Better Metrics for Evaluating Explainable Artificial Intelligence

Blue Sky Ideas Track

Avi Rosenfeld
Department of Computer Science
Jerusalem College of Technology, Jerusalem, Israel, 91160
rosenfa@jct.ac.il

ABSTRACT

This paper presents objective metrics for how explainable artificial intelligence (XAI) can be quantified. Through an overview of current trends, we show that many explanations are generated post-hoc and independent of the agent's logical process, which in turn creates explanations with limited meaning as they lack transparency and fidelity. While user studies are a known basis for evaluating XAI, studies that do not consider objective metrics for evaluating XAI may have limited meaning and may suffer from confirmation bias, particularly if they use low fidelity explanations unnecessarily. To avoid this issue, this paper suggests a paradigm shift in evaluating XAI that focuses on metrics that quantify the explanation itself and its appropriateness given the XAI goal. We suggest four such metrics based on performance differences, ***D***, between the explanation's logic and the agent's actual performance, the number of rules, ***R***, outputted by the explanation, the number of features, ***F***, used to generate that explanation, and the stability, ***S***, of the explanation. We believe that user studies that focus on these metrics in their evaluations are inherently more valid and should be integrated in future XAI research.

KEYWORDS

Explainable Artificial Intelligence; Interpretable Machine Learning; Human-Agent Systems; System Evaluation

ACM Reference Format:
Avi Rosenfeld. 2021. Better Metrics for Evaluating Explainable Artificial Intelligence: Blue Sky Ideas Track. In *Proc. of the 21th International Conference on Autonomous Agents and Multiagent Systems (AAMAS 2021), Online, May 3–7, 2021*, IFAAMAS, 6 pages.

1 OVERVIEW

As the field of Artificial Intelligence matures and becomes ubiquitous, there is a growing emergence of systems where people and agents work together. These systems, often called Human-Agent Systems or Human-Agent Cooperatives, have moved from theory to reality in the many forms, including digital personal assistants, recommendation systems, training and tutoring systems, service robots, chatbots, planning systems, self-driving cars and medical diagnostic systems [2–4, 6, 14, 21–25, 27, 29, 34, 36, 37, 40–42, 42–45, 45, 47, 48, 50, 52, 53, 55, 56, 58–60, 63, 64, 66, 67, 69, 72, 73]. In this paper we focus on how well agents in these systems explain their logic to the people they interact with – the challenge of quantifying the effectiveness of explainable artificial intelligence (XAI).

Proc. of the 21th International Conference on Autonomous Agents and Multiagent Systems (AAMAS 2021), U. Endriss, A. Nowé, F. Dignum, A. Lomuscio (eds.), May 3–7, 2021, Online.

Please consider the following scenario to better motivate surrounding challenges in quantifying XAI. The XYZ company has just developed a new agent to automate the analysis of medical imaging to diagnose a deadly disease such as cancer. Currently, radiologists are 97% successful in finding the disease using state-of-the-art imaging techniques, but the agent is accurate with 99.5% accuracy. However, both the agent and the human experts miss different types of cancer. As such, 0.5% of the cancers that are missed by the agent are found by the experts, but the agent is overall significantly better than experts at finding cancer. One would ideally hope that the agent and human work in tandem and thus experts will find the cancers the agent missed and the agent will inform the experts of cancers they didn't find to create an 100% overall accuracy. Unfortunately, XYZ's agent provides no explanation and they have also noted that the experts that use their system often trust it when they shouldn't. As a result they have formulated the following questions: Would the experts learn how to find cancer better had the system explained its logic better? Given a set of explanations, how can XYZ quantify the effectiveness of each explanation and identify the best one? Is the agent safe and should it be trusted? How can XYZ quantify bias if the agent mistakes only one ethnic minority's cancer? Should society hold XYZ legally liable for this agent bias or might it be the expert's responsibility to avoid this problem?

The goal of this paper is to highlight how ambiguities in XAI definition and goals impact how agent designers quantify XAI. We argue that many current XAI methods are based on a wrong assumption that agents must maximize their performance using certain machine learning techniques even if they are not readily and fully understood by the intended user. This disconnect highlights a potentially poor fit between the motivation for why explanations are needed and how those algorithms are currently being evaluated by XAI researchers. As we will see in the following sections, many such explanations are not capable of instilling trust in the system [55] and others even hurt the user's ability to understand the agent's decisions [38]. Also, user studies to date typically measure XAI based on the user's performance and how it's impacted by explanations [12, 22, 37, 38, 51, 65]. However, not only are user studies relatively hard to run, they may be of limited value [38] and may suffer from confirmation bias [68]. Instead, we present four objective XAI measures to quantify XAI effectiveness either on their own or in conjunction with user studies:

- ***D***, the performance difference between the agent's model and the performance of the logic presented as an explanation
- ***R***, the number of rules in the agent's explanation
- ***F***, the number of features used to construct the explanation
- ***S***, the stability of the agent's explanation

Elements of these metrics exist in other papers [28, 46, 47, 49, 51, 55], and throughout the paper we point out similarities and differences to previous works. In order to better understand the novelty of this work, we first briefly overview state-of-the-art approaches for generating explanations.

2 HOW EXPLANATIONS ARE GENERATED

Unfortunately, no consensus currently exists about the meaning of various terms related to explainability including interpretability and transparency. Part of the confusion is likely complicated by the fact that the terms, "explainability, interpretability and transparency" are often used synonymously while others implicitly define these terms differently [11, 12, 16, 17, 30, 51, 55, 57]. Previous work by Rosenfeld and Richardson defined interpretability as a technical term focusing on the clarity of the system's internal logic and explainability as the ability of human user to understand that logic [51]. In contrast, Rudin defined explanations as agent attempts to explain its logic in a post-hoc fashion without necessarily being tied to the agent's true decision model, while interpretations are inherently tied to the agent's logic [55]. Both works agree that the XAI goal is to completely, accurately and clearly quantify the agent's logic, something that Rosenfeld and Richardson refer to as transparency [51] and Rudin terms fidelity [55]. To avoid terminology confusion, we will use these terms synonymously as both focus on the same paramount XAI goal.

The level of agent transparency depends on which of three basic approaches are used to generate XAI: directly from a transparent machine learning algorithm, through feature selection and/or analysis of the inputs, or by using an algorithm to create a post-hoc modeling, outcome, or visualization tool. The first approach is to only use certain types of machine learning methods, such as decision trees or other rule-based approaches, that transparently output a model that can readily be understood by the user. For example, if a decision tree outputs a relatively small set of rules, this output can then be directly implemented as the agent's logic **and** serves as the explanation presented to the system's user [51, 55, 58].

A second approach is to use feature selection and analysis to establish which data elements should be focused upon. Even if transparent models are then not used, using a limited set of features can help clarify the agent's logic [19, 26, 51, 55]. The advantage of this approach is that the information presented to the user is generated directly from the mathematical relationship between a small set of features and the target being learned. Additionally, even if the agent uses more complex models machine learning models, this approach helps the user better understand the underlying relationships between the input and output of the system even if she does not fully grasp the full interplay of all input possibilities and the resultant model. This in turn allows the agent to use more accurate models without sacrificing significant fidelity levels [51].

The last, and possibly most prevalent, approach uses mechanisms external to the system's logic to help describe the inner working of a black-box system that is not inherently understood [37, 55]. This approach is often used in conjunction with state-of-the-art prediction models obtained from neural networks and ensemble methods that are not transparent [63]. One group of approaches within this category create proxy models secondary to the agent's logic to approximate the agent's logic via transparent models such as decision trees [9, 20, 62, 71]. Other approaches, such as saliency maps, highlight which portion of the input features, such as areas of a picture, are important based of the structure of model being used– typically in a neural network [1, 61, 70]. A third approach highlights which inputs are important based on model perturbations to query the system for how the agent's performance would be impacted without those inputs [10, 13, 31, 32, 39]. Popular examples of XAI algorithms within this approach are LIME (Local Interpretable Model-Agnostic Explanations) [39] and SHAP (SHapley Additive exPlanation) [31, 32]. While the post-hoc explanation is not a full representation of the system's logic, they do enable people to better understand the system's logic [51] – even in black-box systems.

3 MATCHING XAI WITH ITS NEED

An effective evaluation metric must quantify the benefit of XAI towards achieving the system's goals. These goals stem from various needs including legal, ethics, safety, trust, and knowledge discovery considerations [3, 12, 51, 55]. It is important to differentiate between different types of entities requiring the explanations – whether it is a user interacting with the system or an outside societal or legal body. While XAI research is typically geared for individual users, there is a growing need to address legal and governmental concerns. Both the EU and UK governments have adopted guidelines requiring agent designers to provide users information about computer decisions. In the words of the EU's "General Data Protection Regulation" (GDPR), users are legally entitled to obtain "meaningful explanation of the logic involved" of these decisions and additional legislation exists to ensure that automated decisions are not biased against any ethnic or gender groups [12, 17]. However, demonstrating that a system is generally unbiased or even provides "meaningful" explanations is not the same as providing transparency and full fidelity about the logical process of the system for every possible situation.

Explainability has also been suggested to help the system designer evaluate the system or to confirm that the system is functioning properly and safely. This requirement is particularly acute within life-and-death human-agent systems including the medical application built by the XYZ company. Without XAI, both the medical practitioner and the patient might fear that the agent's recommendations might be adopted wrongly at times and thus put people's lives in danger.

In contrast, certain explanation goals are less critical– such as using explanations for knowledge discovery to help researchers gain understanding of various medical phenomena. Explainability can similarly be useful in building trust between the user and system especially when mistakes were made [8]. XAI is not critical in these cases but could help improve the total utility of the human-agent system. Assuming the agent effectively conveys its logic in the XYZ application, the user could potentially understand when to accept the agent's recommendation and when to ignore them. This would create an ideal decision support system (DSS) by leveraging agent and user strengths.

Some explanations goals are relatively easy to evaluate. We generally believe that explanations built to address legal, ethics, safety, and knowledge discovery considerations can be typically evaluated with a simple or binary score– either the system addresses these

considerations or not. If the explanation lacks fidelity, and thus doesn't truly quantify the agent's logic, then there is no basis to consider the system safe or non-biased. Similarly, if the knowledge discovery is rooted in explanations based on simplified logic that is not being used, then the XAI is independent of the agent's logic, making any gain from this knowledge minimal [55].

More commonly, explanations are important to create better human-agent interactions – either to develop trust or to foster better performance in a DSS [51]. In these cases, evaluating the explanation is more difficult as it must quantify the complex relationship between the agent and human components of the system. As is the case in the XYZ system, the agent's ability to create accurate models is critical for the overall success of the system. As such, any evaluation metric should reason about the joint agent and human performance with classic metrics such accuracy, ROC and/or precision. Assuming the agent performance achieved from black-box models is higher, these models could then be used– even at a potential loss of transparency. Many XAI researchers assume that these black-box models are inherently better due to their superior performance and explainability must be created given this constraint [10, 13, 17, 31, 32, 37, 39]. Accordingly, it is tempting to suggest that transparent models and their explanations be used for life-or-death decisions such as the ones made by the XYZ system, or when legally required, but explanations with less fidelity might be acceptable in other situations [51, 55]. User studies with classic performance and satisfaction metrics can then be used to weigh the effectiveness of various explanations [12, 22, 37, 38, 51, 65]. However, for reasons we now detail, we instead suggest using the ***D, R, F, S*** metrics to quantify XAI.

4 BETTER METRICS FOR XAI EVALUATIONS

In this section we argue that the "gold-standard" for evaluating XAI system through user studies may at times be inherently flawed due to several reasons. First, there is an assumption that user studies can properly capture the complex dynamics between user performance and explanations. This has already been shown to not always be correct. Second, XAI research has assumed that providing better explanations aid group behavior. However, one large-scale study involving 3800 participants did *not* find this to be the case and providing more detailed explanations hurt performance [38]. Even studies seeming to support the benefit from a given explanation algorithm may suffer from confirmation bias, where user studies are constructed to confirm the effectiveness of a wrong hypothesis, here a poor explanation given the XAI's goals [68].

We generally claim that existing user studies that evaluate explanations generated post-hoc with a separate logic and low fidelity to validate legal, ethics or safety concerns are inherently flawed. As Rudin points out, if the post-hoc explanation is based on a fundamentally different logic than the one used by the agent, then what is being evaluated? If they contain the same logic, then why not use that model instead of the black-box model [55]? This critique is particularly an issue regarding explanations created post-hoc via proxy methods as these explanations are a known oversimplification of the agent's logic [9, 20, 28, 62, 71]. Saliency maps have also been found to be equally problematic as these visualization tools are often the same regardless of the specific input used– making their general worth questionable [1]. Thus, even if a user is satisfied with explanations of these types, the positive result is potentially based on confirmation bias. Even when trust and performance needs to be evaluated, and the complex interplay between XAI and performance must be considered, one should question the validity of explanations generated with low fidelity. Given the logical gap between the agent's logic and the explanation, classic user metrics cannot necessarily quantify if a positive result is due to the explanation or confirmation bias. Objective metrics are critical for quantifying the tradeoff between agent fidelity and performance.

To address this challenge, we present four XAI evaluation metrics, ***D, R, F, S***, that are not dependant on the task being performed or the XAI algorithm developed. As a result, these metrics cannot suffer from any confirmation bias. Consequently, ***D, R, F,*** and ***S*** can be used to quantify XAI similarly to how the NASA-TLX [54] and the System Usability Scale [5] quantify elements of user performance.

The first metric, ***D***, is predicated upon the assumption that human-agent system designers used black-box models for the agent because they provided a significant improvement in agent performance [18, 51]. It is still unclear if this tradeoff is typically necessary for most applications or even for specialized tasks such as image processing where neural networks are typically used [55]. To evaluate this tradeoff, ***D*** quantifies the change of agent performance, δ, between the black box model, and the best observed transparent model. This measure is similar to the disagreement metric previously developed [28] but uses δ to quantify performance differences between models to facilitate comparing different types of explanations and the potential improvement in performance versus the loss of fidelity. Even if a user is happy with an explanation, δ helps measure if the tradeoff between this explanation and one with transparency was warranted by comparing the performance, P_t of the transparent model and the performance, $P_b - \delta$, of the black box model. For example, assuming XYZ's black box model is 99.5% accurate, but a transparent model is 95.5% accurate, δ would have to be less than 0.04 to justify using the black box model. Similarly, if a model based on feature construction is 99% accurate, δ would have to be less than 0.005 to justify using the black box model. User studies could then focus on what value for δ is most justified for a given XAI goal and specific task.

The ***D*** metric can be useful both in cases where binary and non-binary evaluation of XAI is warranted, but will typically be more helpful in the later case. Assuming binary evaluation is needed due to legal, ethics or safety issues, any value for δ greater than a trivial value ϵ shows that the explanation and agent are not synonymous, and thus any benefit from the explanation is likely nil. For example, if an explanation is used to show lack of bias in XYZ's system, but the logic based on the explanation is different for certain cases, then close inspection is needed to evaluate if these cases represent a bias or can be ignored. Conversely, if δ is zero then the two models are equivalent and the more transparent model should be used regardless. Thus, we believe ***D*** is more useful when XAI is beneficial to the system, but not necessary.

While ***D*** focuses on performance differences between models, the second metric, ***R***, focuses on the size of the agent's explanation without comparison to other models. ***R*** quantifies explanations based on their simplicity – the fewer rules in the explanation, the

better. This metric is built upon an assumption that simpler explanations should be preferred as per Occam's Razor [51] and work by Gigerenzer and Brighton about the bias-variance tradeoff [15]. These works, among others, assume that the world is inherently orderly and understandable by relatively simple rules. Thus, complex rules should be penalized.

While many utility functions are possible for ***R***, similar to Rudin [55] we suggest using a parameter λ to quantify the number of rules within the agent's model. In contrast to their work, we suggest that a penalty of $\lambda * \mathcal{L}$ be used to penalize the performance metric where $\mathcal{L} = size(m) - c$. We define size(m) as the number of rules in the explanation. We set c=1 by default but this value can be set to larger values to show that explanations with this number of rules are fully explainable and should not be penalized. Formally, we define $\mathcal{L} = size(m) - c$ where (size(m)-c)>0, and zero otherwise. For example, assuming the size of the model, size(m), is 1, no penalty is added regardless of the value for λ. If $\lambda = 0.005$, the default value for c is used and 5 rules exist in the explanation, then a 0.02 performance penalty is added for the four rules above the base size of 1. Alternatively, no penalty may be desired for any model less with fewer than 5 rules, and c=5 can be set. As such, a penalty will only exist when 6 or more rules are in the explanation as per $\mathcal{L}$'s definition. In all cases the $\lambda * \mathcal{L}$ penalty could be optimized and evaluated based on theoretical or user studies.

We assume that the ***R*** metric is most useful for transparent methods, but it can also be in conjunction with other XAI algorithms as well. While decision trees are generally assumed to be transparent models [51], one would be pressed to consider a decision tree with thousands of rules as being explainable. Thus, $\lambda * \mathcal{L}$ can quantify the impact of this complexity on the system's performance. Conversely, while neural networks and ensemble methods are typically assumed to be non-transparent, given a small enough model size they may be understood by the intended user. We suggest user studies based on ***R*** be performed to further analyze these claims.

While the ***R*** metric focuses on quantifying the number of rules in the agent's output, the ***F*** metric focuses on the number of features inputted by agent to create its explanation. This metric was particularly built for those explanations based on feature analysis. Even when the agent's model is based on a complex learning model with lower fidelity, the assumption is that explainability will be higher if the user can focus on a smaller number of features, thus making the XAI clearer. To quantify this relationship, we again suggest that a penalty of $\lambda * \mathcal{L}$ be defined similarly to ***R***, but here we define size(m) as the number of inputs to the model instead of the rules outputted. The threshold c can again be used to quantify a maximum number of feature inputs where no penalty should be applied. It is possible that a magical number of around 7 [33] be used for c in both ***R*** and ***F***, but further studies are needed given a specific XAI goal/task.

We posit that the size(m) value in the ***R*** and ***F*** metrics need not based on a single feature or field, and more complex constructed features be considered as one unit for purposes of size(m). For example, image processing typically currently focuses on pixels inputs, but some constructed image features such as edges are inherently more interpretability and might be considered as a single feature for purposes of these metrics. Similarly, a complex driving style feature was previously found to be useful in quantifying people's use of adaptive cruise control [45] and the maximum cancer core length feature was found to be an important feature in quantifying the existence of prostate cancer [7]. As a general rule, we suggest that derived features be treated the same as non-derived ones as long as they are equally understood by the intended user.

The last evaluation metric we present, ***S***, quantifies the stability of the agent's explanation. Feature stability is a metric central to feature selection analysis and refers to its ability to robustly handle *small* noise perturbations. Finding stable features is important as it indicates that the feature selection is unlikely to be overfitted to the specific data being considered. Unstable feature inputs have been linked to poor explainability [35]. We suggest boostrapping the data and then observing its impact on the outputted agent explanations. The similarity between the bootstraps' explanations can be quantified using Jaccard and/or Tanimoto similarity measures. We stress that only small perturbations should be used by the bootraps such that the class labels are not changed and small amounts of resampling noise are used to check that the explanations are stable and thus general. At the other extreme, "the data randomization test", was created that randomly permutes all labels within the training classes [1]. This check also uses similarity to evaluate if an explanation is useful, but here a lack of similarity shows that explanations are not dependent on the data being used, rendering them without value. In all cases, and as was the case for the ***R*** and ***F*** metrics, a penalty cost, $\lambda * \mathcal{L}$ can be used to penalize the performance metric. As similarity metrics are typically between 0-1, with 1 being complete similarity, we suggest directly defining ***S***=λ*(1-similarity) such that performance-(λ*(1-similarity)) could be used as a penalty. Once again, user studies could be used to help compare different methods and appropriately set the values for λ.

While we have presented each of these metrics individually, we believe that they are complimentary in many cases. It is likely desirable that a transparent explanation be both stable, perform similarly to agent, and contain relatively few rules. As such one might expect it to have high scores for the ***D, R, S*** metrics. Conversely, any agent using black box methods might score lower for the ***D, F*** metrics especially if a proxy model is used for generating explanations, but if this explanation is relatively simple and stable it would still achieve higher scores for the ***R, S*** metrics. Thus, composite scores could be constructed and focused user studies should be conducted.

5 CONCLUSION

In this paper we argue that many XAI studies wrongly assume low fidelity explanations should be accepted for certain tasks. We also argue that user studies may also have confirmation bias in their evaluation of XAI. To remove these concerns, we advocate using four general metrics, ***D, R, F, S***, to quantify XAI explainability based on the difference in the agent's performance using models with higher fidelity versus lower fidelity, the number of rules in the outputted explanation, the number of features used by the agent to generate the explanation, and the stability of the agent's explanation. The advantage of these measures is that they make no a-priori assumption about the relatively advantage of using an XAI algorithm with higher or lower fidelity, yet facilitate comparison without any potential confirmation bias from user studies. We hope that these metrics will be considered in the future for what we consider to be more significant evaluations of XAI.

REFERENCES

[1] Julius Adebayo, Justin Gilmer, Michael Muelly, Ian Goodfellow, Moritz Hardt, and Been Kim. 2018. Sanity checks for saliency maps. In *Advances in Neural Information Processing Systems*. 9505–9515.

[2] Ofra Amir and Kobi Gal. 2013. Plan recognition and visualization in exploratory learning environments. *ACM Transactions on Interactive Intelligent Systems (TiiS)* 3, 3 (2013), 16.

[3] Alejandro Barredo Arrieta, Natalia Díaz-Rodríguez, Javier Del Ser, Adrien Bennetot, Siham Tabik, Alberto Barbado, Salvador García, Sergio Gil-López, Daniel Molina, Richard Benjamins, et al. 2020. Explainable Artificial Intelligence (XAI): Concepts, taxonomies, opportunities and challenges toward responsible AI. *Information Fusion* 58 (2020), 82–115.

[4] Amos Azaria, Zinovi Rabinovich, Claudia V Goldman, and Sarit Kraus. 2015. Strategic information disclosure to people with multiple alternatives. *ACM Transactions on Intelligent Systems and Technology (TIST)* 5, 4 (2015), 64.

[5] Aaron Bangor, Philip T Kortum, and James T Miller. 2008. An empirical evaluation of the system usability scale. *Intl. Journal of Human–Computer Interaction* 24, 6 (2008), 574–594.

[6] Samuel Barrett, Avi Rosenfeld, Sarit Kraus, and Peter Stone. 2017. Making friends on the fly: Cooperating with new teammates. *Artificial Intelligence* 242 (2017), 132–171.

[7] Echeverria LM Carmona, A Haider, A Freeman, U Stopka-Farooqui, A Rosenfeld, BS Simpson, Y Hu, D Hawkes, H Pye, S Heavey, V Stavrinides, JM Norris, AE-S Bosaily, Barrena C Cardona, S Bott, L Brown, N Burns-Cox, T Dudderidge, A Henderson, R Hindley, R Kaplan, A Kirkham, R Oldroyd, M Ghei, R Persad, S Punwani, D Rosario, I Shergill, M Winkler, HU Ahmed, M Emberton, and HC Whitaker. 2020. A critical evaluation of visual proportion of Gleason 4 and maximum cancer core length quantified by histopathologists. *Sci Rep* 10 (2020).

[8] Jessie Y Chen, Katelyn Procci, Michael Boyce, Julia Wright, Andre Garcia, and Michael Barnes. 2014. *Situation awareness-based agent transparency*. Technical Report. Army Research Lab Aberdeen Proving Ground MD Human Research and Engineering Directorate.

[9] Houtao Deng. 2014. Interpreting tree ensembles with intrees. *arXiv preprint arXiv:1408.5456* (2014).

[10] Amit Dhurandhar, Pin-Yu Chen, Ronny Luss, Chun-Chen Tu, Paishun Ting, Karthikeyan Shanmugam, and Payel Das. 2018. Explanations based on the missing: Towards contrastive explanations with pertinent negatives. In *Advances in Neural Information Processing Systems*. 592–603.

[11] Derek Doran, Sarah Schulz, and Tarek R. Besold. 2017. What Does Explainable AI Really Mean? A New Conceptualization of Perspectives. In *Proceedings of the First International Workshop on Comprehensibility and Explanation in AI and ML*.

[12] Finale Doshi-Velez and Been Kim. 2017. Towards a rigorous science of interpretable machine learning. *arXiv preprint arXiv:1702.08608* (2017).

[13] Ruth C Fong and Andrea Vedaldi. 2017. Interpretable explanations of black boxes by meaningful perturbation. In *2017 IEEE international conference on computer vision (ICCV)*. 3449–3457.

[14] Maria Fox, Derek Long, and Daniele Magazzeni. 2017. Explainable Planning. *CoRR* abs/1709.10256 (2017).

[15] Gerd Gigerenzer and Henry Brighton. 2009. Homo heuristicus: Why biased minds make better inferences. *Topics in cognitive science* 1, 1 (2009), 107–143.

[16] Leilani H. Gilpin, David Bau, Ben Z. Yuan, Ayesha Bajwa, Michael Specter, and Lalana Kagal. 2018. Explaining Explanations: An Approach to Evaluating Interpretability of Machine Learning. *CoRR* abs/1806.00069 (2018).

[17] Riccardo Guidotti, Anna Monreale, Salvatore Ruggieri, Franco Turini, Fosca Giannotti, and Dino Pedreschi. 2018. A Survey of Methods for Explaining Black Box Models. *ACM Comput. Surv.* 51, 5 (Aug. 2018), 93:1–93:42.

[18] David Gunning. 2017. Explainable artificial intelligence (xai). *Defense Advanced Research Projects Agency (DARPA), nd Web* 2, 2 (2017).

[19] Isabelle Guyon and André Elisseeff. 2003. An introduction to variable and feature selection. *Journal of machine learning research* 3 (2003), 1157–1182.

[20] Satoshi Hara and Kohei Hayashi. 2016. Making tree ensembles interpretable. *arXiv preprint arXiv:1606.05390* (2016).

[21] Nicholas Hoernle, Kobi Gal, Barbara J. Grosz, Leilah Lyons, Ada Ren, and Andee Rubin. 2020. Interpretable Models for Understanding Immersive Simulations. In *Proceedings of the Twenty-Ninth International Joint Conference on Artificial Intelligence, IJCAI 2020*, Christian Bessiere (Ed.). ijcai.org, 2319–2325.

[22] Robert R Hoffman, Shane T Mueller, Gary Klein, and Jordan Litman. 2018. Metrics for explainable AI: Challenges and prospects. *arXiv preprint arXiv:1812.04608* (2018).

[23] Yoshimasa Horie, Toshiyuki Yoshio, Kazuharu Aoyama, Shoichi Yoshimizu, Yusuke Horiuchi, Akiyoshi Ishiyama, Toshiaki Hirasawa, Tomohiro Tsuchida, Tsuyoshi Ozawa, Soichiro Ishihara, et al. 2019. Diagnostic outcomes of esophageal cancer by artificial intelligence using convolutional neural networks. *Gastrointestinal endoscopy* 89, 1 (2019), 25–32.

[24] Nicholas R Jennings, Luc Moreau, David Nicholson, Sarvapali Ramchurn, Stephen Roberts, Tom Rodden, and Alex Rogers. 2014. Human-agent collectives. *Commun. ACM* 57, 12 (2014), 80–88.

[25] Akiva Kleinerman, Ariel Rosenfeld, and Sarit Kraus. 2018. Providing explanations for recommendations in reciprocal environments. In *Proceedings of the 12th ACM Conference on Recommender Systems*. ACM, 22–30.

[26] Igor Kononenko. 1999. Explaining classifications for individual instances. In *In Proceedings of IJCAI'99*. 722–726.

[27] Sarit Kraus, Amos Azaria, Jelena Fiosina, Maike Greve, Noam Hazon, Lutz Kolbe, Tim-Benjamin Lembcke, Jorg P Muller, Soren Schleibaum, and Mark Vollrath. 2020. AI for explaining decisions in multi-agent environments. In *Proceedings of the AAAI Conference on Artificial Intelligence*, Vol. 34. 13534–13538.

[28] Himabindu Lakkaraju, Ece Kamar, Rich Caruana, and Jure Leskovec. 2017. Interpretable & explorable approximations of black box models. *arXiv preprint arXiv:1707.01154* (2017).

[29] Pat Langley, Ben Meadows, Mohan Sridharan, and Dongkyu Choi. 2017. Explainable Agency for Intelligent Autonomous Systems.. In *AAAI*. 4762–4764.

[30] Zachary Chase Lipton. 2016. The Mythos of Model Interpretability. *arXiv preprint arXiv:1606.05390* (2016).

[31] Scott M Lundberg, Gabriel Erion, Hugh Chen, Alex DeGrave, Jordan M Prutkin, Bala Nair, Ronit Katz, Jonathan Himmelfarb, Nisha Bansal, and Su-In Lee. 2020. From local explanations to global understanding with explainable AI for trees. *Nature machine intelligence* 2, 1 (2020), 2522–5839.

[32] Scott M Lundberg and Su-In Lee. 2017. A unified approach to interpreting model predictions. In *Advances in Neural Information Processing Systems*. 4765–4774.

[33] George A Miller. 1956. The magical number seven, plus or minus two: Some limits on our capacity for processing information. *Psychological review* 63, 2 (1956), 81.

[34] Shamim Nemati, Andre Holder, Fereshteh Razmi, Matthew D Stanley, Gari D Clifford, and Timothy G Buchman. 2018. An interpretable machine learning model for accurate prediction of sepsis in the ICU. *Critical care medicine* 46, 4 (2018), 547.

[35] Sarah Nogueira, Konstantinos Sechidis, and Gavin Brown. 2017. On the stability of feature selection algorithms. *The Journal of Machine Learning Research* 18, 1 (2017), 6345–6398.

[36] Erfan Pakdamanian, Shili Sheng, Sonia Baee, Seongkook Heo, Sarit Kraus, and Lu Feng. 2021. DeepTake: Prediction of Driver Takeover Behavior using Multimodal Data. In *Proc. of CHI-21*.

[37] Dino Pedreschi, Fosca Giannotti, Riccardo Guidotti, Anna Monreale, Salvatore Ruggieri, and Franco Turini. 2019. Meaningful Explanations of Black Box AI Decision Systems. *Proceedings of the AAAI Conference on Artificial Intelligence* 33 (Jul. 2019), 9780–9784.

[38] Forough Poursabzi-Sangdeh, Daniel G Goldstein, Jake M Hofman, Jennifer Wortman Vaughan, and Hanna Wallach. 2018. Manipulating and measuring model interpretability. *arXiv preprint arXiv:1802.07810* (2018).

[39] Marco Tulio Ribeiro, Sameer Singh, and Carlos Guestrin. 2016. Why Should I Trust You?: Explaining the Predictions of Any Classifier. In *Proceedings of the 22Nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*. 1135–1144.

[40] Ariella Richardson and Avi Rosenfeld. 2018. A Survey of Interpretability and Explainability in Human-Agent Systems. *XAI 2018* (2018), 137–143.

[41] Hanan Rosemarin, Ariel Rosenfeld, and Sarit Kraus. 2019. Emergency department online patient-caregiver scheduling. In *Proceedings of the AAAI Conference on Artificial Intelligence*, Vol. 33. 695–701.

[42] Ariel Rosenfeld, Noa Agmon, Oleg Maksimov, and Sarit Kraus. 2017. Intelligent agent supporting human–multi-robot team collaboration. *Artificial Intelligence* 252 (2017), 211–231.

[43] Avi Rosenfeld, Zevi Bareket, Claudia V Goldman, Sarit Kraus, David J LeBlanc, and Omer Tsimhoni. 2012. Learning Driver's Behavior to Improve the Acceptance of Adaptive Cruise Control.. In *IAAI*.

[44] Avi Rosenfeld, Zevi Bareket, Claudia V Goldman, Sarit Kraus, David J LeBlanc, and Omer Tsimhoni. 2012. Towards adapting cars to their drivers. *AI Magazine* 33, 4 (2012), 46–46.

[45] Avi Rosenfeld, Zevi Bareket, Claudia V Goldman, David J LeBlanc, and Omer Tsimhoni. 2015. Learning drivers' behavior to improve adaptive cruise control. *Journal of Intelligent Transportation Systems* 19, 1 (2015), 18–31.

[46] Avi Rosenfeld and Matanya Freiman. 2020. Explainable Feature Ensembles through Homogeneous and Heterogeneous Intersections. In *IJCAI-PRICAI 2020 Workshop on XAI*.

[47] Avi Rosenfeld, David G. Graham, Rifat Hamoudi, Rommel Butawan, Victor Eneh, Saif Khan, Haroon Miah, Mahesan Niranjan, and Laurence B. Lovat. 2015. MIAT: A Novel Attribute Selection Approach to Better Predict Upper Gastrointestinal Cancer. In *International Conference on Data Science and Advanced Analytics*.

[48] Avi Rosenfeld, David G Graham, Sarah Jevons, Jose Ariza, Daryl Hagan, Ash Wilson, Samuel J Lovat, Sarmed S Sami, Omer F Ahmad, Marco Novelli, et al. 2020. Development and validation of a risk prediction model to diagnose Barrett's oesophagus (MARK-BE): a case-control machine learning approach. *The Lancet Digital Health* 2, 1 (2020), e37–e48.

[49] Avi Rosenfeld, Ron Illuz, Dovid Gottesman, and Mark Last. 2018. Using discretization for extending the set of predictive features. *EURASIP Journal on Advances in Signal Processing* 7, 1 (2018), 1–11.

[50] Ariel Rosenfeld and Sarit Kraus. 2016. Providing Arguments in Discussions on the Basis of the Prediction of Human Argumentative Behavior. *ACM Trans. Interact. Intell. Syst.* 6, 4 (2016), 30:1–30:33.

[51] Avi Rosenfeld and Ariella Richardson. 2019. Explainability in human-agent systems. *Auton. Agents Multi Agent Syst.* 33, 6 (2019), 673–705.

[52] Avi Rosenfeld, Vinay Sehgal, David G. Graham, Matthew R. Banks, Rehan J. Haidry, and Laurence B. Lovat. 2014. Using Data Mining to Help Detect Dysplasia: Extended Abstract. In *2014 IEEE International Conference on Software Science, Technology and Engineering*. IEEE, 65–66.

[53] Avi Rosenfeld, Inon Zuckerman, Erel Segal-Halevi, Osnat Drein, and Sarit Kraus. 2016. NegoChat-A: a chat-based negotiation agent with bounded rationality. *Autonomous Agents and Multi-Agent Systems* 30, 1 (2016), 60–81.

[54] Susana Rubio, Eva Díaz, Jesús Martín, and José M Puente. 2004. Evaluation of subjective mental workload: A comparison of SWAT, NASA-TLX, and workload profile methods. *Applied Psychology* 53, 1 (2004), 61–86.

[55] Cynthia Rudin. 2019. Stop explaining black box machine learning models for high stakes decisions and use interpretable models instead. *Nature Machine Intelligence* 1, 5 (2019), 206–215.

[56] Maha Salem, Gabriella Lakatos, Farshid Amirabdollahian, and Kerstin Dautenhahn. 2015. Would you trust a (faulty) robot?: Effects of error, task type and personality on human-robot cooperation and trust. In *Proceedings of the Tenth Annual ACM/IEEE International Conference on Human-Robot Interaction*. 141–148.

[57] Wojciech Samek, Thomas Wiegand, and Klaus-Robert Müller. 2017. Explainable artificial intelligence: Understanding, visualizing and interpreting deep learning models. *arXiv preprint arXiv:1708.08296* (2017).

[58] Vinay Sehgal, Avi Rosenfeld, David G Graham, Gideon Lipman, Raf Bisschops, Krish Ragunath, Manuel Rodriguez-Justo, Marco Novelli, Matthew R Banks, Rehan J Haidry, et al. 2018. Machine learning creates a simple endoscopic classification system that improves dysplasia detection in Barrett's oesophagus amongst non-expert endoscopists. *Gastroenterology Research and Practice* (2018).

[59] Raymond Sheh. 2017. why did you do that?" explainable intelligent robots. In *AAAI Workshop on Human-Aware Artificial Intelligence*.

[60] Maarten Sierhuis, Jeffrey M Bradshaw, Alessandro Acquisti, Ron Van Hoof, Renia Jeffers, and Andrzej Uszok. 2003. Human-agent teamwork and adjustable autonomy in practice. In *Proceedings of the seventh international symposium on artificial intelligence, robotics and automation in space (I-SAIRAS)*.

[61] Karen Simonyan, Andrea Vedaldi, and Andrew Zisserman. 2013. Deep inside convolutional networks: Visualising image classification models and saliency maps. *arXiv preprint arXiv:1312.6034* (2013).

[62] Hui Fen Tan, Giles Hooker, and Martin T Wells. 2016. Tree Space Prototypes: Another Look at Making Tree Ensembles Interpretable. *arXiv preprint arXiv:1611.07115* (2016).

[63] Erico Tjoa and Cuntai Guan. 2020. A Survey on Explainable Artificial Intelligence (XAI): Toward Medical XAI. *IEEE Transactions on Neural Networks and Learning Systems* (2020).

[64] David Traum, Jeff Rickel, Jonathan Gratch, and Stacy Marsella. 2003. Negotiation over tasks in hybrid human-agent teams for simulation-based training. In *Proceedings of the second international joint conference on Autonomous agents and multiagent systems*. ACM, 441–448.

[65] Jasper van der Waa, Elisabeth Nieuwburg, Anita Cremers, and Mark Neerincx. 2021. Evaluating XAI: A comparison of rule-based and example-based explanations. *Artificial Intelligence* (2021).

[66] Kurt VanLehn. 2011. The relative effectiveness of human tutoring, intelligent tutoring systems, and other tutoring systems. *Educational Psychologist* 46, 4 (2011), 197–221.

[67] Alfredo Vellido. 2019. The importance of interpretability and visualization in machine learning for applications in medicine and health care. *Neural Computing and Applications* (2019), 1–15.

[68] Danding Wang, Qian Yang, Ashraf Abdul, and Brian Y Lim. 2019. Designing theory-driven user-centric explainable AI. In *Proceedings of the 2019 CHI conference on human factors in computing systems*. 1–15.

[69] Bo Xiao and Izak Benbasat. 2007. E-commerce product recommendation agents: use, characteristics, and impact. *MIS quarterly* 31, 1 (2007), 137–209.

[70] Bolei Zhou, Aditya Khosla, Agata Lapedriza, Aude Oliva, and Antonio Torralba. 2016. Learning deep features for discriminative localization. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. 2921–2929.

[71] Yichen Zhou and Giles Hooker. 2016. Interpreting Models via Single Tree Approximation. *arXiv preprint arXiv:1610.09036* (2016).

[72] Guangming Zhu, Bin Jiang, Hui Chen, Elizabeth Tong, Yuan Xie, Tobias D Faizy, Jeremy J Heit, Greg Zaharchuk, and Max Wintermark. 2020. Artificial Intelligence and Stroke Imaging: A West Coast Perspective. *Neuroimaging Clinics* 30, 4 (2020), 479–492.

[73] Inon Zuckerman, A Rosenfeld, Sarit Kraus, and E Segal-Halevi. 2013. Towards automated negotiation agents that use chat interfaces. In *The sixth international workshop on agent-based complex automated negotiations (ACAN)*.