Explainability in Autonomous Pedagogical Agents

Silvia Tulli [1*]

[1]Department of Computer Science and Engineering
INESC-ID and Instituto Superior Técnico,
Universidade de Lisboa, 2744-016, Porto Salvo, Portugal

Abstract

The research presented herein addresses the topic of explainability in autonomous pedagogical agents. We will be investigating possible ways to explain the decision-making process of such pedagogical agents (which can be embodied as robots) with a focus on the effect of these explanations in concrete learning scenarios for children. The hypothesis is that the agents' explanations about their decision making will support mutual modeling and a better understanding of the learning tasks and how learners perceive them. The objective is to develop a computational model that will allow agents to express internal states and actions and adapt to the human expectations of cooperative behavior accordingly. In addition, we would like to provide a comprehensive taxonomy of both the desiderata and methods in the explainable AI research applied to children's learning scenarios.

Problem Identification

Research in explainable AI is spreading in different communities in pair with the importance of making trustworthy systems that are capable of cooperating with humans. Minimizing the difference between expected agent behavior and actual agent behavior allows for a more efficient human-agent cooperation in these systems. (Rader, Cotter, and Cho 2018).

Learning from Reflections on the Agents' Decision Making Children growing up in the digital era require a renewed type of educational setting that provides them the tools to analyze and learn in a faster and dynamic way. However, the constructivist perspective of learning through disassembling and assembling by adding a reflection process to deepen the knowledge is not applicable for decision-making systems embedded in machines (Wheatley 1991). Also, humans' learning is grounded in social interaction; while interacting, humans enlarge their personal experiences with the experiences of others (Palincsar 1998). Consider a mathematical problem as a learning task that involves logical thinking and mental computation capabilities; the interaction with the agent that explains how and why the problem is solved in a certain way would drive the understanding of the main concepts of the learning task. An agent companion would provide the benefit of social interaction while conveying its decisions towards solving the learning task.

Explainable Agency for Pedagogical Agents The explanation given by a system about the decisions of the agent (in particular pedagogical agents) should be designed considering the specificity of decision making algorithm or tied with how the decision-making algorithm operates. Therefore, to be explainable, the system should provide meaningful explanations that match the mental model of the children about how the system operates and how to solve the problem. The challenge is therefore to identify the class of machine learning models or heuristic algorithms, apply them into learning scenarios, understand their properties, find what about them, if made explainable, will lead to benefits to the children, and validate it later. We would like to investigate which part of that process, made the child understand the importance of looking ahead at the global scene, and concepts of optimal search. We hypothesize that the children would learn that going through all the ways blindly is not as good as using a goal heuristic.

Evaluating the Effects of agents' Explainability The level of explainability of an autonomous system depends on how far humans understand the underlying processes of that system, its decisions and reasons for such decisions; hence, it is only possible to measure the system's explainability considering its effects on other aspects that should be affected by it (e.g., trust, teamwork) (Wortham 2018; Gong and Zhang 2018; Wang, Pynadath, and Hill 2016). When autonomous systems move from being tools to being teammates, or companions in educational settings, an expansion of the interaction model is needed to support the paradigms of teamwork, which require two-way transparency (Chen et al. 2018). Moreover, agents' transparency facilitates the understanding of the responsibilities that different group members might take in collaborative tasks.

*Acknowledges the EU Horizon 2020 research and innovation program for grant agreement No 765955 ANIMATAS project. This work was supported by national funds through Fundao para a Ciłncia e a Tecnologia (FCT) with reference UID/CEC/50021/2019.

We investigated how strategy and transparency of artificial agents can influence human behavior in collaborative game settings. Our results establish that transparency has significant effects on trust, group identification and human likeness. This aspect turns out to be interesting in the context of collaborative learning and the design of relational and social capabilities in intelligent systems (Tulli et al. 2019).

A Collaborative Game Scenario

To validate the hypothesis that the agent's transparency can play a role in fostering logical thinking revealing the agent's decision-making process, we implemented a game scenario in which the child and the agent learn together how to play a zero-sum game that requires logical and mathematical thinking. The game, called Minicomputer Tug of War, is based on Papy's Minicomputer, a non-verbal language to introduce children to mechanical and mental arithmetic through decimal notation with binary positional rules [1]. The agent introduces the game by notifying the child that it is still figuring out how to play. While playing, the agent uses a sub-optimal strategy defined by a minimax search algorithm. The optimal strategy is to iterate over the nodes to find the one with the best terminal state and the sub-optimal strategy is to randomly decide a node with an action that is different from that of the best node. The explanation is generated by comparing the optimal and sub-optimal actions and is meant to give hints to the learner for predicting the agent's behavior and helping the child to make informed decisions. The child-agent play should foster reflections upon the agent's decision-making process and increase the child's understanding of the learning task (Jones, Bull, and Castellano 2018).

Conclusion and Future Work

Autonomous agents have been used to provide automated and personalized teaching and assessment to students (Baraka et al. 2019). We would like to explore the topic of explainability in autonomous pedagogical agents and consider the adaptation of the explanation of the agent to the learner mental model (Conati, Porayska-Pomsta, and Mavrikis 2018). Future work will investigate interactive task learning scenarios in which the agent actually learns the goal or some rules of the task from the child. In the case of the game scenario, the agent should query or asking the child to explain and demonstrate possible actions in the game. The explainable agency could improve the learning experience by revealing to the child, teacher or peer, what is known and what is unclear (Tabrez, Agrawal, and Hayes 2019; Chao, Cakmak, and Thomaz 2010). This work aims to contribute to:

- a methodology for determining the explanations of the planning problem given by the agent in respect to the properties of the heuristic algorithm or machine learning model and the human (observer);
- an evaluation of the effectiveness of the agent's plan explanations in terms of children's learning of logical thinking and mental computation.

[1] Minicomputer Games, http://stern.buffalostate.edu/CSMPProgram/String, consulted on June 2019

References

Baraka, K.; Couto, M.; Melo, F. S.; and Veloso, M. 2019. An optimization approach for structured agent-based provider/receiver tasks. AAMAS '19, 95–103. Richland, SC: International Foundation for Autonomous Agents and Multiagent Systems.

Chao, C.; Cakmak, M.; and Thomaz, A. L. 2010. Transparent active learning for robots. In *2010 5th ACM/IEEE International Conference on Human-Robot Interaction (HRI)*, 317–324.

Chen, J. Y. C.; Lakhmani, S. G.; Stowers, K.; Selkowitz, A. R.; Wright, J. L.; and Barnes, M. 2018. Situation awareness-based agent transparency and human-autonomy teaming effectiveness. *Theoretical Issues in Ergonomics Science* 19(3):259–282.

Conati, C.; Porayska-Pomsta, K.; and Mavrikis, M. 2018. Ai in education needs interpretable machine learning: Lessons from open learner modelling.

Gong, Z., and Zhang, Y. 2018. Behavior explanation as intention signaling in human-robot teaming. 1005–1011.

Jones, A.; Bull, S.; and Castellano, G. 2018. "i know that now, i'm going to learn this next" promoting self-regulated learning with a robotic tutor. *International Journal of Social Robotics* 10(4):439–454.

Palincsar, A. S. 1998. Social constructivist perspective on teaching and learning. *Annual Review of Psychology* 49(1):345–375. PMID: 15012472.

Rader, E.; Cotter, K.; and Cho, J. 2018. Explanations as mechanisms for supporting algorithmic transparency. 103:1–103:13.

Tabrez, A.; Agrawal, S.; and Hayes, B. 2019. Explanation-Based Reward Coaching to Improve Human Performance via Reinforcement Learning. In *2019 14th ACM/IEEE International Conference on Human-Robot Interaction (HRI)*, 249–257.

Tulli, S.; Correia, F.; Mascarenhas, S.; Gomes, S.; Melo, F. S.; and Paiva, A. 2019. Effects of agents' transparency on teamwork. In *Explainable, Transparent Autonomous Agents and Multi-Agent Systems*, 22–37. Cham: Springer International Publishing.

Wang, N.; Pynadath, D. V.; and Hill, S. G. 2016. The impact of pomdp-generated explanations on trust and performance in human-robot teams. In *Proceedings of the 2016 International Conference*, AAMAS '16, 997–1005. Richland, SC: International Foundation for Autonomous Agents and Multiagent Systems.

Wheatley, G. H. 1991. Constructivist perspectives on science and mathematics learning. *Science Education* 75(1):9–21.

Wortham, R. 2018. *Using Other Minds: Transparency as a Fundamental Design Consideration for Artificial Intelligent Systems*. Ph.D. Dissertation.