EXPLAINABLE ARTIFICIAL INTELLIGENCE – THE NEW FRONTIER IN LEGAL INFORMATICS

Bernhard Waltl / Roland Vogl

Research Associate, Technical University of Munich, Department of Informatics,
Software Engineering for Business Information Systems
Boltzmannstraße 3, 85748 Garching bei München, DE
b.waltl@tum.de; http://wwwmatthes.in.tum.de

Executive Director of CodeX – the Stanford Center for Legal Informatics, and
Executive Director of the Stanford Program in Law, Science and Technology
Crown Quadrangle, 559 Nathan Abbott Way, Stanford, CA 94305-8610, USA
rvogl@law.stanford.edu; http://codex.stanford.edu/

Keywords: ***Explainable artificial intelligence, next-generation decision support systems, AI governance, interpretation, explanation, algorithmic decision-making***

Abstract: *In recent years, mainstream media coverage on artificial intelligence (AI) has exploded. Major AI breakthroughs in winning complex games, such as chess and Go, in autonomous mobility, and many other fields show the rapid advances of the technology. AI is touching more and more areas of human life, and is making decisions that humans frequently find difficult to understand. This article explores the increasingly important topic of «explainable AI» and addresses the questions why we need to build systems that can explain their decisions and how should we build them. Specifically, the article adds three additional dimensions to capture transparency to underscore the tremendous importance of explainability as a property inherent to machine learning algorithms. It highlights that explainability can be an additional feature and dimension along which machine learning algorithms can be categorized. The article proposes to view explainability as an intrinsic property of an AI system as opposed to some external function or subsequent auditing process. More generally speaking, this article contributes to legal informatics discourse surrounding the so-called «third wave of AI» which leverages the strengths of manually designed knowledge, statistical inference, and supervised machine learning techniques.*

1. Introduction

Artificial Intelligence (AI) covers a broad range of concepts and terms, and its essence is hard to define. In 1978 Bellman defined AI as the «[The automation of] activities that we associate with human thinking, activities such as decision-making, problem solving, learning …» (Bellman, 1978). While this is a very broad and abstract definition, it is still valid today. Russel and Norvig differentiated a variety of concurrent and complementary definitions, which are organized along the categories of thinking like a human, thinking rationally, acting like a human, acting rationally (Russell & Norvig, 2009). The sheer explosion of AI systems and Algorithmic Decision-Making (ADM) affecting our daily-lives can be traced back to the following three developments:

1. Data that are used especially for training and evaluation of machine learning based systems are more easily available in digital form
2. High performance computing infrastructure
3. Efficient algorithms that can automate increasingly complex tasks

Section 2 of this article will provide a discussion of the meaning of explainability and transparency. These concepts are then woven into our discussion of ADM in Section 3. We will show that explainability of ADM has to be viewed as an intrinsic property of an ADM system, much as the system's performance is an intrinsic property of the system, with commonly accepted metrics for evaluation (see Sections 3.2. and 3.3.). Finally, the paper will briefly discuss explainability along the lifecycle of an ADM procedure, which goes beyond the application of a machine learning classifier including also the data acquisition and pre-processing phases (see Section 4).

1.1. AI in the Legal Domain and the Expanding Use of Algorithmic Decision Making

The idea of formalizing decision-making processes so that they can be automated by algorithms has been an appealing idea for many legal scholars and practitioners for a long time. The field of legal informatics has concerned itself with the many questions surrounding uses of AI in legal settings since its early days. The International Association for Artificial Intelligence and Law and other AI communities have substantially increased the understanding of possibilities, challenges, and limitations of AI applications in the legal field. According recent scholarship in the field, we can generally distinguish the following different of AI reasoning approaches: Deductive reasoning (legal expert systems, classical logic programming, etc.), Case-based reasoning (induction of rules based on prior cases and precedents, etc.), Abductive reasoning (semantic entailment, finding simple and likely explanations, etc.), Defeasible reasoning (non-monotonic logics, and argumentation, etc.), Probabilistic reasoning (fuzzy logic, reasoning on indeterminate and vague terms, etc.), Reasoning on ontologies (formal knowledge representations, semantic web, OWL, etc.), Statistical reasoning including machine learning approaches (un-/supervised ML, etc.), Advanced machine learning (active, interactive, and reinforcement learning, etc.) (Ashley, 2017; Bench-Capon et al., 2012; Russell & Norvig, 2009). In recent years, AI research based on statistics, machine learning and data mining has exploded. Among other uses, these techniques are specifically leveraged for predictive analytics purposes. Companies in a number of different industries, such as advertising, financial services, insurance, telecommunication, logistics, health care are using predictive models to gain strategic advantage over their competition. Software-supported forecasting of legal decisions (Waltl, Bonczek, Scepankova, Landthaler, & Matthes, 2017), and numerous other applications of artificial intelligence to predict legal outcomes have been launched in recent years around the world (Vogl, 2017).

Already, in the heydays of legal expert systems in the 1990s, Herbert Fiedler observed that there are almost mythological expectations when it comes to algorithmic decision-making (Fiedler, 1990). For the field of legal expert systems, he discussed six distinct expectations that people have with regard to ADM, including the following two pertaining to the concept of explainability:

1. The ability to easily understand and follow knowledge representation, and
2. The explainability and transparency of decisions.

Today, little attention is given to the possibilities of using pre-defined rules based on deductive and ontological reasoning techniques that are the main techniques underlying legal expert systems as a means for explaining automated decision-making processes. Current research on ADM is more focused on the techniques of the machine learning field, rather than the more static and human engineered decision structures of legal expert systems. Nevertheless, it is worth noting that handcrafted reasoning tools, such as the Oracle Policy Automation tool[1], are used by organizations all around the world to represent decision structures and to enable automated (legal) reasoning. But, as Fiedler points out, even the decisions of those hand-crafted systems are hard to understand and explain. In light of that, and in light of the prevalence of the opaquer machine-learning based automated decision-making systems, more attention to the actual algorithmic processing underlying an

[1] https://www.oracle.com/applications/oracle-policy-automation/index.html (all websites last accessed on 12 January 2018).

automated decision is warranted. Consequently, we are focusing our attention on providing a more nuanced view of explanation representations.

1.2. The Need for More Algorithmic Transparency

At their core, AI-based systems implement highly formalized processes while avoiding ambiguities and uncertainties. However, the complexity of these systems and their internal structure makes their inner workings difficult to understand even for experts. This is troubling as more and more advanced AI systems are making decisions that affect our daily lives, including decisions on our credit score, on recruiting, health care related decisions, transportation (including autonomous driving), logistics, and many more areas. In light of a growing list of examples of AI systems that produced unfair or discriminatory outcomes, public discourse about AI has recently turned from euphoria to concern (e.g., Cathy O'Neill, 2016). These cases have made clear that we can no longer view ADM systems as mysterious black-boxes, and in order to earn a human's trust, the functioning of ADM systems will have to be understandable. It should be noted that explainable artificial intelligence does not mean the same as having systems that prevent discrimination in algorithmic decision making. Instead it represents a new property of machine learning techniques. One could of course argue that explanations are required to determine whether discrimination occurred. However, answering that question also requires an analysis of attributes and properties of the term «discrimination».

1.3. The Opacity of Algorithmic-Decision-Making Software

From an academic computer science perspective, the currently increased public awareness around artificial intelligence on one hand has the potential to promote innovation and new research in the field. On the other hand, the danger of heightened expectations is of course that if expectations are not met, we may again enter a period of reduced interest and investment in artificial intelligence, or in other word another «AI winter»[2] might be on the horizon.

We believe that part of the media frenzy around AI is due to the opaque and «black-box»-like nature of AI methods and procedures, including methods that are mathematically complex and encapsulated in «ready-to-use» software packages such as MLib of Apache Spark[3], Scikit-learn[4], or Deeplearning4j[5]. These are commonly used to train, test, and evaluate the performance of a machine learning classifier that is subsequently used to predict the outcome of a given task, such as classification of images, categorization of text, the next move in a chess game, or the likelihood of winning a trial.

These software modules operate in a way that suggests even to the human expert user that something magical is happening once a classifier is trained and it can by itself produce seemingly intelligent answers. As long as these software components are viewed as black-boxes with easy to use interfaces, the sentiment that humanly unexplainable things are happening inside the box is perpetuated. However, because machine learning and other forms of artificial intelligence rely on the application of mathematical insights and processes, such as optimization operations, or Naïve Bayes probabilities, their internal procedures are in fact fully deterministic[6] and can at least in principle be reproduced by humans. However, because many of these machine learning algorithms deal with high-dimensional data and a representation of the features that are not commonly understood by humans, the mere inspection of the classifier might not provide answers in in the form that humans would expect. The insights that one can gain by analyzing the machine learning classifier – by for example

[2] Not too long ago the AI winter was a consequence of the highly raised expectations that could not be met by technology. AI winter is a metaphor and describes a period of reduced funding and interest in AI. The period includes a lack of trust in AI technology due to broken promises and only partially fulfilled expectations.

[3] https://spark.apache.org/mllib/.

[4] http://scikit-learn.org/stable/.

[5] https://deeplearning4j.org/.

[6] Unless no random variables are used within the classification process.

looking at the internal states representing a combination of the classifier type and the training data – are in a form is not suitable to be interpreted by humans. This in turn begs the question as to how these internal states can be represented in a way that humans can still understand and interpret the internal structure of artificially intelligent systems. Finding a way that allows us to interpret the internal structure of AI systems, we believe, will contribute to a broader acceptance and a better understanding of the potential and limitations of AI and algorithmic decision making.

In recent years, we have witnessed the birth of the subfield of interpretable machine-learning. We are already seeing important advances in the field. Jung et al., for example, proposed simple rules for bail decisions[7]. Corbett-Davies et al. recently discussed the fairness of such algorithms[8]. Legislators are also working towards building legal frameworks that can help prevent algorithms from creating undesired results, such as discrimination[9]. Latest discussions in politics and societal pressure even caused legislators to examine the potentials of governing ADM in the field of discrimination. This can, for example, be observed in Germany, the ministry of justice announced a project on a feasibility study on governing ADM[10].

1.4. The Role of Explanations within a Complex Process

There are commonly adapted frameworks to describe the phases involved in creating a system, which can decide autonomously. Figure 2 shows a process, which consists of five subsequent phases. This represents a model process for illustration purposes, which does not contain any iterations and feedback loops. In practical applications the process is typically more complex (Waltl, Landthaler, et al., 2017). During each phase, activities are automatically performed and intermediate results are produced that influence the overall ADM.

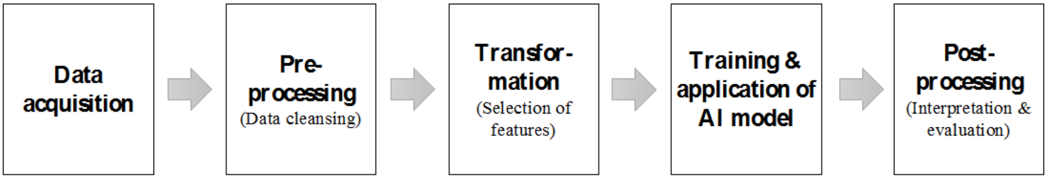

Figure 1: ADM is a complex process of at least five different phases (Fayyad, Piatetsky-Shapiro, & Smyth, 1996).

However, when it comes to explaining the ADM, it is not sufficient to look at the classification result on one instance only. In order to fully understand an automated decision, the whole process needs to be considered. Otherwise the explanation might not be representative and might not reflect all input factors and parameters which led to a particular decision (see Section 3.2.).

[7] https://hbr.org/2017/04/creating-simple-rules-for-complex-decisions.

[8] https://www.nytimes.com/2017/12/20/upshot/algorithms-bail-criminal-justice-system.html.

[9] This is strongly emphasized within Article 21 and 22 of the General Data Protection Regulation (GDPR, Regulation (EU) 2016/679) have been interpreted to provide for a «right to explanation» with regard to algorithmic decision-making that affects an individual. «The data subject shall have the right not to be subject to a **decision based solely on automated processing**, including profiling, […].» and «[…] obtain an **explanation of the decision reached** after such assessment and to challenge the decision.» There is an ongoing debate around the meaning of this rule and to this date there does not seem to be an agreement on the potential consequences of this rule, and how it should be implemented. Some commentators call Article 21 «toothless» and raise important questions regarding this new right (Wachter, Mittelstadt, & Floridi, 2017).

[10] Feasibility study on discrimination in ADM, to be published in 2018 by Gesellschaft für Informatik.

2. A Framework for Explanation and Interpretation

2.1. Trying to Capture the Essence of Explainability

A possible definition for explanation in the context of ADM could be: «A formal and unambiguous descriptive representation of the output of a classifier based on the current and prior input represented as parameters.» This definition consists of four main parts:

1. Formal and unambiguous representation
2. Descriptive nature of the representation
3. Output of a classifier with regard to a decision
4. Current and prior input of the classifier

While satisfying from technical point of view, it is questionable whether this definition actually creates more clarity for a lay person. Even the behaviour of neural networks can be «explained» with mathematical formulas including operations on (very large) matrices in a high-dimensional space. We could theoretically create an unambiguous representation of these systems' decisions, and it would be valid. However, just as clarity around certain legal concepts can be destroyed by excessive legalese in a contract for example, if most people are unable to comprehend these technical representations of an algorithm's inner workings, human subjects of ADM would not be reassured and it would do little to further the goals of the explainable artificial intelligence field.

In a recent article Doshi-Velez et al. stated on the governance of explanations that «when demanding explanation from humans, what we typically want to know is how and whether certain input factors affected the final decision or outcome» (Doshi-Velez et al., 2017). Unfortunately, this definition and finding does not provide much helpful guidance for those who design and implement the algorithms, mostly software developers and computer scientists. A more differentiated perspective on ADM is required.

Gunning provided us with the following set of basic questions that help to assess ADM (Gunning, 2017). The questions can be considered to as guidelines to provide more structure to the development of ADM system and improve their intrinsic explainability:

1. Why did that output happen?
2. Why not some other output?
3. For which cases does the machine produce a reliable output?
4. Can you provide a confidence score for the machine's output?
5. Under which circumstances, i.e. state and input, can the machine's output be trusted?
6. Which parameters effect the output most (negatively and positively)?
7. What can be done to correct an error?

Based on these questions, algorithms can be built to provide explanations for why specific instances or entire classes that were trained for a specific task (in case of supervised machine learning tasks) were classified in the way they were. On a superficial level, these kinds of explanations provide insights on the reasoning process. However, they do not provide answers with regard to other important considerations, such as whether the algorithm may have discriminated against the particular subject of an automated decision. The classification process is in itself just the application of pre-trained machine learning models (see Section 3), which are based on applied mathematics, in other words deterministic and value-free processes. Moral or legal questions – such as whether someone has been discriminated against – should in our opinion not be answered by the algorithm. Answering those questions requires operationalized methods that allow for objective evaluation based on clear conditions. For example, if a decision is significantly influenced by one feature or attributes that are part of individual identity, such as age or sex, this would allow for objective evaluation of an algorithmic decisions discriminatory outcomes. But the ADM process in practice covers more than just the application of a machine learning algorithm, which has been discussed in greater detail in Section 1.4.

Technological considerations on the intrinsic properties of ADM can be used to develop a framework for interpretations and explanations of AI systems. A recent work on the interpretation of machine learning models implies the following useful framework by Lipton (2016):

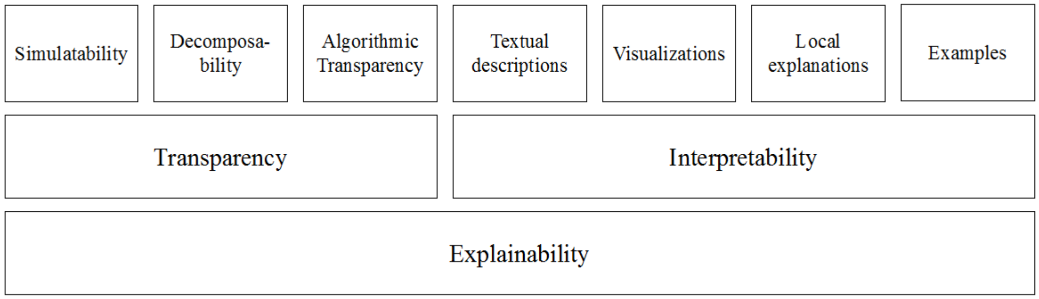

Figure 2: Taxonomy of explainability in the field of ADM.

Explainability is divided along functional and descriptive categories and covers the two main fields transparency and interpretability. Transparency is sub-divided into simulatability (What-if-analysis), decomposability along the whole process of ADM (see Section 3), and algorithmic transparency (see Section 2.5.).

Interpretability is divided into four sub-categories:

1. provision of textual descriptions of how a model behaves and how specific features contribute to a classification,
2. visualizations as dense representations of features and their contexts,
3. local explanations providing information about the impact of one particular feature, and finally, and
4. examples that illustrate the internal structure of a trained model.

In this context, transparency means the process of making a decision-making process visible with all the phases and interactions between components of the algorithms. Complementary to that, interpretability summarize every effort that is made by humans or machines to provide descriptive information, i.e. visualizations, local explanations, that enable humans to understand the decisions made by ADM.

2.2. Expanding Comparability Dimensions of Machine Learning Algorithms for ADM

Machine learning algorithms, which are powering many of the artificial intelligence systems at issue here, can be analyzed along different parameters and dimensions. Kotsiantis proposed a comparison chart which has become widely accepted (Kotsiantis, 2007). Table 1 below shows the dimensions that Kotsiantis proposed. We adapted the table to the research question of this article. The columns list six common machine learning algorithms, and the rows describe the desirable property of the specific machine learning algorithm.

	Decision Trees	Neural Networks	Naïve Bayes	kNN	SVM	Deductive logic based[11]
Accuracy	••	•••	•	••	•••	••
Speed of learning	•••	•	••••	••••	•	••
Speed of classification	••••	••••	••••	•	••••	••••
Tolerance w.r.t. input	•••	••	•	•	••••	••
Transparency of the process	••••	••	•••	••	••	•••
Transparency of the model	••••	•	••	•••	••	••••
Transparency theof classification	••••	•	•••	•••	•	••••

Table 1: Extension of comparison of common machine learning algorithms based on Kotsiantis **(2007).**

Table 1 shows that the main features of interest to end-users are the overall accuracy of an algorithm and the speed of learning or classification. Of course, the overall effort required to train a classifier is also very important measure, especially in practical usage scenarios. We propose to add the three following rows that represent the explainability and transparency of ADM dimensions:

- **Transparency of the process:** The ability of explain refers to the entire process of ADM, which consists of different steps that are all based on individual design decisions. So, to what degree can an ADM process be made transparent and explainable? This refers mainly to the structure of the classification process (see Section 4). Developing a machine learning classifier does not only require the training of a machine learning classifier. It also requires that parameterization of the algorithm, data collection, pre-processing, etc. Throughout the parameterization process many parameters have to be set. For example, decision trees, including random forests, can be parameterized to have a maximal depth. This is known as «pruning», which helps explain the impact of a particular parameter on the trained model. In addition, human biases can influence the data collection which in turn can influence the training and the classification of the model[12].
- **Transparency of the model:** To what degree can the model underlying a trained classifier be made transparent and be communicated clearly to humans? In most implementations of machine learning classifiers, the internal representation of the trained model is stored very monolithically, which does not facilitate deeper understanding of the trained model. Classifiers, such as decision trees or deductive logic based systems, represent their decision structures in rules that can be analyzed and interpreted by humans. More elaborate classifiers such as Naïve Bayes or SVMs[13] rely on a more elaborate mathematical representation, mainly optimizing of probabilities or a so-called kernel function. These internal representations require much more effort to understand and interpretation. Finally, neural networks tend

[11] Technically deductive logic based approaches are mostly implemented with rule-based systems and do belong to the class of machine learning based approaches.

[12] An excellent observation has been made by Goodman & Flaxman (2016): «Machine learning depends upon data that has been collected from society, and to the extent that society contains inequality, exclusion or other traces of discrimination, so too will the data.»

[13] Support Vector Machine.

to have a very complex internal structure, including large amount of so-called «hidden layers,» that make the underlying decision-making process very hard to interpret and comprehend. For every classifier, the reasoning structure is highly rational, albeit in a very formal mathematical sense, which frequently does correspond with what engineers and users would deem acceptable clarity.

- **Transparency of the classification:** To which degree can a classified instance of a trained classifier be made transparent and be communicated? In addition to the transparency of the underlying knowledge on which a model is trained, the transparency of the classification used plays an important role. Given a concrete classification that lead to a specific algorithmic decision, to what degree can the classifier itself make the factors transparent that lead to the specific decision? What were the features that contributed to a specific decision and what was the concrete weight of their contribution? This should also take into account whether a feature contributed directly or indirectly (by influencing other features) to the overall classification. This should not only complement the transparency of the model but should also have an impact on future research and systems engineering. There may be instances where the transparency of the model may not be important at all, while the reproducibility of a decision can be critical.

The addition of three additional dimensions to capture transparency aims to underscore the tremendous importance of explainability as a property inherent to machine learning algorithms. It highlights that explainability can be an additional feature and dimension along which machine learning algorithms can be categorized. The next section will extend the concept of transparency as an inherent property and establish it as a concept that research and industry alike should take into account as a critical aspect in developing new algorithms.

2.3. Explainability as an Intrinsic Property of Machine Learning Algorithms

Every machine learning algorithm has different properties that can be used as a means to categorize and compare them against each other. We have proposed transparency of machine algorithms as an additional important property. In our framework explainability is not treated as an optional characteristic or additional external property that can be «plugged-in» to machine learning classifiers like adding a new component or application to a set of given functions and capabilities. Instead it is an intrinsic property of every ADM that can be improved and enhanced just as other (performance) attributes can be improved.

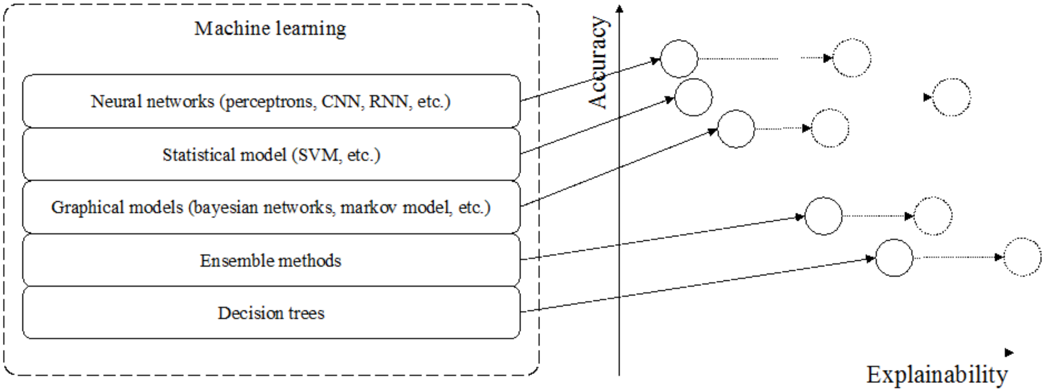
Machine learning
Neural networks (perceptrons, CNN, RNN, etc.)
Statistical model (SVM, etc.)
Graphical models (bayesian networks, markov model, etc.)
Ensemble methods
Decision trees
Accuracy
Explainability

Figure 3: Explainability of machine learning classifiers as desirable property based on Gunning (2017).

Gunning suggested explainability as a desirable property that does not have to negatively affect the performance of a machine learning classifiers (Gunning, 2017). The main idea is illustrated in Figure 3. The figure shows different classes of machine learning algorithms. Although the sets of classifiers are stacked in the figure, this is not meant to create the impression that they are building on each other or that one would be superior to another. One could compare these classifiers along a variety of dimensions such as speed of learning or the time needed for the classification of one instance. The figure maps five different categories of machine learning techniques into a two-dimensional space, illustrating the relationship between explainability and accuracy for each category. The order of the different categories is random and is not meant to invite conclusions about whether the accuracy of a particular category of classifiers is superior to the accuracy of another. This results from the fact that the performance of a particular algorithm heavily depends on the concrete classification task, as well as the training data, and the selected features. Gunning's figure primarily represents a visual effort to express the overarching objective that increasing the explainability of machine learning classifiers should be accomplished without decreasing the accuracy or the computational performance of the classifier. However, most current techniques lack the ability to explain how a particular decision was made. Instead they are mainly focused on performance of the classifiers/algorithms in terms of accuracy and computability. We concur with Gunning's conclusion implied in Figure 3 in that different classifier categories allow for different levels of explainability features and components.

3. Conclusion

As the above analysis shows, furnishing machine learning classifiers and ADM processes in general with functionality that increases their explainability is a non-trivial task. In order to contribute to a more nuanced discourse on the promises and pitfalls of AI, we attempted provide a constructive differentiation and terminological clarification regarding explainable AI. We conclude that explanations of decisions made by AI systems are possible. We differentiated three levels of transparency, specifically transparency on the process, on the model, and the instance.

Artificial intelligence is based on deterministic procedures and mathematical operations. Although it can be observed, that machines outperform humans in many tasks, their decisions are rational and follow strict mathematical, i.e. rational, structures. We need to put more emphasis on making these structures transparent. This would not only allow us to increase their explainability, but it also allows us to optimize these systems and understand the boundaries of artificial intelligence in greater depth. We are already witnessing a trend towards the use of explanatory classification systems in the field of computer vision.[14] We argue that increasing explainability of machine learning techniques more generally, will not only increase the acceptance of ADM based on machine learning (see Section 1), but it will also allow system engineers to improve the classification mechanisms and the algorithmic decision making itself. The increasing demands by regulators to provide secure and transparent systems for consumers, will make the emerging field of interpretable machine learning more and more attractive for academic research and entrepreneurial solutions.

4. References

Ashley, K. D. (2017). *Artificial Intelligence and Legal Analytics: New Tools for Law Practice in the Digital Age*. Cambridge: Cambridge University Press.

Bellman, R. (1978). *An Introduction to Artificial Intelligence: Can Computers Think?*: Boyd & Fraser.

Bench-Capon, T., Araszkiewicz, M., Ashley, K., Atkinson, K., Bex, F., Borges, F., . . . Wyner, A. (2012). A history of AI and Law in 50 papers: 25 years of the international conference on AI and Law. *Artificial Intelligence and Law*.

[14] In that field, modern classification systems for images are mainly built on so-called trained neural networks, which are viewed as the classifiers that are in every dimension the least transparent and the most difficult to explain. Hendricks showed that even state-of-the-art neural network classifiers can be modified to provide a visual explanation for the classification of images and their content (Hendricks et al., 2016). A visual explanation contains both explanations for a concrete image (which correspond to our category of transparency of classification) as well as for a class (corresponding to our category for transparency of the model). Hendricks» article provides a good example of the technologies that can be leveraged to enhance explainability of so-called «deep models», which are considered difficult to interpret.

Doshi-Velez, F., Kortz, M., Budish, R., Bavitz, C., Gershman, S., O'Brien, D., . . . Wood, A. (2017). Accountability of AI Under the Law: The Role of Explanation. *arXiv preprint arXiv:1711.01134*.

Fayyad, U., Piatetsky-Shapiro, G., & Smyth, P. (1996). From data mining to knowledge discovery in databases. *AI magazine, 17*(3), 37.

Fiedler, H. (1990). Entmythologisierung von Expertensystemen In H. Bonin (Ed.), *Einführung in die Thematik für Recht und öffentliche Verwaltung*. Heidelberg: Decker & Müller-Verlag.

Goodman, B., & Flaxman, S. (2016). European Union regulations on algorithmic decision-making and a «right to explanation». *arXiv preprint arXiv:1606.08813*.

Gunning, D. (2017). *Explainable Artificial Intelligence (XAI)*

Hendricks, L. A., Akata, Z., Rohrbach, M., Donahue, J., Schiele, B., & Darrell, T. (2016). *Generating Visual Explanations*. Paper presented at the ECCV.

Kotsiantis, S. B. (2007). *Supervised Machine Learning: A Review of Classification Techniques*. Paper presented at the Proceedings of the 2007 conference on Emerging Artificial Intelligence Applications in Computer Engineering.

Lipton, Z. C. (2016). The mythos of model interpretability. *arXiv preprint arXiv:1606.03490*.

Russell, S., & Norvig, P. (2009). *Artificial Intelligence: A modern approach*.

Vogl, R. (2017). In M. Hartung, M.-M. Bues, & G. Halbleib (Eds.), *Digitalisierung des Rechtsmarkts*: C. H. Beck.

Wachter, S., Mittelstadt, B., & Floridi, L. (2017). Why a right to explanation of automated decision-making does not exist in the general data protection regulation. *International Data Privacy Law, 7*(2), 76–99.

Waltl, B., Bonczek, G., Scepankova, E., Landthaler, J., & Matthes, F. (2017). Predicting the Outcome of Appeal Decisions in Germany's Tax Law. 89–99. *doi:10.1007/978-3-319-64322-9_8*

Waltl, B., Landthaler, J., Scepankova, E., Matthes, F., Geiger, T., Stocker, C., & Schneider, C. (2017). Automated Extraction of Semantic Information from German Legal Documents. *Jusletter IT, Conference Proceedings IRIS 2017*.