A Survey of Interpretability and Explainability in Human-Agent Systems

Ariella Richardson[1], **Avi Rosenfeld**[1],
[1] Jerusalem College of Technology, Jerusalem 91160, Israel
rosenfa@jct.ac.il, richards@jct.ac.il

Abstract

This paper presents a taxonomy of interpretability in Human-Agent Systems. We consider four fundamental questions, "Why, what, when, and how" about interpretability. First, we consider why interpretability is needed in the system. Second, once interpretability is established as being needed, we consider what explanations can be generated to meet this need. Third, we consider when the user should be presented this information. Fourth, we consider the level of detail needed within explanations. Last, we consider how objective and subjective measures can be used to evaluate the entire system including the four fundamental questions regarding interpretability.

1 Introduction

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, natural language processing or chat bots, planning and self-driving cars [35; 39; 40; 41; 32; 23; 18; 30; 9; 26]. One key question surrounding these systems is the type and quality of the information that must be shared between the agents and the human-users during their interactions.

This paper focuses on one aspect of this human-agent interaction — the internal level of interpretability that the agents must present as potential explanations for their decisions within agents that use machine learning. Following previous work by Doshi-Velez and Kim, we refer to interpretability as the ability of an agent to explain or to present its decision to a human user, in understandable terms [8]. Given this definition, a system's explanation constitutes its interpretability and thus we will interchange use of both of these terms. This definition is intentionally left general to encompass the fields of Explainable AI (XAI) [13], where the explanations are geared towards people, and Interpretable Machine Learning [8] which primarily focuses on a machine learning algorithm's output. Despite an emerging interest in these fields, a limited number of works have addressed interpretability within human-agent systems.

We believe that the novelty within this paper is its ability to provide an interpretability framework that unifies elements of previous work [8; 10; 24; 13]. Specifically, we focus on four questions within Human-Agent Systems. We first address, "**Why** should a Human-Agent System be interpretable?" We propose a taxonomy of three reasons of interpretability: not-necessary (or marginally necessary), beneficial, and critical and discuss examples of each of these possibilities. This question must first be asked, as the its answer will likely impact the following issues. Second, we discuss, "**What** explanation can be generated?" We present three possible solutions for how to generate the information needed to make the system interpretable. At times, the explanations can be directly derived from the machine learning algorithm. In other cases, feature selection and/or analysis is used to generate explanations. Last, one might use a secondary explanation algorithm in addition to the machine learning algorithm. The third question, "**When** information should be presented?", addresses when the explanation should be presented: before, during or after the task is completed. We believe that the answer to the fourth question, "**How** much detail is presented in an explanation?" should be tied to the user's level of expertise and the type of system needed. The final issue we discuss is how explanations should be evaluated based on how well the human-agent system answers each of the previous four issues. We posit that both objective and subjective measures should be considered when evaluating such systems. Care must also be taken to address how these explanations may or may not have impacted the accuracy of machine learning algorithms and / or the user's productivity and satisfaction with the system.

Overall, we believe that the solutions presented to all of these issues need to be considered in tandem as they are intertwined. They directly depend on the type of human-agent system needing to be developed and thus directly stem from the first question about the overall reason, or reasons, that the system must be interpretable. Assuming that the system is human-centric, as is the case in recommendation [41], training [39], and tutoring systems [40], then the information will likely need to persuade the person to choose a certain action, for example through arguments about the agent's decision [27] or by using a suited presentation type [2]. If the system is agent-centric, such as in knowledge discovery or

self-driving cars, the agent might need to provide information about its decision to help convince the human participant of the correctness of their solution, aiding in the adoption of these technologies [25]. Furthermore, these explanations might be necessary for legal considerations [8]. In all cases we need to consider and then evaluate how these explanations were generated, presented, and if their level of detail correctly matches the system's need.

2 Why a Human-Agent System should be Interpretable

We posit that one can generalize the need for interpretability with a taxonomy of three levels:

1. Not helpful
2. Beneficial
3. Critical

Adjustable autonomy is a well-established concept within human-agent and human-robot groups that refers to the amount of control an agent/robot has compared to the human-user [11; 42; 31]. Under this approach, the need for interpretability can be viewed as a function of the degree of co-operation between the agent to the human user. Assuming the agent is fully controlled by the human operator (e.g. tele-operated), then no interpretability is needed as the agent is fully an extension of the human participant. Conversely, if the robot is given full control, particularly if the reason for the decision is obvious (a recommendation agent gives advice based on a well-established collaborative filtering algorithm), it again reasons that no interpretability is needed. Additionally, Doshi-Velez and Kim pointed out that explanation at times is not needed if there are no significant consequences for unacceptable results or the agent's decision is universally accepted and trusted [8].

At the other extreme, many Human-Agent systems are built whereby the agent's role is to support a human's task. In many of these cases, we argue that the agent's interpretation is a critical element within the system. For example, Intelligent Tutoring Systems (ITS) typically use step-based granularities of interaction whereby the agent confirms one skill has been learned or uses hints to guide the human participant [40]. The system must provide concrete explanations for its guidance (called *hints* in ITS terminology) to better guide the user. Similarly, explanations form a critical component of many negotiation, training, and argumentation human-agent systems [39; 35; 27; 29]. For example, explanations might be critical to aid a person in making the final life-or-death decision within human-agent systems [35]. Rosenfeld's et. al's NegoChat-A negotiation agent uses arguments to present the logic behind its position [29]. Traum et. al explained the justification within choices of their training agent to better convince the trainee, as well as teach the factors to look at in making decisions [39]. Rosenfeld and Kraus created agents that use argumentation to better persuade people to engage in positive behaviors, such as choosing healthier foods to eat [27]. Azaria et al. demonstrate how an agent that learns the best presentation method for proposals given to a user improves their acceptance rate [2]. Many of these systems can be generally described as Decision Support Systems (DSS). A DSS is typically defined as helping people make semi-structured decisions requiring some human judgment and at the same time with some agreement on the solution method [1]. An agent's effective explanation is critical within a DSS as the system's goal is providing the information to help facilitate improved user decisions.

A middle category in our taxonomy exists when interpretability is beneficial, but not critical. The Merrian-Webster dictionary defines beneficial as something that "produces good or helpful results"[1]. In general, we acknowledge that the value of interpretability can range widely between systems in this category. However, the defining characteristic is that the explanation provided is not inherently needed in order for the system to behave optimally or with peak efficiency. Instead, explanations in these cases can be helpful for **knowledge discovery** beyond the current primary objective of this system and might help further a second, but related goal. For example, a medical diagnostic system may work with peak efficiency exclusively as a *black box*. Nonetheless an agent that provides an explanation for its decision might further human understanding of a medical phenomenon. Along these lines, both the EU and UK governments have expressed a desire to confirm that AI algorithms are not biased against any ethic or gender groups [8]. Here again, the system's explanation is not critical for effective performance of the agent, but instead to confirm that a secondary, legal, requirement is being met. In fact, as we will further explore in the next section, an explainable model might even be done at a sacrifice to the system's performance.

Interpretability is sometimes beneficial to instill feelings of **trust and understanding** within the system's users. People are slow to adapt systems that they do not understand and trust: "If the users do not trust a model or a prediction they will not use it." [25]. Here, while the need for explanations may not arise as a result of the human-agent interaction (as it does in the extremes of full control by the agent or the human), yet these feelings are beneficial to help the human user work more effectively with the system or discard it when it is unreliable. To this end, Ribeiro et al. demonstrate how interpretability is important for identifying models that have high accuracy for the wrong reasons [25]. For example, they show that text classification often are wrongly based on the heading rather than the content. In contrast, image classifiers that capture the main part of the image in a similar manner to the human eye, install a feeling that the model is functioning correctly even if accuracy is not particularly high. Feelings of trust are especially important in certain domains, such as health-care [6], recommender systems [20], planning [9] and human-robot systems [30]. Furthermore, if the agent makes a mistake, the generated reasons could provide *transparency*, thus making it more likely the user will trust the agent in the future [9].

The benefit from this type of explanation will likely vary greatly across systems. If the user will not accept the system without this explanation, then a critical need for interpretabil-

[1] https://www.merriam-webster.com/dictionary/benefit

ity exists. However, in many, if not most, cases, the explanation is beneficial to the system's acceptance and / or to foster better trust. As the goal of interperability is different (e.g. improving user acceptance and trust instead of being vital to the success of the system), it reasons that the type of explanation may be fundamentally different. This in turn may impact the type of information the agent must present, something we address in Sections 4 and 5.

3 What Explanation can be Generated

Once we have established the **why** about explanations, a key related question one must address is **what** explanation can be generated. In addressing this point, we posit that three basic approaches exist as to how explanations can be generated:

1. Directly based on the machine learning algorithm
2. Feature selection and / or analysis
3. Using an explanation algorithm to create a visualization or explanation tool separate from the learning algorithm

The first approach, and possibly the most direct method, is to generate explanations directly from the output of the machine learning algorithm. For example, basic Case Based Reasoning algorithms will typically use k-nn algorithms to provide an explanation that two items are similar [37]. Decision trees have also been shown to be explainable, especially if the size of the tree is limited and have already been established for their ability to explain the model in real-world systems [10; 13; 28]. These two algorithms are potentially clearly explainable machine learning algorithms that could be used to provide critically important levels of explanation. A clear downside to this approach is that one is then limited to a specific algorithm, or a specific algorithmic implementation, that provides the necessary explanations, even when it is less accurate than other alternatives. It has been previously noted that an inverse relationship often exists between machine learning algorithms' accuracy and their explainability [13]. *Black box* algorithms, especially deep neural networks, are often used due to their exceptional accuracy on some problems. However, these types of algorithms are difficult to glean explanations from. Figure 1 is based on previous work [13; 7] and quantifies the general relationship between algorithms' explainability and accuracy. One possible solution to achieve both high accuracy and explainability is to create new machine learning algorithms that explicitly consider the amount of explanation needed by the system. One example of this approach is work by Kim, Rudin and Shah who have suggested a Bayesian Case Model for case-based reasoning [19].

A second approach is to perform feature selection and / or feature analysis before or after a model has been built. Feature selection has long been established as an effective way of building potentially better models which are simpler and thus better overcome the curse of dimensionality [14]. Additionally, models with fewer attributes are potentially easier to understand as the true causal relationship between the dependent and independent variables is clearer and thus easier to present to the user [21]. The strong 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. However, such approaches are inherently limited if models generated by this approach are not sufficiently accurate and / or significantly better models can be generated by less explainable models (e.g. deep learning). Furthermore, even methods that are typically more understandable, such as decision trees, are less interpretable if the size of the tree is too large to fully understand (e.g. thousands of rules) [15].

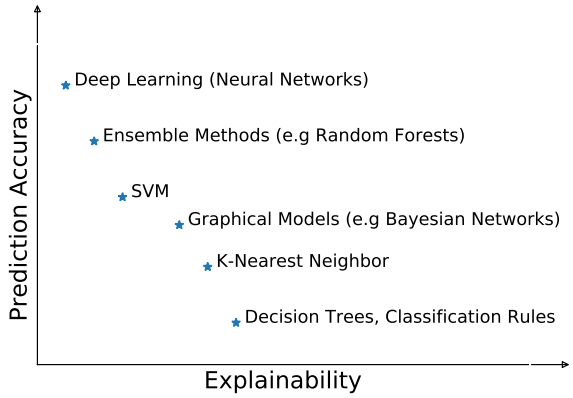

Figure 1: : Prediction accuracy versus explainability

A third approach is to use an algorithm in addition to the agent's machine learning and then use this information to provide a visualization or text-based explanation. This type of approach can be defined as a metacognition process [5], or the process of reasoning about the machine learning process. One example of theses algorithms is LIME (Local Interpretable Model-Agnostic Explanations), which provides explanations without any connection to the type of machine learning algorithm used [25]. It is noteworthy that LIME includes feature engineering as part of its analysis, showing the potential connection between the second and third approaches. A second possibility includes algorithms that provide visualization tools as explanations. For example, saliency maps can provide a visualization tool that is effective in better understanding image classification in deep networks [36]. Visualization of the algorithm output, when possible, is also beneficial to interpretability, such as visualizing a decision tree [32]. A third possibility focuses on developing explanations of algorithms through text explanations. For example, such tools have been suggested for generating explanations of deep networks for image classification [17].

Legal and practical considerations might limit researchers as to what constitutes a sufficient explanation, especially if only certain algorithms provide the critically needed level of interpretability (see Section 2). This in turn impacts what type of explanation approach can be implemented. Additionally, the decision on **when** this information is presented can often be connected to the type of explanation that has been generated, something we further explore in the next section.

4 When Should Information be Presented

Interpretations can be categorized based on when the explanation is made:

1. Before the task
2. Continued explanations during the task
3. After the task

Some agents may present their explanation **before** the task is executed as either justification, conceptualization or explanation for an agent's plan of action. Other agents may present their explanation **during** task execution, especially if this information is important to explain when the agent fails so it will be trusted to correct the error. Others agents provide explanations **after** actions are carried out [23], to be used for retrospective reports or post-hoc analysis [24].

The choice of when to present the explanation is not exclusive. Agents might supply various explanations at various times, before, during and after the task is carried out. Building on the taxonomy in the previous section, if interpretability is critical to the system, then it stands to reason that this knowledge must be presented at the **beginning** of the task, thus enabling the user to determine whether to accept the agents recommendation [32]. However, if it is beneficial to build trust / user acceptance, then it might be directed **during** the task, especially if the agent erred. If the purpose of the explanation is to justify the agent's choice from a legal perspective then we may need to certify that decision before the agent acts (preventative) or after the act (accusatory). But, if the goal is conceptualization, especially in the form of knowledge discovery and / or to support future decisions, then the need for explanation **after** task execution is equally critical.

5 How Much Detail is Presented in an Explanation

The level of explanation detail should depend on why that human-agent systems needs to be interpretable and how the explanation has been generated. If the need for explanation is for legal purposes, then it follows that legal experts need the explanation and not the typically user. Similarly, it reasons that the type of explanation that is given should be directed specifically to this population. If the purpose the explanation is to support expert knowledge discovery, then it reasons that the explanation should be directed towards researchers with knowledge of a specific problem. In these cases, the systems might not even need to present their explanations to the *regular users* and may thus only focus on presenting information to these experts. Most systems will still likely benefit by directing explanation to the *regular users* to help them better understand the system's decisions, thus aiding in their acceptance and / or trust. In these cases, the system should be focused on providing arguments and / or the logic behind their decisions [27] or Case Based Reasoning (e.g. I choose to do this before of that) [22; 4; 19] that help reassure the user about the correctness of the agent's decision.

Similarly, the level of detail of what constitutes an adequate explanation likely depends on the project amount of time the user will study the explanation. If the goal is knowledge discovery and / or legal, then an expert will likely need to spend large amounts of time meticulously studying the inner-workings of the decision making process. In these cases, it seems likely that large amounts of detail regarding the justification of the explanations will be necessary. If a *regular user* is assumed, and the goal is to build user trust and understanding, then shorter, very directed explanations are likely more beneficial. This issues touches upon a larger issues about the potential information overload additional information may pose for a given user [34].

Additionally, the type of interface used for disseminated the system's explanation will likely depend upon the level of the user's expertise. The idea of adaptable interfaces based on people's expertise was previously noted [12; 33; 38]. In these systems, the type of information presented in the interface depends on the user's level of expertise. Accordingly, an interface might consider different types of explanations or explanation algorithms based on who the end-user will be. Consider explanations that can potentially be generated within an image classification task. Most people will be able to understand the comparison of similar pictures or even basic text explanations that have been previously proposed [17]. However, interfaces with saliency maps [36] may be too complex for many users and thus reserved for experts, especially those researching the under-workings of the algorithm. Even among experts, it is reasonable to assume that different users will need different types of information. The different backgrounds of legal experts, scientists, safety engineers, and researchers may necessitate different types of interfaces [8].

6 A Utility for Evaluation of an Explanation Based System

We previously considered issues of **Why**, **What** ,**When** and **How** regarding explanations. These issues are often intrinsically connected. For example, the detail of an explanation is often dependent on **why** that explanation is needed. An expert will likely differ from a *regular user* regarding **why** an explanation is needed, will often need these explanations at different times, e.g. before or after the task (**when**), and may require different types of explanations and interfaces (**what** and **how**). At other times multiple facets of explanation exist even within one category. A DSS system is built to support a user's decision, thus making interpretability a critical issue. However, these systems will still likely benefit from better explanations, so that the user trusts those explanations. Similarly, a scientist pursuing knowledge discovery may need to analyze and interact with information presented before, during and after a task's completion (**when**). Thus, multiple goals must often be considered and evaluated.

We consider three elements that should be considered in evaluating Human-Agent Systems:

1. A score for the performance of the agent's algorithm
2. A score for the user's performance
3. A score for the explanation given the to user

For example, in a movie recommendation system the three scores would be described as follows: The score of the algorithmic component is an evaluation of the algorithm that is used for recommendation (i.e. accuracy, precision and / or recall). The user's performance score evaluates whether the

user chooses more movies when using the system. The score of the explanation refers to user's satisfaction.

A complex interplay exists between these three elements. Both the user's performance and the user's satisfaction can be affected by the explanation. On the other hand they are also both affected by the algorithm. As we discussed in Section 3, there is often a trade-off between the performance of the agent and the ability to provide a high quality explanation. The information generated from simpler machine learning algorithms might be more explainable, but the accuracy of these algorithms might be lower than those of less explainable machine learning algorithms (see Figure 1). Furthermore, different user types and interfaces will be effected by the type of agent design and a total measure is needed to weigh all parameters that are needed by a system into account. For example, an agent that was designed to support an *expert user* is different to that provided to a *regular user*.

We must consider situations when the goals for system are both complementary and in conflict. Earlier, we discussed several instances where the goals are complementary. However, this is not always the case. For example, assuming the explanation goal of a system is to support a person's ability to purchase items in a time-constrained environment (e.g. online stock purchasing). The greater detail contained within the agent's explanations on one hand instill improved confidence within the user, but also will take more time to read and process, which may prevent the user from capitalizing on certain quickly passing market fluctuations. Thus, some measure should likely be introduced to reason about different goals for the explanation and the relative strengths of various explanations, their interfaces, and the algorithms that generate those explanations.

Another equally important element of the system is how well the person performed in the system with the total of all explanations provided. In theory, different such goals may exist for the human user such as immediate performance vs. long-term knowledge acquisition. Furthermore, the performance of the agent's algorithm is an important part of the system and must be included in the system utility, regardless of which specific user or user constraints exist. Clearly, we need to maximize the overall utility of the system across all elements of the system's performance.

To capture these properties, we propose an overall utility to the system which is the following weighted sum:

$$Utility = \sum_{n=1}^{NumGoals} Imp_n * Grade_n \quad (1)$$

We define $NumGoals$ as the number of goals in an the system. Examples of goals could be "explanation during agent activity", "explanation after completion of task or "clarity of visualization for explanation" or "clarity of visualization for explanation". Goals can also refer to the system performance, such as "user performance". A goal can also be related to the algorithmic component, such as "model accuracy". Imp_n is the importance weight we give to the n_{th} goal. Similarly, $Grade_n$ is the score we give to the n_{th} goal. We require that $0 <= Grade_n <= 1$ such that

$$\sum_{n=1}^{NumGoals} Imp_n = 1 \quad (2)$$

While this model helps quantify the interplay of multiple explanation goals, either inherently complimentary or contradictory, a fundamental question lingers about how to set the values of $NumGoals$, Imp_n and $Grade_n$. We assume that either users themselves the system's designer, or outside 3rd party organizations (e.g. governments) can quantify both the goals of the system and their relative importance. The grade for the user's performance is assumed to be quantifiable. Assuming the goal is to support the user's ability to classify an item, then accepted measures such as accuracy, precision, recall, F-measure, and Mean Absolute Errors can be used too. However, it reasons that the score of a given goal may also be boolean (e.g. either the system provided an acceptable legal explanation or it didn't) while others are highly subjective and may vary as the task is performed (e.g. how much trust did the agent generate). If providing a certain explanation is a critically important goal, then this value should be made much larger than any of the other goals to ensure its primacy.

According to this model, some measure of the user's performance with the system is needed to quantify $Grade_n$ for this goal. Assuming the user experience needs to be measured, as is typically done in HCI studies and is likely to be a significant factor in such human-agent systems, then accepted user performance metrics such as the NASA-TLX [16] or the System Usability Scale [3] should be used to measure the utility of the user's satisfaction. In all cases, we concede that for many real-world systems quantifying these elements, especially in the case of the relative values for each $Grade_n$ is an important challenge that needs to be further explored in the future.

In order to help evaluate these systems, some researchers have suggested simplifying the type of domains, experiment setups and evaluation criteria. One possible approach is to objectively quantify the grade for the explanation portion of an agent's model based on its size. The assumption behind this approach is that as the size of machine learning models grow, they are less interpretable. Thus, models with large numbers of nodes / hidden layers (e.g. in deep neural networks), parameter values (for regression and SVM models), the number of rules (rule-based models), or the depth (in decision trees) are less preferable to those with fewer numbers of these values [7]. Following Section 4, we could also create an objective metric based on the number of features generated from feature selection that are used to create the model. The advantage to both of these approaches is the the value of explanation's $Grade_n$ can be made independently to the system's specific task. A second approach is to quantify the relatively value of different approaches and only give a non-zero score to the one that they feel is best [8]. Thus, the scoring function $Grade_n$ could be made boolean, greatly simplifying calculating the system's total value. A third approach is to create simplified accepted tasks, potentially where simulations of human behavior could be used for repeatedly for evaluation across different algorithms, interfaces, and approaches [8]. This could create a standardization for all values of

$NumGoals$, Imp_n and $Grade_n$, again greatly aiding in the evaluation process. Currently, no such canonical tasks have been universally accepted, leaving this issue as an open challenge.

7 Conclusion

We presented a framework designed to enable comparison and evaluation of interpretability in Human-Agent Systems. As Human-Agent Systems are diverse and complex, there is no "one explanation type fits all". Each agent must have its requirements and goals mapped out, and the appropriate explanation chosen. We focused on agents that use machine learning and provided an attempt to define this new field of interpretability and explainability.

Our contribution is a proposed framework that determines the answers to four questions **Why**, **What**, **When** and **How**. These questions define the various aspects of the explanation for the system. In designing an agent one must first establish **why** the system requires explanation, as this will affect the answer to the other questions. Next, one must determine **what** type of explanation is needed, followed by considering **when** to present it. Finally, the question of **how** detailed the explanation should be must be addressed. For each of the first three questions we presented a set of three possible approaches, and discussed when each approach might be appropriate. Various factors affect the answers to these questions. We discussed how the degree of control of the user over the agent affects the need for interpretability. We also discussed how the type of learning that agents perform will affect the explanation that is provided. We then discussed parameters for when to present the information. For the fourth question of **how** much detail to present, we consider a continuous scale, and discuss the different types of detail needed in different systems, and how the answers to the first three questions affect this decision.

Once an explanation has been defined, there is a need to evaluate it. To this end, we presented an evaluation measure. The measure allows for comparing systems while taking into account both the algorithmic performance and the presentation of explanations in order to achieve the highest interpretability and performance. Our proposed utility is capable of combining all the aspects of the system: the machine learning algorithm, user performance and the explanation, into a single measure. We discussed the strengths and limitations of our proposed measure. While the measure provides a means for comparison, its main limitation relates to the elements that can potentially be subjective in determining the values of the parameters. As this remains an open issue, we hope that this unified framework will not only be adopted by researchers when defining and evaluating the important field of interpretability in Human-Agent Systems, but will provide a basis for possible extensions.

References

[1] Frederic Adam, Frederic Adam, and Patrick Humphreys. *Encyclopedia of Decision Making and Decision Support Technologies*. Information Science Reference - Imprint of: IGI Publishing, Hershey, PA, 2008.

[2] Amos Azaria, Ariella Richardson, and Sarit Kraus. An agent for the prospect presentation problem. In *Proceedings of the 2014 international conference on Autonomous agents and multi-agent systems*, pages 989–996. International Foundation for Autonomous Agents and Multiagent Systems, 2014.

[3] John Brooke et al. Sus-a quick and dirty usability scale. *Usability evaluation in industry*, 189(194):4–7, 1996.

[4] Juan M Corchado and Rosalía Laza. Constructing deliberative agents with case-based reasoning technology. *International Journal of Intelligent Systems*, 18(12):1227–1241, 2003.

[5] Michael T Cox and Anita Raja. *Metareasoning: Thinking about thinking*. MIT Press, 2011.

[6] David Crockett and Brian Eliason. What is data mining in healthcare?, 2016.

[7] Hoa Khanh Dam, Truyen Tran, and Aditya Ghose. Explainable software analytics. *CoRR*, abs/1802.00603, 2018.

[8] Finale Doshi-Velez and Been Kim. Towards a rigorous science of interpretable machine learning. 2017.

[9] Maria Fox, Derek Long, and Daniele Magazzeni. Explainable planning. *CoRR*, abs/1709.10256, 2017.

[10] Alex A. Freitas. Comprehensible classification models: A position paper. *SIGKDD Explor. Newsl.*, 15(1):1–10, March 2014.

[11] Michael A Goodrich, Dan R Olsen, Jacob W Crandall, and Thomas J Palmer. Experiments in adjustable autonomy. In *Proceedings of IJCAI Workshop on Autonomy, Delegation and Control: Interacting with Intelligent Agents*, pages 1624–1629. Seattle, WA: American Association for Artificial Intelligence Press, 2001.

[12] Jonathan Grudin. The case against user interface consistency. *Communications of the ACM*, 32(10):1164–1173, 1989.

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

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

[15] Satoshi Hara and Kohei Hayashi. Making tree ensembles interpretable. *arXiv preprint arXiv:1606.05390*, 2016.

[16] Sandra G Hart and Lowell E Staveland. Development of nasa-tlx (task load index): Results of empirical and theoretical research. In *Advances in psychology*, volume 52, pages 139–183. Elsevier, 1988.

[17] Lisa Anne Hendricks, Zeynep Akata, Marcus Rohrbach, Jeff Donahue, Bernt Schiele, and Trevor Darrell. Generating visual explanations. In *Computer Vision - ECCV 2016 - 14th European Conference, Amsterdam, The Netherlands, October 11-14, 2016, Proceedings, Part IV*, pages 3–19, 2016.

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

[19] Been Kim, Cynthia Rudin, and Julie A Shah. The bayesian case model: A generative approach for case-based reasoning and prototype classification. In *Advances in Neural Information Processing Systems*, pages 1952–1960, 2014.

[20] Bart P Knijnenburg, Martijn C Willemsen, Zeno Gantner, Hakan Soncu, and Chris Newell. Explaining the user experience of recommender systems. *User Modeling and User-Adapted Interaction*, 22(4-5):441–504, 2012.

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

[22] Oh Byung Kwon and Norman Sadeh. Applying case-based reasoning and multi-agent intelligent system to context-aware comparative shopping. *Decision Support Systems*, 37(2):199–213, 2004.

[23] Pat Langley, Ben Meadows, Mohan Sridharan, and Dongkyu Choi. Explainable agency for intelligent autonomous systems. In *AAAI*, pages 4762–4764, 2017.

[24] Zachary Chase Lipton. The mythos of model interpretability. *CoRR*, abs/1606.03490, 2016.

[25] Marco Tulio Ribeiro, Sameer Singh, and Carlos Guestrin. 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*, pages 1135–1144. ACM, 2016.

[26] Ariella Richardson, Sarit Kraus, Patrice L Weiss, and Sara Rosenblum. Coach-cumulative online algorithm for classification of handwriting deficiencies. In *AAAI*, pages 1725–1730, 2008.

[27] Ariel Rosenfeld and Sarit Kraus. Strategical argumentative agent for human persuasion. In *ECAI*, volume 16, pages 320–329, 2016.

[28] Avi Rosenfeld, Vinay Sehgal, David G. Graham, Matthew R. Banks, Rehan J. Haidry, and Laurence B. Lovat. Using data mining to help detect dysplasia: Extended abstract. In *2014 IEEE International Conference on Software Science, Technology and Engineering, SWSTE 2014, Ramat Gan, Israel, June 11-12, 2014*, pages 65–66, 2014.

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

[30] Maha Salem, Gabriella Lakatos, Farshid Amirabdollahian, and Kerstin Dautenhahn. 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*, pages 141–148. ACM, 2015.

[31] Paul Scerri, David Pynadath, and Milind Tambe. Adjustable autonomy in real-world multi-agent environments. In *Proceedings of the fifth international conference on Autonomous agents*, pages 300–307. ACM, 2001.

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

[33] Ben Shneiderman. Promoting universal usability with multi-layer interface design. *ACM SIGCAPH Computers and the Physically Handicapped*, (73-74):1–8, 2002.

[34] Tammar Shrot, Avi Rosenfeld, Jennifer Golbeck, and Sarit Kraus. Crisp: an interruption management algorithm based on collaborative filtering. In *Proceedings of the SIGCHI conference on human factors in computing systems*, pages 3035–3044. ACM, 2014.

[35] Maarten Sierhuis, Jeffrey M Bradshaw, Alessandro Acquisti, Ron Van Hoof, Renia Jeffers, and Andrzej Uszok. 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)*, 2003.

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

[37] Frode Sørmo, Jörg Cassens, and Agnar Aamodt. Explanation in case-based reasoning—perspectives and goals. *Artif. Intell. Rev.*, 24(2):109–143, October 2005.

[38] Sebastian Stein, Enrico H. Gerding, Adrian Nedea, Avi Rosenfeld, and Nicholas R. Jennings. Market interfaces for electric vehicle charging. *J. Artif. Intell. Res.*, 59:175–227, 2017.

[39] David Traum, Jeff Rickel, Jonathan Gratch, and Stacy Marsella. 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*, pages 441–448. ACM, 2003.

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

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

[42] Holly A Yanco and Jill Drury. Classifying human-robot interaction: an updated taxonomy. In *Systems, Man and Cybernetics, 2004 IEEE International Conference on*, volume 3, pages 2841–2846. IEEE, 2004.