Artificial Intelligence for Human-Robot Interaction
AAAI Technical Report FS-17-01

Different XAI for Different HRI

Raymond K. Sheh
Intelligent Robots Group, Department of Computing, Curtin University
Building 314, Kent St, Bentley WA 6102, Australia

Abstract

Artificial Intelligence (AI) has become more widespread in critical decision making at all levels of robotics, along with demands that the agent also explain to us humans why they do what they do. This has driven renewed interest in Explainable Artificial Intelligence (XAI). Much work exists on the Human-Robot Interaction (HRI) challenges of creating and presenting explanations to different human users in different applications but matching these up with AI and Machine Learning (ML) techniques that can provide the underlying explanatory information can still be a challenge. In this short paper, we present a categorisation of explanations that communicate the XAI requirements of various users and applications, and the XAI capabilities of various underlying AI and ML techniques.

Introduction

A critical part of Human-Robot Interaction (HRI) is the conveying, to a human, of the reasons behind a robot's decisions. As robotic Artificial Intelligence (AI) agents become more advanced, these reasons become buried in increasingly complex, machine learned (ML) models. Explainable Artificial Intelligence (XAI) is a response to this trend. While the concept of XAI itself is not new, it is increasingly topical (Defense Advanced Research Projects Agency (DARPA) 2016; Ribeiro, Singh, and Guestrin 2016). The European Union's General Data Protection Regulation (GDPR) includes a "Right to an Explanation", showing that governments are also taking an interest in this space (Goodman and Flaxman 2016).

XAI seeks to produce AI agents that can not only make decisions but can also explain them. Robotic XAI agents include the HRI components that generate and present explanations to the user, as well as the underlying AI that, as it makes its decisions, also furnishes the requisite information to support the generation of these explanations.

The topic of how to generate and present explanations to users continues to be the topic of a growing body of work from the HRI and cognitive science perspectives, covering topics including the nature of explanation itself, what level of information should be presented, how it can be structured, issues of persuasion, dialog and the like (Miller, Howe, and Sonenberg 2017; Keil 2003; Lim and Dey 2009; Parsons and McBurney 2003). This existing work tends to assume that the underlying AI systems are white-box, as is the case for systems such as classical planners. This assumption often breaks down for AI agents with a substantial ML component. When building XAI systems where significant decisions are made by ML techniques, it is important to understand the types of explanations that different ML techniques can support. This existing work also tends to focus on users of the system in normal operation. To our knowledge there has been no unified work on the explanatory needs of other users that covers not just ordinary users of the agent, but also those who wish to learn what the agent has discovered and those who wish to troubleshoot or investigate failures in the agent, beyond simply assuming that this information is readily available.

In this short paper, we present a categorisation of explanations from the ML perspective. It extends our previous work (Sheh 2017) and makes concrete the more abstract categorisations that we previously presented. The goal of these categories is to allow developers of XAI agents to trade off the need for explainability against other factors such as efficiency, predictive accuracy and representational flexibility of the underlying ML systems. They also outline the needs of the aforementioned users. These categories complement existing HRI and cognitive science work on what explanations are necessary and how they should be presented. If the agent design is driven by the HRI demands, our categorisation will help to determine the ML techniques that can provide the necessary underlying information. If the agent design is driven by the ML techniques, our categorisation will help to determine which of the aforementioned explanations are possible and the limitations that may be faced in implementing them.

Different Explanations

There is much existing work on defining and categorising explanations from a cognitive science perspective. For example, work such as (Keil 2003) and (Lim and Dey 2009) concern how to focus explanations onto what is most informative or desirable for a given audience and how this changes with application. These categories are useful for producing explanation likely to be appreciated and understood by a given audience in a given application. However,

they focus on the user experience and do not break cleanly on the differences in explanatory information furnished by different ML techniques. Indeed, they often assume a white-box decision making process. Our intermediate categorisations complement these by making the match between XAI-HRI need and ML capability explicit.

In the context of this work, we focus on semantics that are useful in determining what underlying explanatory information can be extracted from an ML system to present to the user. The explanations that this work fits within is perhaps best categorised as being of the type of "Information-Seeking Dialogs" according to (Walton and Krabbe 1995), whereby a human user has questions and believes that the agent has the answers. We consider the presentation to the user – be it through graphical user interfaces, natural language or the like, or the way in which that dialog might be undertaken, such as through negotiation, persuasion or the like (Parsons and McBurney 2003), to be relevant but outside the scope of this specific paper.

Of course it is worth noting that the use of an ML technique that is conducive to explanation is not, by itself, a guarantee of a good explanation. If the underlying phenomena being predicted can't be meaningfully expressed by the ML technique then any explanation will be forced and unuseful, even if the predictive accuracy is good. For example, a decision tree can learn a concept in 2D space with a decision boundary defined by $y = x$ but the splits that define the stairstepped approximation to this function are clearly not a meaningful way of expressing this simple concept.

Our categorisation covers applications beyond just that which focus on a normal user of the system and extend this to the needs of AI experts and developers who construct and debug the system, regulators who need to determine those legally responsible when something goes wrong, and others who may wish to learn about what the agent has discovered in the course if its own learning. We propose that for the purpose of HRI, it is useful to define two dimensions when discussing explanations – **type** and **depth**.

Types of Explanations

We propose the following five types of explanations: *Teaching, Introspective Tracing, Introspective Informative, Post-Hoc* and *Execution*. We have ordered them in terms of how they limit the ML and AI techniques that may be used. At one end, *Teaching* explanations are only possible with a limited number of ML and AI techniques. At the other end, *Execution* explanations can be provided for practically any program. Thus ML techniques capable of supporting explanations of a given type are generally also able to support explanations of subsequent types.

Teaching explanations tend to be ignored by most work on XAI-HRI and yet, with its roots in expert systems, are perhaps the earliest of the explanation types to be studied by the AI community. These explanations convey to the human concepts and knowledge (for some level of granularity) that the agent has gathered. The human may be a domain expert, AI expert, user or other member of the general public. Examples of ML techniques that provide this level of information include decision trees, ripple-down rules and inductive logic programming. These explanations are not necessarily anchored to a given decision. Examples of ways in which teaching explanations may be presented to humans include agent-generated hypotheticals, logic rules, dialog, visualisations of state space or the like.

Teaching explanations may follow on from other explanations discussed below. For example, after asking the agent how it arrived at a given decision, a user may start to ask the agent questions such as – "How much does parameter A need to change by to yield a higher probability of this other decision?", "Show me examples of situations where you would have made a similar decision" or "If this sensor were missing, what decision would you have made?".

Introspective Tracing explanations are perhaps the most commonly thought of explanations in XAI. They are so named because they are based on introspection into the underlying models and provide enough information to trace the decision making process, much like how a classical planner is able to. These explanations contain enough information, at a suitable level of abstraction, that a suitably dedicated human could understand and reconstruct the decision at the desired level of granularity. The aforementioned work in XAI-HRI tends to focus on agents that have this level of knowledge of their decisions.

Introspective Tracing explanations are useful for (human) entities that range from the lay public through to domain and AI experts. Crucially, they provide the requisite information to help investigators assign responsibility when something goes wrong. They may also provide developers with the knowledge to produce solutions to errors that can be verified explicitly (or at least where the solution's effect on the agent's behaviour can be characterised explicitly). This is in contrast to solutions where verification of a change can only be characterised statistically. We have included "Tracing" in the name of this category to highlight this fact, that these explanations can be used to trace the complete (at a given level of granularity) decision making process.

Few ML techniques are able to provide data supporting human-interpretable Introspective Tracing explanations. More can provide data that support explanations that are still faithful to the underlying models and decision making process but are incomplete in some way.

Introspective Informative explanations provide less information than Tracing explanations but at least enough so that when there is a discrepancy between the agent's decisions and those expected by a user (who may be a domain expert but not necessarily an AI expert), the agent can reasonably convince the user that the agent is correct, or that the agent has made an error and where the error is likely to have occurred. These explanations are derived from the AI agent's underlying models and decision making process but may be limited in some way. For example, hybrid models that combine a decision tree with function approximation at the leaves may be able to trace the decision process as far as a given leaf but may then only be able to provide information about what attributes were considered beyond that point.

Post-Hoc explanations, or rationalisations, are explanations that are constructed by the agent that explain a decision without accurately representing the underlying decision

making processes. Typically, these are used for agents where the underlying learned model is a black-box, such as a neural network. To the user, these explanations may be indistinguishable from the Introspective explanations. Thus, Post-Hoc explanations must be used and presented carefully as they have the potential to mislead. The topic of Post-Hoc explanations has gained considerable attention in recent times as the wave of black-box and near-black-box learning techniques associated with Deep Learning meets the increasing demands for explainability (Ribeiro, Singh, and Guestrin 2016; Ross, Hughes, and Doshi-Velez 2017).

These explanations may be produced around a particular decision, a process that might be thought of as akin to linearising a complex function around a given point. It will diverge from the black-box model's decisions as the situation changes, perhaps in ways that are difficult to characterise or understand. Alternatively, they may be produced by a parallel, explainable model that attempts to replicate the underlying black-box model, to some level of coverage and fidelity. There are efforts to train deep learning networks to produce explanatory information alongside decisions. These may result in hybrid explanations that still do not reflect, with fidelity, the black-box model but nevertheless still have some basis in the learned concepts.

Such explanations are akin to explanations humans might provide. Humans are notoriously bad at generating accurate explanations for their actions, especially where skill or experience is involved. Indeed, the field of behavioural cloning (Bratko, Urbančič, and Sammut 1998; Isaac and Sammut 2003; Kadous, Sammut, and Sheh 2006), a variant of Learning from Demonstration (Atkeson and Schaal 2016), grew out of the expert systems community to produce Post-Hoc explanations of expert behaviours in an automated manner. Actually executing these models so they can generate decisions in their own right was a lower priority (although it can certainly be a goal (Sheh 2010), making the explanations Introspective Tracing rather than Post-Hoc).

Execution explanations are a listing of the individual operations that the computer undertakes, and a literal interpretation of its static and dynamic data structures. The level of abstraction may vary from individual CPU instructions up to neural network weights and activations. All AI agents (and all programs, for that matter) are theoretically capable of producing these explanations. However, depending on the complexity of the agent, it may be intractable to do so due to the required computation time or storage space. Furthermore, apart from debugging applications or the simplest agents, these explanations are unlikely to be informative in a practical HRI setting.

Depth of Explanations

Along with the different types of explanations, we propose that it is useful to also consider the depth of the explanation. In this context, we define depth from the perspective of the final decision, as the concept of how far back the agent must go through the decision making process to generate the explanation. We define the following three depths: *Attribute Only* (shallowest), *Attribute Use* and *Model* (deepest).

Attribute Only explanations include only information about the attributes that the model considered in making the decision, perhaps with limited information about how these attributes were used such as their priority or relative weighting. In a HRI context, such explanations may still be useful in the first instance to determine if the agent has based its decisions on factors that are reasonable, rather than those that are irrelevant and likely to be due to overfitting. For example, a robot that is traversing rough terrain main explain that its decision to turn back was being driven by the height of the terrain above and the relative flatness of the terrain next to it, with the tilt of the robot being a minor factor as well.

The attributes that most affected a given decision can be generated or derived for many types of models and ML techniques, even if only in a statistical manner. For example, agents may invite the human user to select between a set of plausible Post-Hoc Attribute Only explanations (Ross, Hughes, and Doshi-Velez 2017).

Attribute Use explanations include information that requires more in-depth knowledge of how those attributes are used by the model. For example, in a decision tree, this would take the form of a trace through the tree and incorporate not just the identity of the attributes that were considered but also the various thresholds. Memory and distribution based techniques may instead report measures based on, for instance, decision boundaries. In a HRI application, explanations at this level allow the agent to present to explanations that include the implications of the values of their attributes. For example, a robot that is traversing rough terrain may be able to explain that its decision to turn back was dominated by the fact that the area ahead of it was more than 30 cm above the surrounding area. ML techniques differ in their ability to provide information about how attributes are used in a decision making process. Decision Trees and other techniques that encode information in a form that can be directly converted into relatively compact logic rules tend to be more amenable to this level of explanation although presentation can still be challenging. In contrast, those that couple attributes together in more complex networks, such as Neural Networks, may make it impossible to extract Introspective or Teaching explanations at this level of depth.

Model explanations include information about how the model itself was generated. In the context of HRI, beyond debugging and accountability, explanations at this level help a user to understand the history and experience behind the decision. For example, the rough terrain traversing robot may explain that historically, situations that included the area ahead of it being 30 cm above the surrounding area had a 70% probability of failing if the robot drove forward.

These explanations favour a clear relationship between the model and the data that generated it. For example, in decision tree learning, the training examples can be pushed through the resulting tree (even after pruning) and the distribution of their arrival at the leaf nodes is generally broadly consistent with the leaf node's decision. Systems built on models such as Inductive Logic Programming (Wicaksono, Sammut, and Sheh 2017) can also provide such explanations. As the complexity and coupling of the model increases, as is the case in such models as Neural Networks and the like, it becomes harder to generate meaningful

Model explanations, especially for Introspective and Teaching explanations.

Discussion

We will now discuss how different HRI needs can be met by different XAI approaches using two examples: Shopping Mall Assistants and Self Driving Cars.

Shopping Mall Assistants

HRI roles for a shopping assistant robot might include manipulating and carrying groceries under direction, answering questions about products and making recommendations. In this discussion, we will focus on the role of the recommender. When such an agent makes a recommendation, it is possible that the user will want an explanation for reasons such as the recommendation being unusual or unexpected. These explanations are mostly for the purpose of satisfying the user's curiosity and as a way for the agent to further engage in dialog with the user. Post-Hoc explanations may be quite acceptable at Attribute Only or Attribute Use levels. Most ML techniques may be used. Indeed, less explainable techniques may be more suited to the large quantities of consumer data that these decisions may be based on.

Self Driving Cars

As self-driving cars become more prevalent, eventually an accident will happen that is traced back to a component that includes a learned model, such as a pedestrian detector. Investigators and regulators will want to know that the cause and extent of the error be found and a fix implemented that not only fixes its response to that exact situation but also any similar situation. In this scenario, the HRI interaction is with the investigation and development teams.

In such a setting, it is vital that any explanation be both an accurate and intelligible reflection of the underlying decision processes of the agent. Introspective explanations would be required that allow the investigators to step through the decisions and to determine where the error occurred. This may also be one of few cases where Execution explanations may be needed, just as it is sometimes necessary to use a debugger to step through a program instruction by instruction.

As such, we propose that for such an application, for ML components of the agent that cannot be unit tested to a sufficient level of confidence, it is important to select ML techniques and models that can furnish Introspective explanations that can at least guide human users towards where errors may have occurred.

Conclusion

We have presented a way of categorising explanations with a focus on matching ML capabilities with HRI requirements to produce effective XAI agents. These categories extend from users of the agent under normal operation through to investigators, developers and those who wish to learn concepts that the agent has discovered. These explanations complement existing work that focuses on the HRI and cognitive science aspects of explanations. The ML techniques and models may need to be amended or "patched" in response to these explanations, again in a way that is predictable and explainable. Development of these, and the semantics that allow the nuances of their requirements in different applications, is the subject of ongoing work.

References

Atkeson, C. G., and Schaal, S. 2016. Robot learning from demonstration. In *Proc. Int. Conf. on Machine Learning (ICML)*, volume 97.

Bratko, I.; Urbančič, T.; and Sammut, C. 1998. Behavioural cloning of control skill. *Machine Learning and Data Mining* 335–351.

Defense Advanced Research Projects Agency (DARPA). 2016. Broad Agency Announcement: Explainable Artificial Intelligence (XAI). online.

Goodman, B., and Flaxman, S. 2016. European Union regulations on algorithmic decision-making and a "right to explanation". In *ICML-16 Workshop on Human Interpretability in Machine Learning*.

Isaac, A., and Sammut, C. 2003. Goal-directed learning to fly. In *Proc. Int'l. Conf. on Machine Learning*, 258–265.

Kadous, M. W.; Sammut, C.; and Sheh, R. 2006. Autonomous traversal of rough terrain using behavioural cloning. In *Proc. 3rd Int. Conf. on Autonomous Robots and Agents*.

Keil, F. C. 2003. Folkscience: coarse interpretations of a complex reality. *Trends in Cognitive Sciences* 7(8):368–373.

Lim, B. Y., and Dey, A. K. 2009. Assessing demand for intelligibility in context-aware applications. *Proc. of the 11th Int'l Conf. on Ubiquitous computing*.

Miller, T.; Howe, P.; and Sonenberg, L. 2017. Explainable AI : Beware of Inmates Running the Asylum. In *IJCAI-17 Workshop on Explainable AI (XAI-17)*.

Parsons, S., and McBurney, P. 2003. Argumentation-based communication between agents. In *Communication in Multiagent Systems*, 164–178. Springer.

Ribeiro, M. T.; Singh, S.; and Guestrin, C. 2016. "Why Should I Trust You?": Explaining the Predictions of Any Classifier. *arXiv*.

Ross, A. S.; Hughes, M. C.; and Doshi-Velez, F. 2017. Right for the right reasons: Training differentiable models by constraining their explanations. In *Proc. 26th Int'l Joint Conf. on AI (IJCAI-17)*.

Sheh, R. 2010. *Learning Robot Behaviours by Observing and Envisaging*. Ph.D. Dissertation, School of Computer Science and Engineering, The University of New South Wales, UNSW Sydney.

Sheh, R. 2017. "Why did you do that?" Explainable Intelligent Robots. In *AAAI-17 Workshop on Human-Aware Artificial Intelligence (HAAI-17)*.

Walton, D., and Krabbe, E. C. W. 1995. *Commitment in Dialog*. SUNY Press.

Wicaksono, H.; Sammut, C.; and Sheh, R. 2017. Towards explainable tool creation by a robot. In *IJCAI-17 Workshop on Explainable AI (XAI-17)*.