artificial Intelligence (DFKI)
Trippstadter Str. 122
67663 Kaiserslautern
kalthoff@dfki.uni-kl.de

Abstract. The definition of an explainable artificial intelligence heavily depends on the use-case, whether one is focusing on the technical knowledge-management component [30, 33, 37, 43] or rather the more social interaction including speech acts and conversations [27, 31, 33]. Since the uprising debate of the unknown outcome on the development of AI in general using Deep Learning [4, 34, 35, 44] and recent legal restrictions (for example the GDPR [19]), the need on developing an explainable AI is rapidly increasing, especially since the last two years. Additionally, the goal to increase the users trust towards AI has still to be achieved. Thus, this contribution aims to provide an overview on the current topics especially since 2018 with a focus on case-based explanations[3] up until today.

Keywords: Explanation, XAI, Framework, Case-Based Explanation

1 Introduction

Seemingly any general discussion on artificial intelligence contains at least some sort of statement that explainable artificial intelligence (XAI) will be a crucial component in future systems [8, 9, 15, 17, 21, 23]. These mentions are usually on a rather general level - without becoming too specific on how an explanation can actually be automatically generated by any kind of algorithm or methodology [13, 27, 33, 43]. This is also reflected by the very small amount of practically used and evaluated systems. However, some approaches seem to be promising, e. g.: Black Box Explanations through Transparent Approximations (BETA) [28], Local Interpretable Model-Agnostic Explanations (LIME) [38] and additive models with pairwise interactions (GAMs) [6] (see [23]). Depending on the point of view,

[3] For surveys before 2018, we refer to the interested reader to [1, 2, 10, 24, 46]

different systems can be considered as the "first" XAI. One among those in terms of "making sense" as defined by Schank [40] is SWALE [41]. Others might argue that explainable AI has always been a part in developing an AI - thus referring to expert systems in general, ranging back to Weizenbaums ELIZA in 1966 [45]. However, the common goal has not changed: to create a component that understands and makes sense of the underlying data in a certain context [23]. This goal is financially supported by the European commission by investing an additional 1.5 billion EUR (a total of 20 billion EUR by 2020) "*and more than 20 billion euro per year from public and private investments over the following decade*" [12, 23]. For this survey we define XAI in the following way:

"*An explainable artificial intelligence enables an user to learn a transparent, relevant and justified information at the right time using an appropriate size.*"

Each approach dealing with explanations has two common tasks as an explanation foundation: A knowledge management and maintenance task to solve and an appropriate interface to socially interact with the user, even on a one-sided level, by providing an explanation. Whichever methodology is used to solve these tasks, the cited works have shown that each of these can be used in multiple domains and thus can learn from each other.

2 Related Work

2.1 Results of previous surveys

The movement from the different fields of AI to XAI can also be proven by the rising numbers of surveys on XAI. Holzinger [24], Došilovič [10], and Adadi [2] surveyed (among others, e. g. [1, 20, 46]) in 2018 the current trends of XAI and how to move from black-box machine learning to glass-box XAI. Adadi et al. provided a comprehensive overview on key related concepts of XAI with a schematic view of XAI related concepts [2]. The authors motivate on different explanation goals and how they are used in certain domains: to control, to improve, to discover, to justify.

Holzinger motivates different trend indicators, i. e. multiple global industrial companies using AI - especially in recommendation systems, since the overall goal still remains to convince the user to buy another suitable product or to remain longer on their website by recommending other similar series to watch [24]. Other indicators are funding (as motivated and cited in the introduction) as well as conferences. The interest on the Neural Information Processing Systems (NIPS, now name changed to NeurIPS) conference kept its projected success on the expected amount of participants by selling out all available tickets within 11 minutes [32]. The 35^{th} international conference on machine learning (ICML) scored a similar success with selling out before the end of the submission deadline. The main problems on XAI as mentioned by Holzinger [24] and to some extent also by Došilovič [10] (with the addition of ethical and quality-of-life implications), trust, privacy, and security remain to be the core problems as of today.

As the arising problems with the general data protection regulation (GDPR) are well-discussed (i. e. [19, 24]). Additionally, there is no known technical possibility to solve the decision on when an uploaded file, e.g., a video, image, or piece of literature, is copyright-protected which has recently been discussed in Germany [39, 42], XAI could decrease the false-positive rate and provide a first step into the right direction.

2.2 Current research by domain

Fig. 1 depicts a list of recent publications since 2018 with a size of at least five pages. The most publications are applicable to multiple domains (e. g. machine learning in medical domain), but are listed only in one category - depending on its focus. This choice was made to illustrate the manifold, different domains, in which XAI currently experiences development. The spike in Machine Learning is not surprising (Deep Learning is considered to be part of Machine Learning), due to its recent success and popularity as illustrated in Fig. 2. Especially this year in 2019, almost every conference dealing with AI is either advertising XAI as their main theme or have at least one workshop attached to it [7, 14, 25, 26, 36]. Fig. 1 is expected to drastically change at the end of this year after the proceedings on the XAI centring conferences and workshops have been published, possibly also with the addition of novel domains.

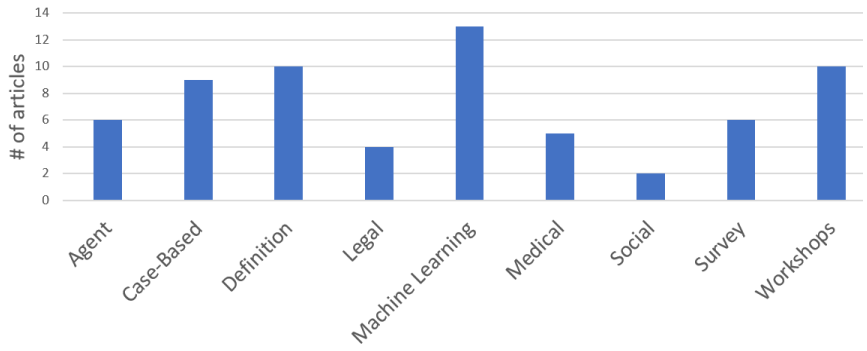

Fig. 1. Distinct publications since 2018 (without claim of completeness).

3 Machine Learning

The success of the probabilistic and statistical approaches of machine learning during the last five years is undeniable [3] and led to an increased interest in machine learning in general (see Fig. 2). But still, the decisions driven by these algorithms are mainly black-box approaches which are critical regarding trust of the user in how the decision has been made, especially in the medical domain and for decision support systems in general [5, 18, 24]. This is one of the key-challenges which are recently targeted by machine learners, which is furthermore illustrated by presenting a few exemplary suggestions (without claim of completeness).

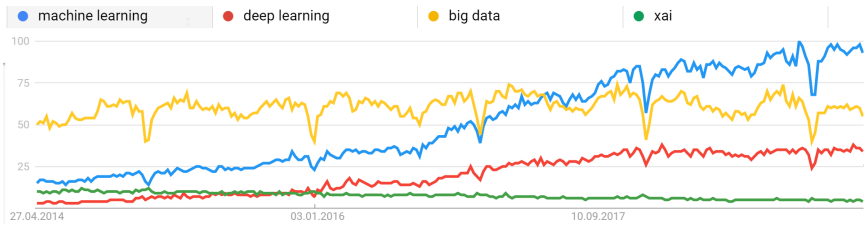

Fig. 2. Google Trends worldwide for the last five years with an increasing interest in machine learning (blue).

Ghosal et al. combined their deep convolutional neural network for leaf image classification with an additional explanation phase [16]. An image of a leaf is used as input and will be further analyzed throughout multiple feature maps to process the image. These identifications and classifications (of a diagnosis, i. e. herbicide injury or septoria brown spot) are then further used in combination to present the given solution to the user, based on the identified pairs of features and diagnosis. This is basically the same argumentation structure as an expert would use as well (explaining the diagnosis based on the identified visual features). The authors envision that this approach could be extended to animal and human diseases [16], but the acceptance proofed by an user-centring study is missing - even though the approach seems to be very promising. To provide visual explanations can also be found in the recent case-based reasoning approach by Lamy et al [29] in the medical domain (identifying breast cancer). It should be further investigated to combine these approaches, if the proposed methodology should be extended to the medical domain, since the reuse of knowledge by treating the pairs of features and diagnosis can easily be treated as a case.

Another explainable machine learning application in the medical domain is provided by Lundberg et al. to support anesthesiologists in predicting the possibility of hypoxaemia during surgery [30]. To achieve this, 20+ static features (age, BMI, ...) and 45 dynamic features are used in real-time to build a predictive model of hypoxaemia events. These features are color-encoded (pink for increased

risk, green for decreased risk) and are further combined mathematically to calculate the size of the prediction window, which supports the anaesthesiologist by knowing which attributes of the patient and procedure contributed to the current risk [30]. Each feature contains a certain range of impact (i. e. weight) and will be treated accordingly while building the explanation. The explanation itself is a real-time graph containing each relevant feature and its impact. One might argue that this can not be considered as an explanation, but during a surgery, there is no time for reading/listening to a textual representation of an explanation but rather the relevant information needs to be understandable on first sight. The approach has been evaluated against practicing anesthesiologists and achieves superior performance when predicting hypoxaemia risk from electronically recorded intraoperative data [30].

The last presented approach is a result of the further development of the mentioned GAM [6] to understand how data scientists understand machine learning models by Hohman et al, since they "...have different reasons to interpret models and tailor explanations for specific audiences..." [22]. Here, explanations are divided into six classes, local instance explanations, instance explanation comparisons, counterfactuals, nearest neighbors, reigns of error and feature importance. The distinction is important since depending on the actual use-case one class might be a better fit than the other. To decide which class is used, GAMs (generalized additive models [6]) are used and smoothed by shape functions f_i. The generated explanation itself is presented as a waterfall chart for two data instances and reflects the impact of each attribute in its current domain and use-case (here: housing). The approach has been evaluated by 12 selected professional data scientists (out of 33 replies on 200 invitations). It needs to be highlighted, that the target audience is proficient in terms of artifical- and explainable AI - thus, they are more likely to understand a models domain and how the model work within this domain. Nevertheless, the participants agreed on an enjoyable and easy to use experience [22].

4 Case-Based Explanation

Case-Based Reasoning (CBR) is a methodology that reuses knowledge of previous occured situations. In the medical domain, a case can consist of the symptoms of a sickness and the corresponding solution to cure the sickness of a patient. For another patient with similar symptoms, the CBR cycle proposes the most similar case to the new patient and thus can be justified. Nevertheless, some patients might argue that even the most similar case is not similar enough due to the individuality of a human being and the high domain complexity. Nevertheless, the possibility to use CBR as a baseline approach and build explanations upon it has received new attention due to the general interest in XAI.

Due to the lack of reasoning *why* the proposed case is the most similar case, Lamy et al added a visual interface for an explanation on which decision (and thus which therapy) is better for a breast cancer patient [29]. In the healthcare domain, after most of the symptoms have been raised, usually only a few rea-

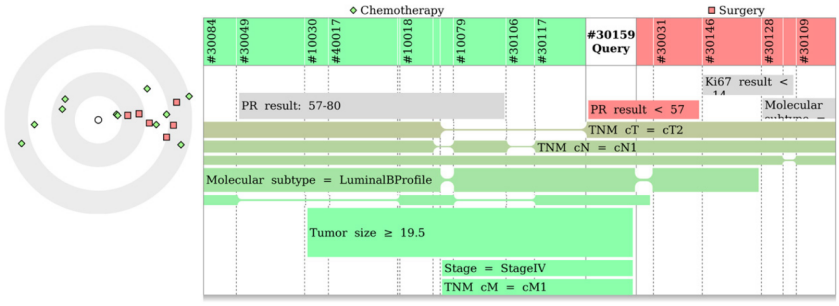

Fig. 3. Screenshot of the visual interface for CBR, proposed by Lamy et al. [29]

sonable options are left to be considered. For each option, a dimension will be opened and the symptoms and their values are weighted accordingly. Whenever a query has been issued to the system, the most similar cases are retrieved. The visual interface uses these data and chooses the most relevant attributes. These will be displayed and ordered in a way so that the user can comprehend the influence to the proposed solutions (the case number next to the query) as proposed in Fig. 3. As mentioned earlier, this is a very similar presentation to the approach in machine learning by Ghosal et al [16]. The approach has been tested on three public datasets and the size of the case base has been limited to 315 cases. It remains to be surveyed how the approach would fare on real datasets and especially a larger case base to increase its proficiency.

Eisenstadt et al deployed explanation patterns using an agent-based system module within a case-based assistance framework to provide human understandable insights on the system behavior in the architectural domain [11]. The underlying data sets are semantic fingerprints of MetisCBR, for example the room count as unconnected vertices in a graph so that the position of the rooms are also known. To model the accessibility of two rooms, these vertices can be connected via a corresponding edge where the edge also represents a possibility to move from one room to another (e. g through a door). These fingerprints then are used to create an explanation using explanation patterns (see Fig. 4) [11].

These two works have been picked as examples for the two current situations which most case-based explanation systems are facing: The first is to manage and combine manifold known and measured attributes in such a way that the user gains trust and can cognitively understand the systems choice of one of many possible medical treatments. Whilst most companies do have the ability to measure attributes, the most natural approach is to use fossiled explanations or canned explanations as suggested back then by R. Schank [41]. Nevertheless the challenge remain to loosen up the rather static nature of predefined templates to generate an explanation on the fly which is a difficult problem to solve, given the huge possibility of valid and invalid combinations which have to be identified. Most approaches identified in this survey using CBE are focusing

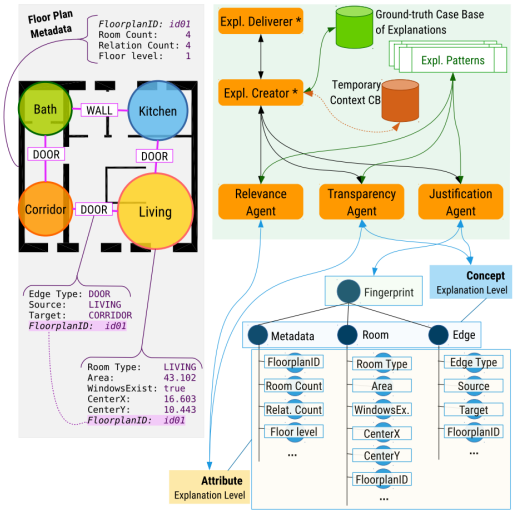

Fig. 4. Domain model, exemplary case, general structure of an explanation tree and the current explainer module, proposed by Eisenstadt et al. [11]

on using CBR with focus of learning the users feedback as an explanation to the given problem. The other challenge which current approaches are rightfully face is to implement explanations into areas where no explanations have been given before. As stated before, the cold start problem can to some extent be avoided by using the knowledge about how to structure the foundation of an explanation-aware system (knowledge management, social interaction), but the domain knowledge still needs to be connected to the explanation component which might be difficult depending on which architecture has been used.

5 Conclusion

As the literature review hints, there are multiple manifold distinct domains in which XAI can be used. Most of these domains have interfaces to benefit from each other. It remains open to find a general valid formulation on what an explanation actually is - despite the efforts of defining it. This remains to be changed for each specific situation a possible user is currently in. But the lack of formalism is probably not even the problem. The main goal is still to explain a given decision to the user individually and the individuality of each user increases the complexity by a large margin. A lot of recent approaches and implementations seem to be promising, yet, actual user-centring results on conducted case-studies are missing. These would be very interesting to measure the acceptance of users and to adjust the direction of developing an XAI accordingly. However, moving from black-box decision making to a glass-box decision making is a step in the

right direction to support the acceptance of AI in the everyday life, including the introduction of IT in schools and other social areas to enable a larger group of people in using the advantages of (X)AI.

References

1. Abdul, A., Vermeulen, J., Wang, D., Lim, B.Y., Kankanhalli, M.: Trends and Trajectories for Explainable, Accountable and Intelligible Systems: An HCI Research Agenda. In: Proceedings of the 2018 CHI Conference on Human Factors in Computing Systems - CHI 18. pp. 118. ACM Press, Montreal QC, Canada (2018). https://doi.org/10.1145/3173574.3174156.
2. Adadi, A., Berrada, M.: Peeking Inside the Black-Box: A Survey on Explainable Artificial Intelligence (XAI). IEEE Access. 6, 5213852160 (2018). https://doi.org/10.1109/ACCESS.2018.2870052.
3. The AlphaStar team: AlphaStar: Mastering the Real-Time Strategy Game StarCraft II. https://deepmind.com/blog/alphastar-mastering-real-time-strategy-game-starcraft-ii/. Last access: 04/22/2019. 24 January 2019.
4. Cellan-Jones, R.: Stephen Hawking warns artificial intelligence could end mankind. URL: https://www.bbc.com/news/technology-30290540. Last access: 04/22/2019. 2014.
5. Binder, A., Bach, S., Montavon, G., Müller, K.-R., Samek, W.: Layer-Wise Relevance Propagation for Deep Neural Network Architectures. In: Kim, K.J. and Joukov, N. (eds.) Information Science and Applications (ICISA) 2016. pp. 913922. Springer Singapore, Singapore (2016).
6. 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 - KDD '15. pp. 17211730. ACM Press, Sydney, NSW, Australia (2015).
7. CD-MAKE-19: Cross Domain Conference for Machine Learning and Knowledge Extraction. CD-MAKE 2019 Workshop on explainable Artificial Intelligence. https://cd-make.net/special-sessions/make-explainable-ai/. Last access: 04/24/2019.
8. Choo, J., Liu, S.: Visual Analytics for Explainable Deep Learning. IEEE Computer Graphics and Applications. 38, 8492 (2018).
9. Conati, C., Porayska-Pomsta, K., Mavrikis, M.: AI in Education needs interpretable machine learning: Lessons from Open Learner Modelling. arXiv:1807.00154 [cs]. (2018).
10. Došilović, F. K., Brčić, M., Hlupić, N.: Explainable artificial intelligence: A survey. 41st International Convention on Information and Communication Technology, Electronics and Microelectronics (MIPRO), Opatija, 2018, pp. 0210-0215. 2018
11. Eisenstadt, V., Espinoza-Stapelfeld, C., Mikyas, A., Althoff, K.-D.: Explainable Distributed Case-Based Support Systems: Patterns for Enhancement and Validation of Design Recommendations. In: Cox, M.T., Funk, P., and Begum, S. (eds.) Case-Based Reasoning Research and Development. pp. 7894. Springer International Publishing, Cham (2018). https://doi.org/10.1007/978-3-030-01081-2_6.
12. European Commission: Artificial intelligence. https://ec.europa.eu/commission/-news/artificial-intelligence-2018-dec-07_en. Last access: 04/22/2019. Published 7 December 2018.

13. Escalante, H.J., Guyon, I., Escalera, S., Jacques, J., Madadi, M., Baro, X., Ayache, S., Viegas, E., Gucluturk, Y., Guclu, U., van Gerven, M.A.J., van Lier, R.: Design of an explainable machine learning challenge for video interviews. In: 2017 International Joint Conference on Neural Networks (IJCNN). pp. 36883695. IEEE, Anchorage, AK, USA (2017).
14. EXTRAAMAS-19: EXplainable TRansparent Autonomous Agents and Multi-Agent Systems. https://extraamas.ehealth.hevs.ch/index.html. Last access: 04/24/2019.
15. Gandhi, P.: Explainable Artificial Intelligence. https://www.kdnuggets.com/2019/01/explainable-ai.html. Last access: 04/22/2019. 2019.
16. Ghosal, S., Blystone, D., Singh, A.K., Ganapathysubramanian, B., Singh, A., Sarkar, S.: An explainable deep machine vision framework for plant stress phenotyping. Proceedings of the National Academy of Sciences. 115, 46134618 (2018).
17. Pagel, P., Portmann, E., Vey, K., (2018). Cognitive Computing - Teil 2. Informatik Spektrum: Vol. 41, No. 2. Berlin Heidelberg: Springer-Verlag. (S. 81-84).
18. Goebel, R., Chander, A., Holzinger, K., Lecue, F., Akata, Z., Stumpf, S., Kieseberg, P., Holzinger, A.: Explainable AI: The New 42? In: Holzinger, A., Kieseberg, P., Tjoa, A.M., and Weippl, E. (eds.) Machine Learning and Knowledge Extraction. pp. 295303. Springer International Publishing, Cham (2018).
Clearance Domain. ICCBR-18 Workshop Proceedings. XCBR: Case-Based Reasoning for the Explanation of Intelligent Systems. 2018.
19. Goodman, B., Flaxman, S.: European Union Regulations on Algorithmic Decision-Making and a "Right to Explanation". AI Magazine. 38, 50 (2017).
20. Guidotti, R., Monreale, A., Ruggieri, S., Turini, F., Pedreschi, D., Giannotti, F.: A Survey Of Methods For Explaining Black Box Models. arXiv:1802.01933 [cs]. (2018).
21. Gunning, D.: Explainable Artificial Intelligence (XAI). 36 (2017).
22. Hohman, F., Head, A., Caruana, R., DeLine, R., Drucker, S.M.: Gamut: A Design Probe to Understand How Data Scientists Understand Machine Learning Models. 13. 2019.
23. Holzinger, A., (2018). Explainable AI (ex-AI). Informatik Spektrum: Vol. 41, No. 2. Berlin Heidelberg: Springer-Verlag. (S. 138-143).
24. Holzinger, A., Kieseberg, P., Weippl, E., Tjoa, A.M.: Current Advances, Trends and Challenges of Machine Learning and Knowledge Extraction: From Machine Learning to Explainable AI. In: Holzinger, A., Kieseberg, P., Tjoa, A.M., and Weippl, E. (eds.) Machine Learning and Knowledge Extraction. pp. 18. Springer International Publishing, Cham (2018).
25. ICCBR-19: 27th International conference on case-based reasoning, September 8-12 in Otzenhausen, Germany. The theme for ICCBR 2019 is Explainable AI. http://iccbr2019.com/. Last access: 04/24/2019.
IJCAI-ECAI-18, Stockholm, Sweden, July 13-19, 2018. XCBR: First Workshop on case-based reasoning for the explanation of intelligent systems. http://gaia.fdi.ucm.es/events/xcbr/. Last access: 04/24/2019.
26. ISWC-10: SEMEX 2019: 1st Workshop on Semantic Explainability colocated with the 18th International Semantic Web Conference (ISWC 2019). https://scdemo.techfak.uni-bielefeld.de/semex2019/. Last access: 04/24/2019.
27. Kirsch, A.: Explain to whom? Putting the User in the Center of Explainable AI. Proceedings of the First International Workshop on Comprehensibility and Explanation in AI and ML 2017 co-located with 16th International Conference of the Italian Association for Artificial Intelligence (AI*IA 2017), 2017, Bari, Italy.

28. Lakkaraju, H., Kamar, E., Caruana, R., Leskovec, J.: Interpretable & Explorable Approximations of Black Box Models. arXiv:1707.01154 [cs]. (2017).
29. Lamy, J.-B., Sekar, B., Guezennec, G., Bouaud, J., Sroussi, B.: Explainable artificial intelligence for breast cancer: A visual case-based reasoning approach. Artificial Intelligence in Medicine. 94, 4253 (2019).
30. Lundberg, S.M., Nair, B., Vavilala, M.S., Horibe, M., Eisses, M.J., Adams, T., Liston, D.E., Low, D.K.-W., Newman, S.-F., Kim, J., Lee, S.-I.: Explainable machine-learning predictions for the prevention of hypoxaemia during surgery. Nature Biomedical Engineering. 2, 749760 (2018).
31. Madumal, P., Miller, T., Vetere, F., Sonenberg, L.: Towards a Grounded Dialog Model for Explainable Artificial Intelligence. arXiv:1806.08055 [cs]. (2018).
32. Synced: NIPS Tickets Sell Out in Less Than 12 Minutes. https://medium.com/syncedreview/nips-tickets-sell-out-in-less-than-12-minutes-e3aab37ab36a. Last access: 04/23/2018. 2018.
33. Mittelstadt, B., Russell, C., Wachter, S.: Explaining Explanations in AI. Proceedings of the Conference on Fairness, Accountability, and Transparency - FAT* '19. 279288 (2019).
34. Musk, E., Isaacson, W.: Elon Musk: Artificial Intelligence Could Wipe Out Humanity. URL: https://sagaciousnewsnetwork.com/elon-musk-artificial-intelligence-could-wipe-out-humanty/. Last access: 04/22/2019. 2014.
35. Musk, E.: AI "vastly more risky than North Korea" URL: https://twitter.com/elonmusk/status/896166762361704450. Last access: 04/22/2019. 2017.
36. NeurIPS-19: NeurIPS 2019 Expo Workshop: Fairness and Explainability: From ideation to implementation. https://nips.cc/Expo/Conferences/2018/Schedule?workshop_id=5. Last access: 04/24/2019.
37. Nushi, B., Kamar, E., Horvitz, E.: Towards Accountable AI: Hybrid Human-Machine Analyses for Characterizing System Failure. 10. 2018
38. Peltola, T.: Local Interpretable Model-agnostic Explanations of Bayesian Predictive Models via Kullback-Leibler Projections. arXiv:1810.02678 [cs, stat]. (2018).
39. Reda, J.: Unofficial consolidated version: trialogue outcome. Article 13 + related definition. 2019.
40. Schank, R. C.: Explanation Patterns: Understanding Mechanical and Creatively. L. Erlbaum Assoc. Inc., Hillsdale, NJ, USA. 1986.
41. Schank, R.C., Leake, D.B.: Creativity and learning in a case-based explainer. Artificial Intelligence. 40, 353385 (1989).
Investigating the solution space for online iterative explanation in goal reasoning agents. AI Commun. 31(2): 213-233 (2018)
42. Vincent, J.: Europes controversial overhaul of online copyright receives final approval. https://www.theverge.com/2019/3/26/18280726/europe-copyright-directive. Last access: 04/22/2019. 2019.
43. Wang, D., Yang, Q., Abdul, A., Lim, B.Y.: Designing Theory-Driven User-Centric Explainable AI. 15. 2019.
44. Waltl, B., Vogl, R.: Explainable Artificial Intelligence - The new frontier in legal informatics. 10. 2018.
45. Weizenbaum, J.: ELIZA–A Computer Program For the Study of Natural Language Communication Between Man and Machine. Communications of the ACM. 7. 1966.
46. Zhang, Y., Chen, X.: Explainable Recommendation: A Survey and New Perspectives. arXiv:1804.11192 [cs]. (2018).