Why, Who, What, When and How about Explainability in Human-Agent Systems

JAAMAS Track

Avi Rosenfeld
Jerusalem College of Technology
Jerusalem
rosenfa@jct.ac.il

Ariella Richardson
Jerusalem College of Technology
Jerusalem
richards@jct.ac.il

ABSTRACT

This paper presents a survey of issues relating to explainability in Human-Agent Systems. We consider fundamental questions about the *Why, Who, What, When and How* of explainability. First, we define explainability and its relationship to the related terms of interpretability, transparency, explicitness, and faithfulness. These definitions allow us to answer *why* explainability is needed in the system, *whom* it is geared to and *what* explanations can be generated to meet this need. We then consider *when* the user should be presented with this information. Last, we consider *how* objective and subjective measures can be used to evaluate the entire system. This last question is the most encompassing as it needs to evaluate all other issues regarding explainability.

KEYWORDS

Human–agent systems ; XAI ; Machine learning interpretability ; Machine learning transparency

ACM Reference Format:
Avi Rosenfeld and Ariella Richardson. 2020. Why, Who, What, When and How about Explainability in Human-Agent Systems. In *Proc. of the 19th International Conference on Autonomous Agents and Multiagent Systems (AAMAS 2020), B. An, N. Yorke-Smith, A. El Fallah Seghrouchni, G. Sukthankar (eds.), Auckland, New Zealand, May 2020,* IFAAMAS, 4 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, chat bots, planning systems and self-driving cars [2–4, 7, 10–12, 14–18, 20–23, 25, 26, 28]. 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.

We focus on one aspect of this human-agent interaction — the internal level of explainability that agents using machine learning must have regarding the decisions they make. Our overall goal is to provide an extensive study of this issue in Human-Agent Systems. Towards this goal, our first step is to formally and clearly define explainability, as well as the concepts of interpretability, transparency, explicitness, and faithfulness that make a system explainable. Through using these definitions, we provide a clear taxonomy regarding the *Why, Who, What, When, and How* about explainability and stress the relationship of interpretability, transparency, explicitness, and faithfulness to each of these issues.

This paper's first contribution is a clear definition for explainability and for the related terms: interpretability and transparency. In defining these terms we also define how explicitness and faithfulness are used within the context of Human-Agent Systems. A summary of these definitions is found in Table 1. In defining these terms, we focus on the features and records that are used as training input in the system, the supervised targets that need to be identified, and the machine learning algorithm used by the agent. We define $\mathcal{L}$ as the machine learning algorithm that is created from a set of training records, R. Each record $r \in R$ contains values for a tuple of ordered features, F. Each feature is defined as $f \in F$. Thus, the entire training set consists of $R \times F$. While this model naturally lends itself to tabular data, it can as easily be applied to other forms of input such as texts, whereby f are strings, or images whereby f are pixels. The objective of $\mathcal{L}$ is to properly fit $R \times F$ with regard to the labeled targets $t \in T$.

To help visualize the relationship between explainability, interpretability and transparency, please note Figure 1. Note that interpretability includes six methods, including transparent models, and also the non-transparent possibilities of model and outcome tools, feature analysis, visualization methods, and prototype analysis. Feature analysis can serve as a basis for creating transparent models, on its own as a method of interpretability, or as a interpretable component within model, outcome and visualization tools. Similarly, visualization tools can help explain the entire model as a global solution or as a localized interpretable element for specific outcomes of $t \in T$. Prototype analysis uses R as the basis for interpretability, and not F, and can be used for visualization and/or outcome analysis of $r \in R$. Interpretability is a means for providing explainability, as per these terms' definitions in Table 1.

To date, many reasons have been suggested for making systems explainable [1, 5, 6, 8, 9, 13, 24]: to justify its decisions so the human participant can decide to accept them (provide control), to explain the agent's choices and guarantee safety concerns are met, to build trust in the agent's choices, especially if a mistake is suspected or the human operator does not have experience with the system, to explain the agent's choices that ensure fair, ethical, and/or legal decisions are made, to explain the agent's choices and better evaluate or debug the system in previously unconsidered situations,

Proc. of the 19th International Conference on Autonomous Agents and Multiagent Systems (AAMAS 2020), B. An, N. Yorke-Smith, A. El Fallah Seghrouchni, G. Sukthankar (eds.), May 2020, Auckland, New Zealand.

Term	Notation	Short Description
Feature	F	One field within the input.
Record	R	A collection of one item of information (e.g. picture, row in datasheet).
Target	T	The labeled category to be learned. Can be categorical or numeric.
Algorithm	$\mathcal{L}$	The algorithm used to predict the value of T from the collection of data (all features and records).
Interpretation	$\mathbb{I}$	A function that takes as its input F, R, T, and $\mathcal{L}$ and returns a representation of $\mathcal{L}$'s logic.
Explanation	$\mathbb{E}$	The human-centric objective for the user to understand $\mathcal{L}$ using $\mathbb{I}$.
Explicitness		The extent to which $\mathbb{I}$ is understandable to the intended user.
Fairness		The lack of bias in $\mathcal{L}$ for a field of importance (e.g. gender, age, ethnicity).
Faithfulness		The extent to which the logic within $\mathbb{I}$ is similar to that of $\mathcal{L}$.
Transparency		The connection between $\mathbb{I}$ and $\mathcal{L}$ is both explicit and faithful.

Table 1: Notation and short definition of explainability, interpretability, transparency, fairness, and explicitness.

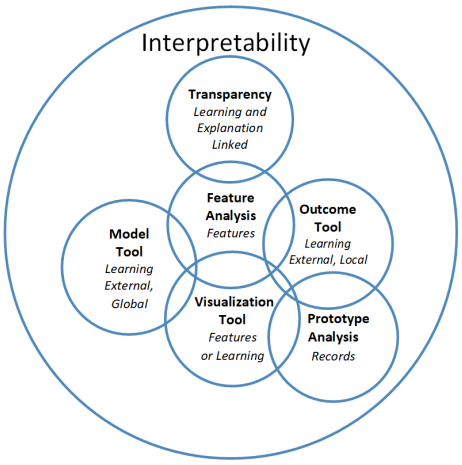

Figure 1: A Venn Diagram of the relationship between Explainability, Interpretability and Transparency. Notice the centrality of Feature Analysis to 4 of the 5 elements.

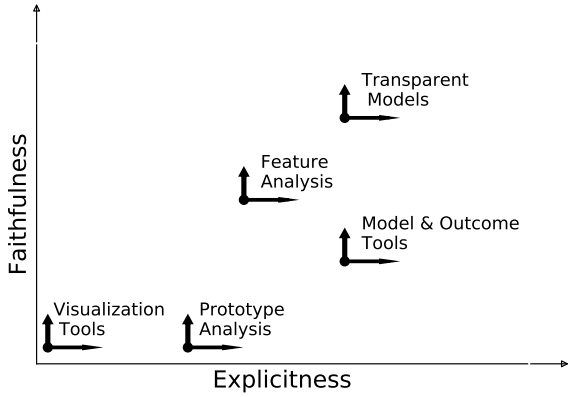

Figure 2: Faithfulness versus explicitness within the six basic approaches for generating interpretations

and knowledge / scientific discovery. Once we have established the *why* and *who* about explanations, a key related question one must address is *what* interpretation can be generated as the basis for the required explanation. Different users will need different types of explanations, and the interpretations required for effective explanations will differ accordingly [27]. We posit that six basic approaches exist as to how interpretations can be generated: directly from a transparent machine learning algorithm, feature selection and/or analysis of the inputs, using an algorithm to create a post-hoc model tool, using an algorithm to create a post-hoc outcome tool, using an interpretation algorithm to create a post-hoc visualization of the agent's logic and using an interpretation algorithm to provide post-hoc support for the agent's logic via prototypes.

In Figure 2 we describe how these various methods for generating interpretations have different degrees of faithfulness and explicitness. Each of these methods contains some level of trade-off between their explicitness and faithfulness. Transparent models are inherently more explicit and faithful than other possibilities. Nonetheless, we present this figure only as a guideline, as many implementations and possibilities exists within each of these six basic approaches. These differences will impact the levels of both faithfulness and explicitness, something we indicate via the arrows pointing to both higher levels of faithfulness and explicitness for a specific implementation.

Creating a general evaluation framework is still an open challenge as 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 explainability 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 hope that the definitions presented here and in the extended version of this paper [19] will serve as a basis for future studies about the five questions about explainability that we present, particularly in the proper evaluation of explainability in Human-Agent Systems.

REFERENCES

[1] Ashraf Abdul, Jo Vermeulen, Danding Wang, Brian Y. Lim, and Mohan Kankanhalli. 2018. Trends and Trajectories for Explainable, Accountable and Intelligible Systems: An HCI Research Agenda. In *Proceedings of the 2018 CHI Conference on Human Factors in Computing Systems*. Article 582, 18 pages.

[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] 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.

[4] 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.

[5] 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*.

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

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

[8] Shirley Gregor and Izak Benbasat. 1999. Explanations from intelligent systems: Theoretical foundations and implications for practice. *MIS quarterly* (1999), 497–530.

[9] 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, Article 93 (Aug. 2018), 42 pages.

[10] 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.

[11] 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.

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

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

[14] Ariella Richardson, Sarit Kraus, Patrice L Weiss, and Sara Rosenblum. 2008. COACH-Cumulative Online Algorithm for Classification of Handwriting Deficiencies.. In *AAAI*. 1725–1730.

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

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

[17] 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*.

[18] 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.

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

[20] 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.

[21] 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.

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

[23] 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)*.

[24] Frode Sørmo, Jörg Cassens, and Agnar Aamodt. 2005. Explanation in case-based reasoning–perspectives and goals. *Artificial Intelligence Review* 24, 2 (2005), 109–143.

[25] 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.

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

[27] Luca Viganò and Daniele Magazzeni. 2018. Explainable Security. *arXiv preprint arXiv:1807.04178* (2018).

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