Explanation and Justification in Machine Learning: A Survey

Or Biran
n-Join
or@n-join.com

Courtenay Cotton
n-Join
courtenay@n-join.com

Abstract

We present a survey of the research concerning explanation and justification in the Machine Learning literature and several adjacent fields. Within Machine Learning, we differentiate between two main branches of current research: interpretable models, and prediction interpretation and justification.

1 Introduction

A key component of an artificially intelligent system is the ability to *explain* the decisions, recommendations, predictions or actions made by it and the process through which they are made. Explanation is closely related to the concept of *interpretability*: systems are interpretable if their operations can be understood by a human, either through introspection or through a produced explanation. In the case of machine learning models, explanation is often a difficult task since most models are not readily interpretable. A related concept is *justification*: intuitively, a justification explains why a decision is a good one, but it may or may not do so by explaining exactly how it was made. Unlike introspective explanations, justifications can be produced for non-interpretable systems.

Explanation has been shown to be important for user acceptance and satisfaction in a number of studies. In one early study, physicians rated the ability to explain decisions as the most highly desirable feature of a decision-assisting system [Teach and Shortliffe, 1981]. [Ye and Johnson, 1995] experimented with three types of explanations for an expert system - trace, justification and strategy - and found that explanations in general and justifications in particular make the generated advice more acceptable to users, and that justification (defined as showing the rationale behind each step in the decision) was the most effective type of explanation in changing users' attitudes towards the system. Later studies that empirically tested the importance of explanation to users, in various fields, consistently showed that explanations significantly increase users' confidence and trust [Herlocker *et al.*, 2000; Sinha and Swearingen, 2002; Bilgic and Mooney, 2005; Symeonidis *et al.*, 2009] as well as their ability to correctly assess whether a prediction is accurate [Kim *et al.*, 2016; Gkatzia *et al.*, 2016; Biran and McKeown, 2017].

2 History and Adjacent Fields

Work on producing explanations comes from multiple fields. In this section, we focus on historical background and current work in fields adjacent to machine learning.

2.1 Expert Systems and Bayesian Networks

Historically, explanations first appeared in the context of rule-based expert systems, and were mostly treated as a systems design task (i.e., the task of designing a system capable of producing drill-down into its decisions). The need for explaining the decisions of expert systems was discussed as early as the 1970's [Shortliffe and Buchanan, 1975]. [Swartout, 1983] described a framework for creating expert systems with explanation capabilities, and was one of the first to stress the importance of explanations that are not merely traces, but also contain justifications. [Swartout *et al.*, 1991] is a later example of such a framework. Both were exclusively for rule-based systems and relied on a domain-specific taxonomic knowledge base and a separate strategic knowledge base. [Barzilay *et al.*, 1998] further separated the knowledge into three layers, adding the *communication* layer to the previously described *domain* and *strategic* layers. Separating the communication layer from the rest of the system was intended to allow a communication expert to create solutions that were independent of the specific system and domain.

In some domains probabilistic decision-making systems, often based on Bayesian Networks (BN), are still referred to as expert systems and regarded as successors of earlier rule-based systems. The (scarce) work on explanation for these BN systems self-describes as expert systems explanation. [Lacave and Díez, 2002] present a survey of methods of explanation for Bayesian networks and an excellent analysis of the methods in terms of several properties of explanation. Of particular interest is their classification of the focus of explanation into an explanation of the *reasoning*, the *model*, and the *evidence* for the decision. Most work on explanation in Bayesian networks has been within the narrow context of a particular system, and relies on producing canned text showing the actual posterior probabilities of each node and providing no explanation for what the nodes themselves symbolize, assuming that their names are enough (individual nodes are often symptoms, in the medical domain, or physical evidence, e.g. "valve open", in other domains) [Druzdzel, 1996; Haddawy *et al.*, 1997; Yap *et al.*, 2008].

2.2 Recommender Systems

Recommender systems are online services that serve a large number of users and provide individualized recommendations for media or products. It is usually desirable to produce a short and intuitive justification to help the users decide whether to follow the recommendation or not.

[Herlocker *et al.*, 2000] conducted an experiment measuring user satisfaction with a variety of justification types for a collaborative filtering (neighbor-based) movie recommendation system. They found that the most satisfying were simple and conclusive methods, such as stating the neighbors' ratings or showing one strong feature like a favorite actor. Justifications using ML concepts such as model confidence and complex justifications such as a full neighbor graph scored significantly lower. Regardless of type, 86% of users wanted the justifications they were shown added to the system. Other studies from the early 2000's have also shown that users are overwhelmingly more satisfied with systems that contain some form of justification [Sinha and Swearingen, 2002].

[Symeonidis *et al.*, 2009] presented a style of justification that focused on the most important feature along with the user's past history with regards to that feature. A user study showed that this justification style was significantly more satisfying to users than previous methods. [Papadimitriou *et al.*, 2012] defined a classification of recommender system explanations into three types: those based on previous items chosen by the user, those based on choices of similar users, and those based on features. They also defined a hybrid type which combines two or more of the above, and following a user study concluded that feature-based explanations were the best of the three core types, and that hybrid explanations were best overall. [Bilgic and Mooney, 2005] noted that previous studies have often evaluated the persuasiveness of the justification and not its justifiability. Their experiments showed that for justifiability, feature-based justifications were superior to neighbor-based and user-history-based ones.

2.3 Other adjacent Fields

Generating explanations for users has also been explored in *constraint programming*, specifically for problems where the user may have an interactive role in solving the problem and therefore needs to understand why certain choices were made. In [Freuder *et al.*, 2001] and [Wallace and Freuder, 2001] the authors explored a method of presenting explanations for any assignment decision made by their program. An assignment can be explained by identifying a set of previous assignments that form a sufficient basis to justify the current one. Applying this at each subsequent assignment step forms what the authors call an explanation tree.

Context-aware systems are those that are capable of sensing environmental changes and responding to them. Some work has been done on providing users with explanations for the behavior of these systems. [Tullio *et al.*, 2007] studied mental models that users developed of a system for predicting their managers' interruptibility. However, they concluded that the low level feature contributions that they presented to users were only moderately helpful in improving users understanding of the system, and recommended using higher level concepts instead. [Lim and Dey, 2010] developed a toolkit for use in context-aware applications that provides eight types of explanation for four of the most common model types (rules, decision trees, naïve Bayes and HMMs).

There has been some work on explanation of *Markov Decision Processes* (MDPs). In the context of a particular state in a MDP, it is sometimes desirable to explain to a user what is the best current course of action and why. [Elizalde *et al.*, 2007] describe an explanation system that assists plant operators in executing necessary operations; [Khan *et al.*, 2009] explore the minimal explanation sufficient for tasks such as picking the next course in a college curriculum; [Dodson *et al.*, 2011] propose a dialog system, instead of a single fixed explanation, which allows the user to argue and ask questions.

The *case-based reasoning* community has also explored explanation of probabilistic systems. One example is [Nugent *et al.*, 2009] who proposed a case-based method of explanation for decision support systems, where alternative samples are selected and the explanation focuses on how they differ (if the decision is different) or their similarities (otherwise).

Causal discovery is concerned with determining the direction of causality between variables in a model, which can help explain the behavior of the model. [Hoyer *et al.*, 2009] exploited both non-linearity and non-Gaussianity of real data to identify causality between variables, even in the presence of additive noise. Demonstrating causal relationships is useful for justifying predictions based on these models to users.

In *forensic science*, [Vlek *et al.*, 2016] explained legal cases by combining Bayesian networks with a narrative idiom they call a scenario, taking statistical evidence into account while also maintaining a narrative framework which helps a judge or jury understand the assumptions being made and the relationships among them. This allows insight into the structure of the statistical model, which is crucial for humans to make an informed decision in a legal case. [Timmer *et al.*, 2017] used a somewhat similar approach to generate explanations for Bayesian networks in legal cases. Their work relies on defining a support graph directed toward the variable of interest, then using it to construct an argument.

Interpretability has also been studied in the context of *communicating agents*. [Lazaridou *et al.*, 2017] experimented with neural agents which learn to communicate with each other about images. They then leverage a human-supervised task to ground the learned communication in a way that would be understandable to humans. [Andreas *et al.*, 2017] also studied messages passed between agents in systems with learned deep communicating policies. They developed a strategy for translating these messages into natural language based on the underlying beliefs implied by messages. This is similar to understanding the beliefs implied by any model.

A field that is particularly closely related to explanation is Natural Language Generation (NLG). Much of the work discussed in this survey uses NLG (of varying sophistication) to produce explanations. In addition to explaining ML and other AI systems, however, there has been work on explanations of other kinds. For example, [Pace and Rosner, 2014] produce explanations of user interactions with a software system, intended for administrators, while [Gkatzia *et al.*, 2016] explain how a weather forecast was produced and show that it helps readers decide whether or not to believe the forecast.

Other related work includes [McGuinness and Borgida, 1995], who proposed generating explanations as a debugging tool for the developers and users of a Description Logic-based system. They first break down inference rules into atomic descriptions; corresponding atomic explanations are then created using subsumption rules, and chained to form proofs supporting the system's conclusions.

2.4 Theoretical Work

There has also been some theoretical work on explanation. [Chajewska and Halpern, 1997] proposed a formal definition of explanation in general probabilistic systems, after examining two contemporary ideas and finding them incomplete. In expert systems, [Johnson and Johnson, 1993] presented a short survey of accounts of explanation in philosophy, psychology and cognitive science and found that they fall into three categories: associations between antecedent and consequent; contrasts and differences; and causal mechanisms. In recommender systems, [Yetim, 2008] proposed a framework of justifications which uses existing models of argument to enumerate the components of a justification and provide a taxonomy of justification types. [Corfield, 2010] aims to formalize justifications for the accuracy of ML models by classifying them into four types of reasonings, two based on absolute performance and two rooted in Bayesian ideas.

More recently, [Doshi-Velez and Kim, 2017] considered how to evaluate human *interpretability* of machine learning models. They proposed a taxonomy of three approaches: application-grounded, which judges explanations based on how much they assist humans in performing a real task; human-grounded, which judges explanations based on human preference or ability to reason about a model from the explanation; and functionally-grounded, which judges explanations without human input, based on some formal proxy for interpretability. For this third approach, they hypothesized that matrix factorization of result data (quantized by domain and method) may be useful for identifying common latent factors that influence interpretability.

3 Machine Learning

In the machine learning literature, early work on explanation often focused on producing visualizations of the prediction in order to assist machine learning experts in evaluating the correctness of the model. One very common visualization technique is *nomograms*. It was first applied to logistic regression models by [Lubsen *et al.*, 1978], and later to Naive Bayes [Možina *et al.*, 2004], SVM [Jakulin *et al.*, 2005] and other models. [Szafron *et al.*, 2003] proposed a visualization-based explanation framework for Naive Bayes classifiers.

More recently, visualization techniques have focused on visualizing the hidden states of neural models [Tzeng and Ma, 2005], most notably of Convolutional Neural Nets (CNNs) in image classification [Simonyan *et al.*, 2013; Zeiler and Fergus, 2013] and of Recurrent Neural Nets (RNNs) in Natural Language Processing (NLP) applications [Karpathy *et al.*, 2015; Li *et al.*, 2016; Strobelt *et al.*, 2016].

Beyond visualization, research has focused on two broad approaches to explanation. The first is *prediction interpretation and justification*, where a (usually non-interpretable) model and prediction are given, and a justification for the prediction must be produced. The second is *interpretable models*, which aims to devise models that are intrinsically interpretable and can be explained through reasoning.

3.1 Prediction Interpretation and Justification

This approach has focused on interpreting the predictions of complex models, often by proposing to isolate the contributions of individual features to the prediction. Such proposals were made for Bayesian networks [Suermondt, 1992], multilayer Perceptrons [Feraud and Clerot, 2002], RBF networks [Robnik-Šikonja *et al.*, 2011] and general hierarchical networks [Landecker *et al.*, 2013]. [Martens *et al.*, 2008] proposed to interpret the predictions of an SVM classifier by extracting conjunctive rules using a small subset of features.

In addition to model-specific methods, there have been a few suggestions for model-agnostic frameworks. [Robnik-Šikonja and Kononenko, 2008] proposed measuring the effect of an individual feature on an unknown classifier's prediction by checking what the prediction would have been if that feature value was absent and comparing the two using various distance measures. The effects are then displayed visually to explain the main contributors towards a prediction or to compare the effect of the feature in various models. This method was extended to include regression models in [Kononenko *et al.*, 2013]. [Baehrens *et al.*, 2010] described an alternative approach using *explanation vectors* (class probability gradients) which highlight the effect of the most important features.

Other work, especially in the NLP literature, has focused on using a small portion of input as evidence to justify the prediction result, and often explored alternative definitions of evidence and styles of explanation. [Martens and Provost, 2014] describe a framework of linguistic explanations for document classification with bag-of-words features. Their method shows removal-based explanations of the type "the classification would change to [alternative class] if the words [list of words] were removed from the document", which can help a domain expert intuitively assess how solid the prediction is. [Kim *et al.*, 2016] select two subset of training samples: *prototypes* - samples of different types that the model represents well; and *criticisms* - samples that are most misrepresented by the model. They show that this model-level explanation makes users more likely to correctly predict the model's success with new samples. [Lei *et al.*, 2016] select small snippets of the input text of text classification tasks as justification for the decision. The justification model is separate from the prediction model, but trained on the same data, with the constraint that the prediction of the main model for the (much shorter) justification should be very similar to that for the full text. [Biran and McKeown, 2017] define evidence as the intersection of a feature's actual contribution and expected contribution, and categorize features that are important to the prediction based on that definition. Their work therefore shows not only actual evidence but also *missing evidence*, an important part of human reasoning, and differentiates between expected and unexpected evidence.

Work on *model approximation* focuses on deriving a simple, interpretable model (such as a shallow decision tree, rule list, or sparse linear model) that approximates a more com-

plex, uninterpretable one (e.g., a neural net). Early work described approximations of the entire model [Thrun, 1995; Craven and Shavlik, 1999]; the disadvantage of these approaches is that for even moderately complex models, a good global approximation cannot generally be found. [Ribeiro *et al.*, 2016] introduce an approach that focuses on local approximations, which behave similarly to the global model only in the vicinity of a particular prediction. Their algorithm is agnostic to the details of the original model.

In image classification, there has been work on secondary neural models, inspired by neural caption generation, that learn to generate textual justifications for classifications of the primary neural model. [Hendricks *et al.*, 2016] use an LSTM caption generation model with a loss function that encourages class discriminative information to generate justifications for the image classification of a CNN. [Park *et al.*, 2016] produce both a textual justification and a visual attention map, making their approach a combination of an interpretable model (See Section 3.2) and an external justification model. [Vedantam *et al.*, 2017] produce captions that are locally discriminative, in the context of other images.

3.2 Interpretable Models

An alternative to methods for interpreting or justifying otherwise black-box models is to produce models that are inherently interpretable. One family of models that are readily interpretable by humans are shallow rule-based models: decision lists and decision trees. [Rudin *et al.*, 2013] introduced classifiers that use association rules [Agrawal *et al.*, 1993], which can be learned efficiently from sparse data. This family of models includes Bayesian Rule Lists [Letham *et al.*, 2015], an algorithm that generates a posterior distribution of decision lists that encourages sparsity as well as accuracy; Bayesian Or's of And's [Wang *et al.*, 2015], highly efficient disjunctive rule lists; and Falling Rule Lists [Wang and Rudin, 2015], where the order of rules implies both domain-level importance and estimated probability of success. Other approaches have focused on creating sparse models via features selection or extraction that aims to optimize interpretability. Examples include Supersparse Linear Integer Models [Ustun and Rudin, 2016] and Mind-the-Gap Model [Kim *et al.*, 2015].

In deep learning, attention mechanisms which allow a model to focus on a subset of its vector representation were found to improve accuracy on many tasks, particularly within NLP [Bahdanau *et al.*, 2014] and image classification [Xu *et al.*, 2015], and at the same time result in significantly more intuitive hidden states that appear semantically appropriate to humans when inspected.

Finally, there has been some work on *compositional generative models* which are constrained or encouraged to learn hierarchical, semantically meaningful representations of data. [Si and Zhu, 2013] learn compositional models of objects in images: an object (e.g., a cat) contains a mandatory set of parts (e.g., ears), but each part can come in many forms (pointed, round...), and each form is represented using a pixel-level generative model. [Lake *et al.*, 2015] learn a generative model of linguistic character images from sparse data. Their approach infers motor programs (line strokes) from sample images and learns a prior generative model of generative models, which can then produce a reasonable model from even one sample of a new character.

4 Conclusion

While eXplainable AI (XAI) is only now gaining widespread visibility, the ML literature and that of allied fields contain a long, continuous history of work on explanation and can provide a pool of ideas for researchers currently tackling the task of explanation. Despite this history, current efforts face unprecedented difficulties: contemporary models are more complex and less interpretable than ever; they are used for a wider array of tasks, and are more pervasive in everyday life than in the past; and they are increasingly allowed to make (and take) more autonomous decisions (and actions). Justifying these decisions will only become more crucial, and there is little doubt that this field will continue to rise in prominence and produce exciting and much needed work in the future.

References

[Agrawal *et al.*, 1993] R Agrawal, T Imieliński, and A Swami. Mining association rules between sets of items in large databases. *SIGMOD*, 22(2):207–216, June 1993.

[Andreas *et al.*, 2017] Jacob Andreas, Anca Dragan, and Dan Klein. Translating neuralese. In *Proceedings of the Association for Computational Linguistics (ACL)*, 2017.

[Baehrens *et al.*, 2010] David Baehrens, Timon Schroeter, Stefan Harmeling, Motoaki Kawanabe, Katja Hansen, and Klaus-Robert Müller. How to explain individual classification decisions. *JMLR*, 11, August 2010.

[Bahdanau *et al.*, 2014] D Bahdanau, K Cho, and Y Bengio. Neural machine translation by jointly learning to align and translate. *CoRR*, abs/1409.0473, 2014.

[Barzilay *et al.*, 1998] Regina Barzilay, Daryl Mccullough, Owen Rambow, Jonathan DeCristofaro, Tanya Korelsky, and Benoit Lavoie. A new approach to expert system explanations. In *International Workshop on NLG*, 1998.

[Bilgic and Mooney, 2005] Mustafa Bilgic and Raymond J. Mooney. Explaining recommendations: Satisfaction vs. promotion. In *Workshop on the Next Stage of Recommender Systems Research*, San Diego, CA, 2005.

[Biran and McKeown, 2017] Or Biran and Kathleen McKeown. Human-centric justification of machine learning predictions. In *IJCAI*, Melbourne, Australia, 2017.

[Chajewska and Halpern, 1997] U Chajewska and J Y Halpern. Defining explanation in probabilistic systems. In *Uncertainty in artificial intelligence*, 1997.

[Corfield, 2010] David Corfield. Varieties of justification in machine learning. *Minds and Machines*, 20(2):291–301, 7 2010.

[Craven and Shavlik, 1999] Mark Craven and Jude Shavlik. Rule extraction: Where do we go from here?, 1999.

[Dodson *et al.*, 2011] Thomas Dodson, Nicholas Mattei, and Judy Goldsmith. A natural language argumentation interface for explanation generation in markov decision processes. In *Algorithmic Decision Theory*, 2011.

[Doshi-Velez and Kim, 2017] Finale Doshi-Velez and Been Kim. A roadmap for a rigorous science of interpretability. *arXiv preprint arXiv:1702.08608*, 2017.

[Druzdzel, 1996] Marek J Druzdzel. Qualitiative verbal explanations in bayesian belief networks. *AISB QUARTERLY*, pages 43–54, 1996.

[Elizalde *et al.*, 2007] F. Elizalde, L. E. Sucar, A. Reyes, and P. deBuen. An MDP approach for explanation generation. In *Explanation-Aware Computing Workshop at AAAI*, pages 28–33, Vancouver, BC, Canada, 2007.

[Feraud and Clerot, 2002] Raphael Feraud and Fabrice Clerot. A methodology to explain neural network classification. *Neural Networks*, 15(2):237 – 246, 2002.

[Freuder *et al.*, 2001] E. Freuder, C. Likitvivatanavong, and R. Wallace. Deriving explanations and implications for constraint satisfaction problems. In *Principles and Practice of CP*, pages 585–589. Springer, 2001.

[Gkatzia *et al.*, 2016] D Gkatzia, O Lemon, and V Rieser. Natural language generation enhances human decision-making with uncertain information. In *ACL*, 2016.

[Haddawy *et al.*, 1997] P. Haddawy, J. Jacobson, and C. E. Kahn. BANTER: a Bayesian network tutoring shell. *Artificial Intelligence in Medicine*, 10(2):177–200, June 1997.

[Hendricks *et al.*, 2016] L.A Hendricks, Z Akata, M Rohrbach, J Donahue, B Schiele, and T Darrell. Generating visual explanations. In *ECCV*, 2016.

[Herlocker *et al.*, 2000] J Herlocker, J Konstan, and J Riedl. Explaining collaborative filtering recommendations. In *Computer Supported Cooperative Work (CSCW)*, 2000.

[Hoyer *et al.*, 2009] P Hoyer, D Janzing, J Mooij, J Peters, and B Schölkopf. Nonlinear causal discovery with additive noise models. In *NIPS*, 2009.

[Jakulin *et al.*, 2005] A Jakulin, M Možina, J Demšar, I Bratko, and B Zupan. Nomograms for visualizing support vector machines. In *KDD*, 2005.

[Johnson and Johnson, 1993] H. Johnson and P. Johnson. Explanation facilities and interactive systems. In *IUI*, pages 159–166, New York, NY, USA, 1993.

[Karpathy *et al.*, 2015] Andrej Karpathy, Justin Johnson, and Fei-Fei Li. Visualizing and understanding recurrent networks. *CoRR*, abs/1506.02078, 2015.

[Khan *et al.*, 2009] Omar Zia Khan, Pascal Poupart, and James P. Black. Minimal sufficient explanations for factored markov decision processes. In *ICAPS*, 2009.

[Kim *et al.*, 2015] Been Kim, Julie Shah, and Finale Doshi-Velez. Mind the gap: A generative approach to interpretable feature selection and extraction. In *NIPS*, 2015.

[Kim *et al.*, 2016] Been Kim, Rajiv Khanna, and Oluwasanmi Koyejo. Examples are not enough, learn to criticize! criticism for interpretability. In *NIPS*. 2016.

[Kononenko *et al.*, 2013] Igor Kononenko, Erik Strumbelj, Zoran Bosnic, Darko Pevec, Matjaz Kukar, and Marko Robnik-Šikonja. Explanation and reliability of individual predictions. *Informatica (Slovenia)*, 37(1):41–48, 2013.

[Lacave and Díez, 2002] C. Lacave and F. J. Díez. A review of explanation methods for Bayesian networks. *Knowledge Engineering Review*, 17:107–127, 2002.

[Lake *et al.*, 2015] Brenden M. Lake, Ruslan Salakhutdinov, and Joshua B. Tenenbaum. Human-level concept learning through probabilistic program induction. *Science*, 350(6266):1332–1338, December 2015.

[Landecker *et al.*, 2013] W. Landecker, M.D. Thomure, L.M.A. Bettencourt, M. Mitchell, G.T. Kenyon, and S.P. Brumby. Interpreting individual classifications of hierarchical networks. In *CIDM*, 2013.

[Lazaridou *et al.*, 2017] A Lazaridou, A Peysakhovich, and M Baroni. Multi-agent cooperation and the emergence of (natural) language. In *ICLR*, Toulon, France, 2017.

[Lei *et al.*, 2016] T Lei, R Barzilay, and T.S Jaakkola. Rationalizing neural predictions. In *EMNLP*, 2016.

[Letham *et al.*, 2015] B Letham, C Rudin, T.H McCormick, and D Madigan. Interpretable classifiers using rules and bayesian analysis: Building a better stroke prediction model. *CoRR*, abs/1511.01644, 2015.

[Li *et al.*, 2016] Jiwei Li, Xinlei Chen, Eduard H. Hovy, and Dan Jurafsky. Visualizing and understanding neural models in NLP. In *NAACL-HLT*, 2016.

[Lim and Dey, 2010] Brian Lim and Anind Dey. Toolkit to support intelligibility in context-aware applications. In *Ubiquitous Computing (Ubicomp)*, 2010.

[Lubsen *et al.*, 1978] J. Lubsen, J. Pool, and E. van der Does. A practical device for the application of a diagnostic or prognostic function. *Methods of information in medicine*, 17(2):127–129, April 1978.

[Martens and Provost, 2014] David Martens and Foster Provost. Explaining data-driven document classifications. *MIS Q.*, 38(1):73–100, March 2014.

[Martens *et al.*, 2008] D Martens, J Huysmans, R Setiono, J Vanthienen, and B Baesens. Rule extraction from support vector machines: An overview of issues and application in credit scoring. In *Rule Extraction from SVMs*, volume 80 of *Studies in Comp. Int.*, pages 33–63. 2008.

[McGuinness and Borgida, 1995] Deborah L McGuinness and Alexander Borgida. Explaining subsumption in description logics. In *IJCAI (1)*, pages 816–821, 1995.

[Možina *et al.*, 2004] M. Možina, J. Demšar, M. Kattan, and B. Zupan. Nomograms for visualization of naive bayesian classifier. In *PKDD*, 2004.

[Nugent *et al.*, 2009] Conor Nugent, Dónal Doyle, and Pádraig Cunningham. Gaining insight through case-based explanation. *J. Intell. Inf. Syst.*, 32(3):267–295, June 2009.

[Pace and Rosner, 2014] Gordon J. Pace and Michael Rosner. *Explaining Violation Traces with Finite State Natural Language Generation Models*, pages 179–189. Springer International Publishing, 2014.

[Papadimitriou *et al.*, 2012] A. Papadimitriou, P. Symeonidis, and Y. Manolopoulos. A generalized taxonomy of

explanations styles for traditional and social recommender systems. *Data Min. Knowl. Discov.*, 24(3):555–583, 2012.

[Park *et al.*, 2016] D H Park, L A Hendricks, Z Akata, B Schiele, T Darrell, and M Rohrbach. Attentive explanations: Justifying decisions and pointing to the evidence. *CoRR*, abs/1612.04757, 2016.

[Ribeiro *et al.*, 2016] Marco Tulio Ribeiro, Sameer Singh, and Carlos Guestrin. "why should i trust you?": Explaining the predictions of any classifier. In *KDD*, 2016.

[Robnik-Šikonja and Kononenko, 2008] M. Robnik-Šikonja and I. Kononenko. Explaining classifications for individual instances. *TKDE*, 20(5):589–600, May 2008.

[Robnik-Šikonja *et al.*, 2011] M Robnik-Šikonja, A Likas, C Constantinopoulos, I Kononenko, and E Strumbelj. Efficiently explaining decisions of probabilistic rbf classification networks. In *ICANNGA*, 2011.

[Rudin *et al.*, 2013] C Rudin, B Letham, and D Madigan. Learning theory analysis for association rules and sequential event prediction. *JMLR*, 14:3441–3492, 2013.

[Shortliffe and Buchanan, 1975] Edward H Shortliffe and Bruce G Buchanan. A model of inexact reasoning in medicine. *Mathematical biosciences*, 23(3), 1975.

[Si and Zhu, 2013] Zhangzhang Si and Song-Chun Zhu. Learning and-or templates for object recognition and detection. *IEEE Trans. Pattern Anal. Mach. Intell.*, 35(9):2189–2205, September 2013.

[Simonyan *et al.*, 2013] Karen Simonyan, Andrea Vedaldi, and Andrew Zisserman. Deep inside convolutional networks: Visualising image classification models and saliency maps. *CoRR*, abs/1312.6034, 2013.

[Sinha and Swearingen, 2002] Rashmi Sinha and Kirsten Swearingen. The role of transparency in recommender systems. In *CHI EA*, 2002.

[Strobelt *et al.*, 2016] Hendrik Strobelt, Sebastian Gehrmann, Bernd Huber, Hanspeter Pfister, and Alexander M. Rush. Visual analysis of hidden state dynamics in recurrent neural networks. *CoRR*, abs/1606.07461, 2016.

[Suermondt, 1992] Henri Jacques Suermondt. *Explanation in Bayesian Belief Networks*. PhD thesis, Stanford, CA, USA, 1992. UMI Order No. GAX92-21673.

[Swartout *et al.*, 1991] W Swartout, C Paris, and J Moore. Explanations in knowledge systems: Design for explainable expert systems. *IEEE Expert*, 6(3):58–64, 1991.

[Swartout, 1983] William R. Swartout. Xplain: A system for creating and explaining expert consulting programs. *Artificial Intelligence*, 21(3), September 1983.

[Symeonidis *et al.*, 2009] P. Symeonidis, A. Nanopoulos, and Y. Manolopoulos. Moviexplain: A recommender system with explanations. In *RecSys*, 2009.

[Szafron *et al.*, 2003] D Szafron, R Greiner, P Lu, D Wishart, C Macdonell, J Anvik, B Poulin, Z Lu, and R Eisner. Explaining naive bayes classifications. Technical report, 2003.

[Teach and Shortliffe, 1981] R. Teach and E. Shortliffe. An Analysis of Physician Attitudes Regarding Computer-Based Clinical Consultation Systems. *Computers and Biomedical Research*, 14:542–558, 1981.

[Thrun, 1995] Sebastian Thrun. Extracting rules from artificial neural networks with distributed representations. In *NIPS*, 1995.

[Timmer *et al.*, 2017] S. Timmer, J. Meyer, H. Prakken, S. Renooij, and B. Verheij. A two-phase method for extracting explanatory arguments from bayesian networks. *Int. J. Approx. Reasoning*, 80(C):475–494, January 2017.

[Tullio *et al.*, 2007] Joe Tullio, Anind Dey, Jason Chalecki, and James Fogarty. How it works: A field study of non-technical users interacting with an intelligent system. In *SIGCHI Human Factors in Computing Systems*, 2007.

[Tzeng and Ma, 2005] F. Y. Tzeng and K. L. Ma. Opening the black box - data driven visualization of neural networks. In *IEEE Visualization*, pages 383–390, 2005.

[Ustun and Rudin, 2016] B Ustun and C Rudin. Supersparse linear integer models for optimized medical scoring systems. *Mach. Learn.*, 102(3):349–391, March 2016.

[Vedantam *et al.*, 2017] Ramakrishna Vedantam, Samy Bengio, Kevin Murphy, Devi Parikh, and Gal Chechik. Context-aware captions from context-agnostic supervision. *CoRR*, abs/1701.02870, 2017.

[Vlek *et al.*, 2016] Charlotte S. Vlek, Henry Prakken, Silja Renooij, and Bart Verheij. A method for explaining bayesian networks for legal evidence with scenarios. *Artificial Intelligence and Law*, 24(3):285–324, 2016.

[Wallace and Freuder, 2001] Richard J Wallace and Eugene C Freuder. Explanations for whom. In *CP01 Workshop on User-Interaction in Constraint Satisfaction*, 2001.

[Wang and Rudin, 2015] Fulton Wang and Cynthia Rudin. Falling rule lists. In *AISTATS*, 2015.

[Wang *et al.*, 2015] T Wang, C Rudin, F Doshi-Velez, Y Liu, E Klampfl, and P MacNeille. Or's of and's for interpretable classification, with application to context-aware recommender systems. *CoRR*, abs/1504.07614, 2015.

[Xu *et al.*, 2015] K Xu, J Ba, R Kiros, K Cho, A Courville, R Salakhudinov, R Zemel, and Y Bengio. Show, attend and tell: Neural image caption generation with visual attention. In *ICML*, Lille, France, 2015.

[Yap *et al.*, 2008] Ghim-Eng Yap, Ah-Hwee Tan, and Hwee-Hwa Pang. Explaining inferences in bayesian networks. *Applied Intelligence*, 29(3):263–278, 2008.

[Ye and Johnson, 1995] L. Richard Ye and Paul E. Johnson. The impact of explanation facilities on user acceptance of expert systems advice. *MIS Q.*, 19(2):157–172, June 1995.

[Yetim, 2008] Fahri Yetim. A framework for organizing justifications for strategic use in adaptive interaction contexts. In *ECIS*, pages 815–825, 2008.

[Zeiler and Fergus, 2013] Matthew D. Zeiler and Rob Fergus. Visualizing and understanding convolutional networks. *CoRR*, abs/1311.2901, 2013.