Examination of Current AI Systems within the Scope of Right to Explanation and Designing Explainable AI Systems*

Orhan Gazi Yalçın[1**][0000−0001−7990−6531]

PhD candidate of LAST-JD-RIoE,
University of Bologna (beneficiary),
The Technical University of Madrid,
University of Turin
orhangazi.yalcin2@unibo.it
https://last-jd-rioe.eu/

Abstract. This research aims to explore explainable artificial intelligence, a new sub field of artificial intelligence that is gaining importance in academic and business literature due to increased use of intelligent systems in our daily lives. As part of the research project, first of all, the necessity of the explainability in AI systems will be explained in terms of accountability, transparency, liability, and fundamental rights & freedoms. The latest explainable AI algorithms introduced by the AI researchers will be examined firstly from technical and then, from legal perspectives. Their statistical and legal competencies will be analyzed. After detecting the deficiencies of the current solutions, a comprehensive and technical AI system design will be proposed which satisfies not only the statistical requisites; but also the legal, ethical, and logical requisites.

Keywords: right to explanation · GDPR · explainable AI · XAI · interpretable machine learning · artificial intelligence · machine learning · accuracy-explainability.

1 Introduction

Artificially intelligent (AI) systems offer many benefits to the individuals and the public & private institutions. Thanks to AI systems and software automation, the services which require a high level of human involvement may be provided quickly with low to none human involvement using machine learning. With the help of the applied statistics and affordable computing power, engineers can develop AI

*. This project has received funding from the European Union's Horizon 2020 research and innovation programme under the Marie Skłodowska-Curie ITN EJD "Law, Science and Technology Rights of Internet of Everything" grant agreement No 814177.

systems to complete difficult tasks such as designing driverless cars, building machine translation softwares, or developing algorithmic profiling systems.[1]

Since the primary goal of the AI systems is to increase efficiency and accuracy,[2] machine learning engineers often overlook the explainability of their systems. The assumption of an engineer tends to be that as long as the model accurately predicts the result of a future event, the relevant parties will remain satisfied. Even though this assumption is partially correct, it is certainly deficient. Whenever a decision of the AI system harms a third party or the service receiver due to an incorrect prediction, there may be liabilities and obligations as well as violations of fundamental rights & freedoms. In these situations, the reasoning of the AI systems will be crucial to understand the logic behind the incorrect prediction and to decide on the liability of the parties. For instance, in a recent study in the U.S. Fintech sector, the researchers found that mortgage refinancing algorithms used in the U.S. -as well as the professionals in this field- discriminate against Latin and African American borrowers.[3] Even the legitimate-business-necessity interpretation is taken into account,[4] the research shows that at least 6% of the minority applications are rejected due to purely discriminatory practices.[5] The credit application example is only the tip of the iceberg. In the near future, armed UAVs with AI systems, AI judges, and AI police bots will take over the corresponding traditional jobs where accountability and liability are a significant part of the process. They will make decisions in irreversible matters which involve fundamental rights and freedoms such as right to live, right to bodily integrity, and right to freedom.[6] Therefore, discriminatory or incorrect decisions may cause significant material and moral damages.[7]

---

1. Amitai Etzioni and Oren Etzioni, "Keeping AI Legal," *Vand. J. Ent. & Tech. L.* XIX, no. 1 (2007): 2, https://perma.cc/UQ7F-7VYX]..

2. Pedro Domingos, "A few useful things to know about machine learning," *Communications of the ACM* 55, no. 10 (2012): 3, ISSN: 00010782, doi:10.1145/2347736.2347755.

3. Robert P. Bartlett et al., "Consumer Lending Discrimination in the FinTech Era," *SSRN Electronic Journal*, November 2017, 1, doi:10.2139/ssrn.3063448, https://dx.doi.org/10.2139/ssrn.3063448.

4. Ibid., 3-4.

5. Ibid., 21.

6. Wojciech Samek, Thomas Wiegand, and Klaus-Robert Müller, "Explainable Artificial Intelligence: Understanding, Visualizing and Interpreting Deep Learning Models," *ITU Journal: ICT Discoveries* 1, no. Special Issue 1 (2017): 2-3, arXiv: 1708.08296, http://arxiv.org/abs/1708.08296.

7. Sina Aghaei, Mohammad Javad Azizi, and Phebe Vayanos, "Learning Optimal and Fair Decision Trees for Non-Discriminative Decision-Making," *Proceedings of the AAAI Conference on Artificial Intelligence* 33 (2019): 1-3, ISSN: 2159-5399, doi:10.1609/aaai.v33i01.33011418, arXiv: 1903.10598, www.aaai.org.

2 The Problem

There are a number of different algorithms that may be used for machine learning which have different levels of success on the accuracy metrics. Although the traditional algorithms are highly interpretable, AI engineers are likely to prefer deep learning algorithms over traditional algorithms due to the high performance of deep learning algorithms on accuracy metrics (see Figure 1[8]).

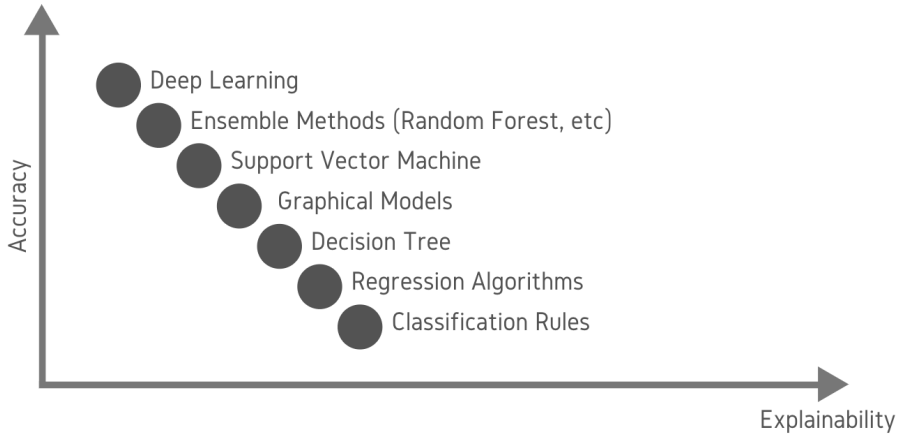

Fig. 1. Accuracy-Explainability Plot of Various AI Algorithms[9][10]

In other words, machine learning algorithms that are highly explainable usually have low accuracy performance in a relative sense, especially when there is an abundance of data. Therefore, as long as there is not a constraint on computational power and there is enough data, ML engineers tend to select models with higher accuracy while ignoring their low-level explainability. In addition, the popularity of predefined machine learning libraries (e.g. Keras, Tensorflow, PyTorch, Scikit Learn) also contributes to the wide-spread use of black-box models and to the negligence of the explainability in the AI systems.[11]

If a new wave of research does not solve the reverse relationship observed between explainability and accuracy, in the near future, AI judges, soldiers, armed drones, police officers, and other sensitive AI systems must have to use the algorithms with high accuracy; therefore, with low explainability. The low level of AI system explainability will certainly be problematic to secure the right to explanation, particularly in areas such as transportation, security, medicine, finance,

8. Preet Gandhi, *Explainable Artificial Intelligence*, 2019, accessed November 12, 2019, `https://www.kdnuggets.com/2019/01/explainable-ai.html`.

11. Bernhard Waltl and Roland Vogl, "Explainable Artificial Intelligence-the New Frontier in Legal Informatics," *Jusletter IT*, 2018, 3.

legal, and military.[12] One may argue that the right to explanation is not a widely accepted and essential right at the moment. However, it is not hard to foresee that with the new advancements in technology, the significance of this right will gain momentum and it will soon become part of the fundamental rights & freedoms. For instance, in the U.S., credit scoring decisions must already be given with reasoning; therefore, algorithms used for credit scoring must be explainable.[13] On the other hand, the decisions made in these fields may constitute a violation of the traditional fundamental rights & freedoms as well. For instance, the decision of an AI judge without reasoning -regardless of its accuracy- will violate the right to a fair trial.[14]

3 Current Developments in Explainable AI from Legal and Technical Perspectives

The knowledge in the novel field of explainable AI may be evaluated from two different perspectives: (i) legal perspective and (ii) data science perspective. From the legal perspective, the significance of the explainability of the AI systems has already caught the attention of European and American lawmakers to some extent. As mentioned above, the right to explanation in credit scoring has been a long-standing right in the U.S. On the European side, with the EU General Data Protection Regulation (GDPR) adopted in 2016, the right to explanation was strengthened with Art. 13-15 of the GDPR -which all read "The data subject shall have... access to ... the existence of automated decision- making... " and "... meaningful information about the logic involved"-.[15] On most occasions, the party influenced by the AI system is not aware of the parameters used in the model and the sampling of the data (train and test data) as well as which parameter is given more weight for the prediction. The E.U. and the U.S. have the aforementioned set of infant rules to mitigate this problem. As mentioned above, fundamental rights and freedoms such as the right to a fair trial or right to live may also be used as a shield against unfair practices in AI systems.

From the data science perspective, designing explainable AI systems is also significant to recognize cause and effect relationships to improve existing systems. In a competition held by The Defense Advanced Research Projects Agency (DARPA), nine U.S. universities in partnership with industrial players and/or

12. Matt Turek, *Explainable Artificial Intelligence (XAI)*, 2016, 1, accessed November 12, 2019, `https://www.darpa.mil/program/explainable-artificial-intelligence`.

13. Jiahao Chen, "Fair lending needs explainable models for responsible recommendation," *FATREC 2018*, 2018, 2, arXiv: `1809.04684v1`.

14. European Court of Human Rights, *Guide on Article 6: Right to a Fair Trial (Criminal Limb)*, technical report (2013), 32, `https://perma.cc/C4XN-AE8N`.

15. Bryce Goodman and Seth Flaxman, "European union regulations on algorithmic decision making and a "right to explanation"," *AI Magazine*, 2017, 6, ISSN: 07384602, doi:`10.1609/aimag.v38i3.2741`, arXiv: `1606.08813`.

European universities have proposed novel explainable AI systems. These participant universities have suggested explainable learners (a combination of explainable models and explanation interfaces) in addition to their research on the psychological model of explanation. Their systems may focus on one of these three subcategories: (i) Deep Explanation, (ii) Interpretable Models, (iii) Model Induction. Briefly, Deep explanation teams aim to develop modified deep learning models in which explainable features may be extracted. Interpretable model teams focus on traditional & causal methods (e.g. And-Or grammars, Hierarchical Bayesian Networks, and Random Forests) and try to come up with more explainable models (more structured, interpretable, and causal). Finally, Model induction teams try to induce novel models by testing the black box models.[16]

CP	Performer	Explainable Model	Performer
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

Fig. 2. The Ongoing XAI Researches in the U.S.[17]

3.1 Previous Studies

The motivation of the previous studies revolves around the different stakeholders or the audience of the explainable AI field who are (i) data subjects, (ii) domain experts, (iii) data scientists & developers, (iv) company managers, and finally, (iv) regulatory entities. The overall literature shows that stakeholders seek different explainability components in their AI systems. While domain experts look for trustworthiness, transferability, and confidence; regulatory entities and the data subjects seek causality, policy awareness, and fairness. On the other hand,

16. David Gunning, *Explainable Artificial Intelligence (XAI)*, technical report (2017), 10-18, `https://www.darpa.mil/attachments/XAIProgramUpdate.pdf`.

informativeness is sought by all the stakeholders. The research will analyse the cluster of the previous explainable AI literature to clarify the needs of the stakeholder.[18]

4 Preliminary Ideas and Proposed Approach for Designing an Explainable AI System

Ideally, when an AI system is used for matters which involve public services, justice, and other heavily regulated areas, the system must be transparent and explainable.[19] In addition to the explainable nature of the machine learning models, this also means that the feature selection process must be accessible publicly or upon request. Furthermore, sampling principles of the training and test datasets must be very well explained. Finally, the design of the model must be explainable by nature which indicates both the use of explainable machine learning algorithms and also the availability of an interface with which the administrator of the AI system or the relevant authorities can analyze the results.

To have a truly explainable system, the scope of explainability must also be examined from a legal standpoint as well as a linguistic standpoint. First of all, this research will address the differences between the concepts of explainability and interpretability which the researchers in the field often use interchangeably. Due to the high increase in the number of publications in the field, the cluster of the explainable AI literature created its own nomenclature with a variety of adjacent terms including, but not limited to, understandability, intelligibility, comprehensibility, transparency as well as interpretability and explainability. This research will also examine these terms in terms of their similarities and differences to clarify the fundamental concepts of this field.[20]

While these concepts shape the nature and the requisites of comprehensive AI explainability, there is more than one way achieving explainability. These techniques may be grouped under two main categories: (i) Ante-hoc approaches and (ii) Post-hoc approaches.[21] Ante-hoc approaches aim to achieve the explainability with the model design and other assisting methods whereas the post-hoc approaches aim to extract explanations from existing models.[22] A third approach

18. Alejandro Barredo Arrieta et al., *Explainable Artificial Intelligence (XAI): Concepts, Taxonomies, Opportunities and Challenges toward Responsible AI*, technical report (2019), 8-10, arXiv: 1910.10045v2.

19. Matt Turek, *Explainable Artificial Intelligence (XAI).*

20. Barredo Arrieta et al., *Explainable Artificial Intelligence (XAI): Concepts, Taxonomies, Opportunities and Challenges toward Responsible AI*, 5.

21. Kacper Sokol and Peter Flach, "Explainability Fact Sheets: A Framework for Systematic Assessment of Explainable Approaches," *FAT* 2020 - Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency*, December 2019, 57-58, doi:10.1145/3351095.3372870, arXiv: 1912.05100, http://arxiv.org/abs/1912.05100%20http://dx.doi.org/10.1145/3351095.3372870.

22. Barredo Arrieta et al., *Explainable Artificial Intelligence (XAI): Concepts, Taxonomies, Opportunities and Challenges toward Responsible AI*, 10-12.

which is usually considered within the post-hoc approaches is the global mimic approach which aims generate models that mimic the overall behavior of complex black-box models.[23]

As mentioned above, the entirety of AI system components (the algorithm, selection of the features, sampling of the train data, test data, and validation data, and all the other aspects) must be transparent and interpretable for a comprehensively explainable system. The explainable nature of the AI system must also correspond to the legal requirements such as accountability and liability. The overall model design contributes to the explainability of all machine learning models. After setting the clear boundaries of the acceptable framework for legally and technically acceptable explainability, the research will move towards examining these approaches and concepts to achieve the desired explainability goals as per the designed framework requires. This algorithmic explainability research will; on the other hand, specifically target the deep learning models because of their increasing popularity.

4.1 Proposed Approach

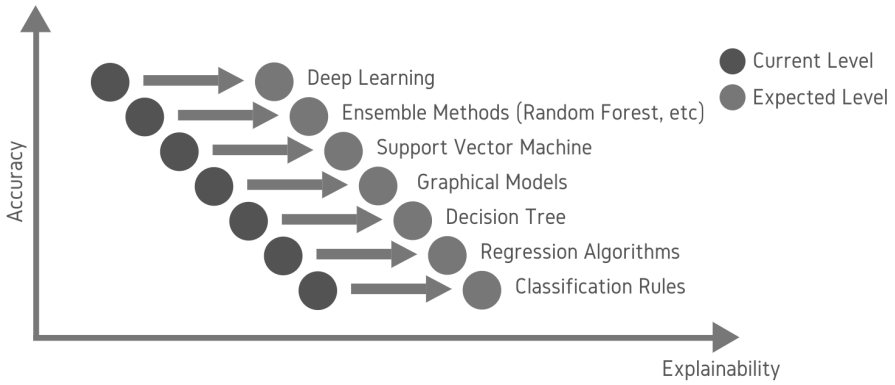

Fig. 3. The Shift Aimed with the Research on the Accuracy - Explainability Plot

The proposed research will start with a scope definition. The legal nature of the problem, concepts of explainability, accountability, and liability and the relationship between them will be clearly defined. The potential threats on fundamental rights & freedoms which the AI systems may carry in their nature will be scrutinized. Then, the existing AI systems will be evaluated based on their accuracy and explainability performances. Finally, the advancements observed in the novel explainable AI systems will be observed, analyzed, and examined. The disadvantages and the advantages of these novel systems will be regarded as a beacon for

23. Sokol and Flach, "Explainability Fact Sheets: A Framework for Systematic Assessment of Explainable Approaches," 58.

designing cutting edge explainable AI systems. Each part of the research will be crucial to design unique explainable AI systems and to offer remedies to increase the explainability of the current machine learning algorithms in which both the research made in legal and technical domains will be taken into account. The workflow process of the research methodology may be visualized as follows:

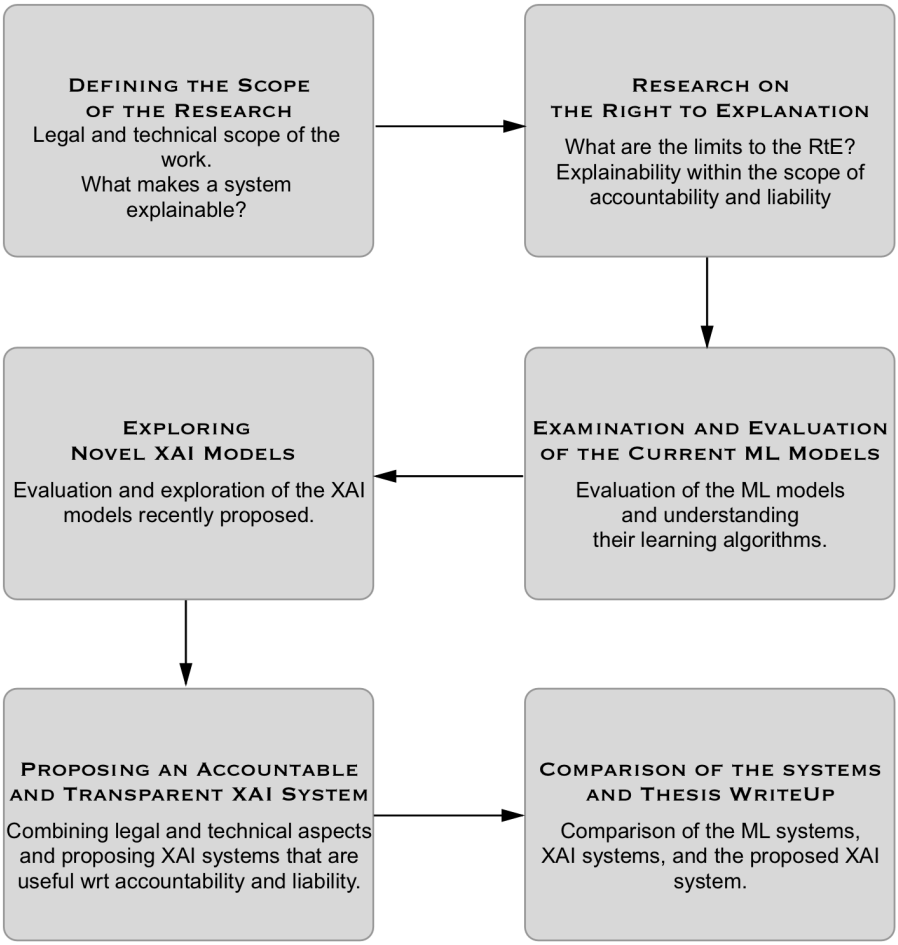

Fig. 4. A Visual Demonstration of the Research Methodology

5 The Contribution and the Unique Nature of the Research

There are two main categories observed when the previous studies are analyzed. Studies that are in the first group are done by the legal scholars analyzing the

possible negative effects of the wide-spread use of AI systems that are not explainable in sensitive domains. Since the usage of AI systems in sensitive fields may raise questions regarding violation of fundamental rights & freedoms and may create significant damages due to incorrect decisions, determining the accountability and the liability of the parties are very significant. Therefore, the legal scholars focus on the accountability and liability of the parties when damages are suffered and rights are violated.

On the other hand, the second group of studies are conducted by machine learning and data science expert and focuses on the statistical analysis of the AI systems, transferability of the trained models to new areas, and cause-and-effect relationships between explanatory and response variables. These groups of researchers focus on understanding the decision-making process of the trained AI system.[24] However, legal reasoning might have to contain more information regarding the event than a technical expert foresees.

Therefore, there must be a bridge between the legal scope of explainability and the field of explainable artificial intelligence. Only with this bridge study, the expectations of the public and law community may be met by the technical researchers. Therefore, this research will act as a bridge between the expectations of the public, the law community and the works of the technical experts building meaningful explainable AI systems.

6 Conclusion

Utilizing the increased network connectivity (thanks to the Internet), robotics & software automation, and cheap computing power, humanity is entering into an era where the mainstreamed and repetitive tasks are fully automated with artificially intelligent systems. Large enterprises and governments has already utilized intelligent systems in many of their tasks. However, this is still the beginning of the AI era. With the advancements in machine learning, Intelligent systems will increasingly be used in sensitive tasks. Therefore, the decisions of the Intelligent systems will be subject to many civil and penal disputes. Therefore, explaining the decision making mechanism of these systems (i.e., explainability) will be a very crucial component for securing the justice in a healthy society. This research aims to satisfy this need by approaching it from a law-oriented perspective as well bearing the technical side in mind. By reviewing the latest academic and business literature and by experimenting on the current explainable AI and AI models, legally acceptable XAI systems will be developed and presented.

7 Acknowledgements

This project has received funding from the European Union's Horizon 2020 research and innovation programme under the Marie Skłodowska-Curie ITN EJD

24. Barredo Arrieta et al., *Explainable Artificial Intelligence (XAI): Concepts, Taxonomies, Opportunities and Challenges toward Responsible AI*, 8-10.

"Law, Science and Technology Rights of Internet of Everything" grant agreement No 814177.

References

Aghaei, Sina, Mohammad Javad Azizi, and Phebe Vayanos. "Learning Optimal and Fair Decision Trees for Non-Discriminative Decision-Making." *Proceedings of the AAAI Conference on Artificial Intelligence* 33 (2019): 1418–1426. ISSN: 2159-5399. doi:10.1609/aaai.v33i01.33011418. arXiv: 1903.10598. www.aaai.org.

Barredo Arrieta, Alejandro, Natalia Díaz-Rodríguez, Javier Del Ser, Adrien Bennetot, Siham Tabik, Alberto Barbado, Salvador Garcia, et al. *Explainable Artificial Intelligence (XAI): Concepts, Taxonomies, Opportunities and Challenges toward Responsible AI.* Technical report. 2019. arXiv: 1910.10045v2.

Bartlett, Robert P., Adair Morse, Richard Stanton, and Nancy Wallace. "Consumer Lending Discrimination in the FinTech Era." *SSRN Electronic Journal*, November 2017. doi:10.2139/ssrn.3063448. https://dx.doi.org/10.2139/ssrn.3063448.

Chen, Jiahao. "Fair lending needs explainable models for responsible recommendation." *FATREC 2018*, 2018. arXiv: 1809.04684v1.

Domingos, Pedro. "A few useful things to know about machine learning." *Communications of the ACM* 55, no. 10 (2012): 78–87. ISSN: 00010782. doi:10.1145/2347736.2347755.

Etzioni, Amitai, and Oren Etzioni. "Keeping AI Legal." *Vand. J. Ent. & Tech. L.* XIX, no. 1 (2007): 133–146. https://perma.cc/UQ7F-7VYX]..

European Court of Human Rights. *Guide on Article 6: Right to a Fair Trial (Criminal Limb).* Technical report. 2013. https://perma.cc/C4XN-AE8N.

Gandhi, Preet. *Explainable Artificial Intelligence*, 2019. Accessed November 12, 2019. https://www.kdnuggets.com/2019/01/explainable-ai.html.

Goodman, Bryce, and Seth Flaxman. "European union regulations on algorithmic decision making and a "right to explanation"." *AI Magazine*, 2017. ISSN: 07384602. doi:10.1609/aimag.v38i3.2741. arXiv: 1606.08813.

Gunning, David. *Explainable Artificial Intelligence (XAI).* Technical report. 2017. https://www.darpa.mil/attachments/XAIProgramUpdate.pdf.

Matt Turek. *Explainable Artificial Intelligence (XAI)*, 2016. Accessed November 12, 2019. https://www.darpa.mil/program/explainable-artificial-intelligence.

Samek, Wojciech, Thomas Wiegand, and Klaus-Robert Müller. "Explainable Artificial Intelligence: Understanding, Visualizing and Interpreting Deep Learning Models." *ITU Journal: ICT Discoveries* 1, no. Special Issue 1 (2017): 1–10. arXiv: 1708.08296. http://arxiv.org/abs/1708.08296.

Sokol, Kacper, and Peter Flach. "Explainability Fact Sheets: A Framework for Systematic Assessment of Explainable Approaches." *FAT* 2020 - Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency*, December 2019, 56–67. doi:10.1145/3351095.3372870. arXiv: 1912.05100. http://arxiv.org/abs/1912.05100%20http://dx.doi.org/10.1145/3351095.3372870.

Waltl, Bernhard, and Roland Vogl. "Explainable Artificial Intelligence-the New Frontier in Legal Informatics." *Jusletter IT*, 2018.