Learning from Explanations and Demonstrations: A Pilot Study

Silvia Tulli
INESC-ID and IST, Portugal
silvia.tulli@gaips.inesc-id.pt

Sebastian Wallkötter
Uppsala University, Sweden
sebastian.wallkotter@it.uu.se

Ana Paiva
INESC-ID and IST, Portugal
ana.paiva@inesc-id.pt

Francisco S. Melo
INESC-ID and IST, Portugal
fmelo@inesc-id.pt

Mohamed Chetouani
ISIR and SU, France
mohamed.chetouani@sorbonne-universite.fr

Abstract

We discuss the relationship between explainability and knowledge transfer in reinforcement learning. We argue that explainability methods, in particular methods that use counterfactuals, might help increasing sample efficiency. For this, we present a computational approach to optimize the learner's performance using explanations of another agent and discuss our results in light of effective natural language explanations for both agents and humans.

1 Introduction

The process of gaining knowledge from the interaction between individuals needs to allow a two-way flow of information, i.e., reciprocally active communication. During this process explainability is key to enabling a shared communication protocol for effective information transfer. To build explainable systems, a large portion of existing research uses various kinds of natural language technologies, e.g., text-to-speech mechanisms, or string visualizations. However, to the best of our knowledge, few works in the existing literature specifically address how the features of explanations influence the dynamics of agents learning within an interactive scenarios.

Interactive learning scenarios are a much less common but similarly interesting context to study explainability. Explanations can contribute in defining the role of each agent involved in the interaction or guide an agent's exploration to relevant parts of the learning task. Here, some of the known benefits of explanability (e.g., increased trust, causality, transferability, informativeness) can improve the learning experience in interactive scenarios.

Although feedback and demonstration have been largely investigated in reinforcement learning (Silva et al., 2019), the design and evaluation of natural language explanations that foster knowledge transfer in both human-agent and agent-agent scenarios is hardly explored.

Our contribution aims to optimize this knowledge transfer among agents by using explanation-guided exploration. We refer to explanations as the set of information that aims to convey a causality by comparing counterfactuals in the task, i.e, providing the reward that could have been obtained if a different action would have been chosen. Instead of providing the optimal solution for the task, this approach lets the learner infer the best strategy to pursue. In this work, we provide (1) *an overview on the topic of natural language explanations in interactive learning scenarios*, and (2) *a preliminary computational experiment* to evaluate the effect of explanation and demonstration on a learning agent performance in a two-agents setting. We then discuss our results in light of effective natural language explanations for both agents and humans.

2 On Natural Language Explanations in Interactive Learning Scenarios

Humans use the flexibility of natural language to express themselves and provide various forms of feedback, e.g., via counterfactuals. To be successful, artificial agents must therefore be capable of both learning from and using natural language explanations; especially in unstructured environments with human presence. Recent advances in grounded-language feedback state that, although there is a conceptual difference between natural language explanations and tuples that hold information about the environment, natural language is still a favorable candidate for building models that acquire world knowledge (Luketina et al., 2019; Schwartz et al., 2020; Liu and Zhang, 2017; Stiennon et al., 2020). Along this line, training agents to learn

Proceedings of the 2nd Workshop on Interactive Natural Language Technology for Explainable Artificial Intelligence (NL4XAI 2020), pages 61–66,
Dublin, Ireland, 18 December 2020. ©2020 Association for Computational Linguistics

rewards from natural language explanations has been widely explored (Sumers et al., 2020; Najar and Chetouani, 2020; Najar et al., 2020; Krening et al., 2017; Knox et al., 2013; Li et al., 2020; Chuang et al., 2020). The interestigness of Sumers et al. (2020) approach lays in grounding the implementation of two artificial agents on a corpus of naturalistic forms of feedback studied in educational research. The authors presented a general method that uses sentiment analysis and contextualization to translate feedback into quantities that reinforcement learning algorithms can reason with. Similarly, (Ehsan and Riedl, 2020) build a training corpus of state-action pairs annotated with natural language explanations with the intent of rationalizing the agent's action or behavior in a way that closely resemble how a human would most likely do.

Existing literature reviews and experimental studies paired natural language feedback with demonstrations of the corresponding tasks to learn the mapping between instructions and actions (Najar and Chetouani, 2020; Taylor, 2018). This aspect has been studied also in the context of real-time interactive learning scenarios in which the guidance and the dialog with a human tutor is often realized by providing explanations (Thomaz et al., 2005; Li et al., 2020).

Following the idea of *AI rationalization* introduced by (Ehsan and Riedl, 2020), our work approaches the generation of explanations as a problem of translation between ad-hoc representations of an agent's behavior and the shape of the reward function. On the contrary, to achieve our goal we use counterfactuals that can be easily encoded in natural language.

2.1 Explanations for Humans

There exists a substantial corpus of research that investigates explanations in philosophy, psychology, and cognitive science. Miller (Miller, 2019) argues that the way humans explain to each other can inform ways to provide explanation in artificial intelligence. In this context, some authors showed that revealing the inner workings of a system can help humans better understand the system. This is often realized by either generating natural language explanations and visualizing otherwise hidden information (Wallkotter, Tulli, Castellano, Paiva, and Chetouani, 2020). Studies on human learning suggest that explanations serve as a guide to generalization. Lombrozo et al. (Lombrozo and Gwynne, 2014) compared the properties of mechanistic and functional explanations for generalizing from known to novel cases. Their results show that the nature of different kinds of explanations can thus provide key insights into the nature of inductive constraints, and the processes by which prior beliefs guide inference.

Above literature highlights the central role of causality in explanation and the vast majority of everyday explanations invoke notions of cause and effect (Keil, 2006). Therefore, we grounded our explanation formalization in this idea of differentiating properties of competing hypothesis (Hoffmann and Magazzeni, 2019) by comparison of contrastive cases (Madumal et al., 2019).

2.2 Explanations for Agents

Several attempts have been made to develop explanations about the decision of an autonomous agent. Many approaches focus on the interpretation of human queries by either mapping inputs to query or instruction templates (Hayes and Shah, 2017; Lindsay, 2019; Krening et al., 2017), by using an encoder-decoder model to construct a general language-based critique policy (Harrison et al., 2018), or by learning structural causal models for identifying the relationships between variables of interest (Madumal et al., 2019).

However, for a model to be considered explainable, it is necessary to account for the observer of the explanation. In this regard, the research of Lage et al. (2019) investigates the effect of the mismatch between the model used to extract a summary of an agent's policy and the model used from another agent to reconstruct the given summary.

Focusing onto experimental work about knowledge transfer between agents, there exist two main approaches to solve this problem: (1) by reusing knowledge from previously solved tasks, (2) by reusing the experience of another agent. The latter is called inter-agent transfer learning, and is often realized thought human feedback, action advising, and learning from demonstration (Argall et al., 2009; Fournier et al., 2019; Jacq et al., 2019). Some authors refer to policy summarization or shaping when the feedback, advice or demonstration summarize the agent's behavior with the objective of transferring information to another agent (Amir and Amir, 2018). Heuristic based approaches extract diverse important states based on state similarity and q-values, while machine teaching and inverse reinforcement learning approaches extrapolate state-action pairs useful for recovering the agent's reward function (Brown and Niekum, 2018). We take inspiration from policy summarization and learning from demonstration approaches, and extend it by considering explanation-based exploration. Differently from Fournier et al. (2019) and Jacq et al. (2019) we investigate the topic of transfer learning having a two-agents setting and a q-learner. Furthermore, in contrast with the existing approaches that evaluate explanation by measuring the accuracy of an agent's prediction about another agent behavior, we focus on the effect of the explanation on the agent learning.

3 Experiments

To operationalize the constructs discussed above, we have created an interactive learning scenario allowing both human-agent, and agent-agent interaction. We present initial results that use this interactive scenario to compare different kinds of information provided to the learner.

3.1 Hypothesis

We hypothesize that the agent receiving both, explanations and demonstrations, will learn faster than agents that only receive one of these additional forms of teaching signals. Additionally, all three agents will learn faster than an agent learning by itself.

3.2 Materials

Environment The environment is based on Papi's Minicomputer[1], a competitive two-player game, and it enables learning from explanations, demonstrations, and own experience. Papi's Minicomputer is a non-verbal language to introduce children to mechanical and mental arithmetic through decimal notation with binary positional rules. This environment can be taken as an example of a dynamic, navigational environment. Previous studies involving children, used the same environment, and compared optimal and suboptimal actions, giving an information about the effect of those actions in a certain amount of future steps (Tulli et al., 2020).

Learning Agent The learning agent is an agent that chooses moves using a Q-table. It learns from own experience using q-learning ($\alpha = 0.8$, $\gamma = 0.99$) to solve a Markov Decision Process (MDP), in which the optimal Q-value function is $Q*(s,a) = max_{\pi} Q^{\pi}(s,a)$ (Sutton and Barto, 2005). Examples from demonstrations are treated in the same way (direct q-learning update). Examples from explanations are converted into a format that allows using a q-learning update by summing the reward from the explainer's actual action with the explained reward difference.

Explainer Agent The explainer agent is model-based and plans moves using the depth limited min-max algorithm with search depth of 3. The agent is also capable of giving demonstrations and explanations (see below).

Demonstrations Demonstrations are additional examples given to the learning agent on top of the self-exploration (plain condition). It allows the agent to learn about states and transitions that it has not explored directly by itself. Concretely, to generate a set of demonstrations, the explainer agent selects 10 random states and generates actions for these states according to its policy. It then uses its task model to compute the corresponding next state and computes the reward obtained by this transition. The explainer then gives this information (state, action, next state, reward) to the learner.

Explanations Similar to demonstrations, explanations are examples given to the learning agent on top of the self-exploration (plain condition). However, differently from demonstrations, explanations contrast alternative actions in the same state and aim to suggest a casual relationship between examples by giving a measure of how good the performed action is.

To generate a set of explanations, the explainer agent first computes the actual action that it will perform in the current state. It computes the next state and the reward associated with this transition. Then, it chooses up to three alternative actions at random and simulates the resulting alternative state and associated reward. Finally, the agent computes the difference between the alternative reward and the reward from the actual action.

All this information (current state, actual action, next state, reward, alternative action, alternative state, reward difference) is then combined and given to the learning agent as an explanation. This is the agent-agent scenario equivalent to a natural language encoding using template sentences. Turned into natural language, such an explanation

[1] http://stern.buffalostate.edu/CSMPProgram/String, consulted on Oct 2020

could take the form of: "I am doing *action* which would give me *reward* and lead to *next state*, because doing *alternative action* would lead to *alternative state* and have *reward difference* points more/less."

3.3 Design

We designed an experiment with four conditions: (1) learning from own experience only [**plain**], (2) learning from experience and demonstrations [**demonstration**], (3) learning from experience and explanations [**explanations**], and (4) learning from experience, demonstrations, and explanations [**both**]. For each condition we let the learning agent play against the explainer agent until it has seen $100,000$ examples in total from any source; i.e., to compute the total number of examples we sum the examples from exploration by itself, from demonstration, and from explanations.

In the condition *plain* the learner agent receives 1 example in each step (self-exploration). In the condition *demonstration*, the learner agent receives 11 examples in each step (1 from self-exploration, 10 from demonstration). In the condition *explanation*, the agent receives up to 4 examples in each step (1 from self-exploration, and up to 3 from explanations, depending on how many alternative actions are available in that state). In the condition *both* the learner agent receives up to 14 examples in each step, one from self-explanation, 10 from demonstrations, and up to 3 from explanations. This means that the number of steps and episodes may differ between conditions, but the total number of samples (i.e., examples) is matched between conditions. This means we are providing the same amount of search-space coverage in each condition.

During a single episode of the game, the learning agent updates its policy at every turn. If it is the learning agent's turn, it performs an update based on its own experience (all conditions). If it is the explainer's turn, the learning agent may receive a set of demonstrations and/or explanations - depending on the condition -, which it uses to update its policy. Then, the learning agent updates its policy again based on the explainer's move (all conditions). The explainer does not update its policy in this setup.

To create a dataset to analyze the performance, we train the agent in each condition for $N = 100$ trials (total of 400 trials). We track the outcome of the game (win/loss) and a rolling average (window size 10) of the current win rate.

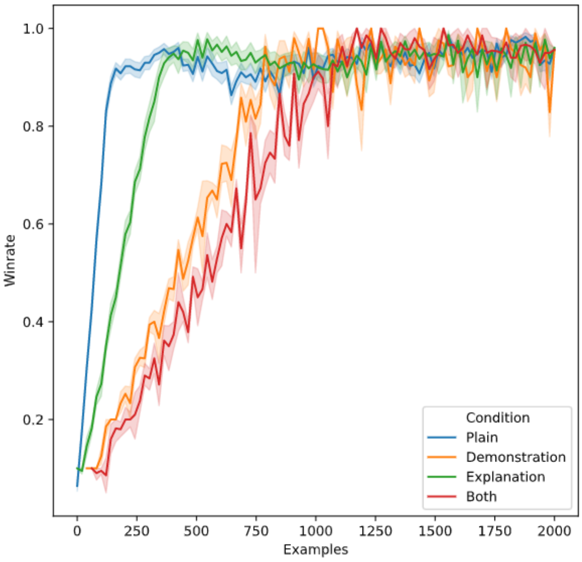

Figure 1: Average (N=100) amount of examples needed to obtain a desired winrate against the explainer agent. The number of examples is calculated as the sum of all examples obtained from self-exploration, demonstrations, and explanations.

3.4 Results

After performing the experiment, we plotted the average number of examples needed for a given winrate grouped by condition (figure 1). The agent begins to perform better than the explainer agent very early in the learning process, which is visualized by a suitable winrate with less than 250 examples. Then, agents from all conditions begin to quickly learn to dominate the explainer agent, with *the agent from the explanation condition requiring the least amount of samples* to win the majority of games. Having access to demonstrations also yields a slight advantage in learning, especially early in the training process. Interestingly, having access to both, demonstrations and explanations, does not lead to improvements.

4 Discussion

In above section we organized the literature on the topic of natural language in interactive learning scenarios involving humans and agents. To date, several excellent works exist on the topic of explainability and natural language technologies, but there it seems to be a gap for experimental work that aims to investigate the concept of explainable AI for transfer learning in both human-agent and agent-agent scenarios.

We expected that the proposed counterfactual structure of an agent's explanations would affect the learning of another agent interacting in the same environment. Overall, the data did not confirm this hypothesis. We assume that the impact of the formalization of the demonstrations and the explanations is less strong than other learning parameters. Furthermore, the access to both demonstrations and the explanations might have influenced erroneously the agent's reasoning about the task. Future work should consider isolating the problem of comparing different types of information employing other rationale that can be suitable, such as inverse reinforcement learning.

Another challenging future direction is represented by the implementation of methods that model the recipient of an explanation. Inferring the learner understanding of the task through partial observations of its state would help in driving the explainer's selection of informative examples.

One of the aspect we neglected in the current study is more realistic and reactive behaviors on both the part of the learner and the explainer. On this subject, while any given agent may not be an expert during learning, accounting for the explainable agency of agents that are not experts remains a topic of future work.

Using counterfactuals to allow agents to understand the effects of their actions seems a promising approach. However, this is not always applicable in complex environment involving humans. If we consider the Hex Game with a number of states of around 10^{92}, generating counterfactuals in natural language might conduct to probabilistic explanations and increase mental overload, leading to performance degradation.

Considering a training corpus of annotated natural language explanations provided by humans appear to be a necessary requirement to extend our findings to human-agent scenarios. Following the same line, testing the effect of agents' explainability on human learning requires challenging long-term studies. The evaluation framework is, in fact, an open challenge. Further evaluation about the effects of the provided explanations on several metrics beyond the human's performance is needed to support our claims.

5 Conclusion

Throughout this paper, we contextualize natural language explanations with a specific focus on learning scenarios. We gave an overview of the existing literature bridging the concept of explanation in humans and artificial agents and showing that explainability is receiving attention in the context of multi-agent settings. We proposed a preliminary computational experiment for comparing demonstrations and explanations and discuss limitations and future work.

Acknowledgements

We acknowledge the EU Horizon 2020 research and innovation program for grant agreement No 765955 ANIMATAS project. This work was supported by national funds through Fundação para a Ciência e a Tecnologia (FCT) with reference UIDB/50021/2020.

References

D. Amir and Ofra Amir. 2018. Highlights: Summarizing agent behavior to people. In *AAMAS*.

Brenna Argall, S. Chernova, M. Veloso, and B. Browning. 2009. A survey of robot learning from demonstration. *Robotics Auton. Syst.*, 57:469–483.

Daniel S. Brown and Scott Niekum. 2018. Machine teaching for inverse reinforcement learning: Algorithms and applications. In *AAAI*.

Y. Chuang, X. Zhang, Yuzhe Ma, M. Ho, Joseph L Austerweil, and Xiaojin Zhu. 2020. Using machine teaching to investigate human assumptions when teaching reinforcement learners. *ArXiv*, abs/2009.02476.

Upol Ehsan and Mark O. Riedl. 2020. Human-centered explainable ai: Towards a reflective sociotechnical approach. *ArXiv*, abs/2002.01092.

P. Fournier, C. Colas, M. Chetouani, and O. Sigaud. 2019. Clic: Curriculum learning and imitation for object control in non-rewarding environments. *IEEE Transactions on Cognitive and Developmental Systems*, pages 1–1.

Brent Harrison, Upol Ehsan, and Mark O. Riedl. 2018. Guiding reinforcement learning exploration using natural language. In *Proceedings of the 17th International Conference on Autonomous Agents and MultiAgent Systems*, AAMAS '18, page 1956–1958, Richland, SC. International Foundation for Autonomous Agents and Multiagent Systems.

Bradley Hayes and Julie A. Shah. 2017. Improving robot controller transparency through autonomous policy explanation. In *Proceedings of the 2017 ACM/IEEE International Conference on Human-Robot Interaction*, HRI '17, page 303–312, New York, NY, USA. Association for Computing Machinery.

Jörg Hoffmann and Daniele Magazzeni. 2019. Explainable ai planning (xaip): Overview and the case of contrastive explanation (extended abstract). In *Reasoning Web*.

Alexis Jacq, Matthieu Geist, Ana Paiva, and Olivier Pietquin. 2019. Learning from a learner. In *Proceedings of the 36th International Conference on Machine Learning*, volume 97 of *Proceedings of Machine Learning Research*, pages 2990–2999, Long Beach, California, USA. PMLR.

F. Keil. 2006. Explanation and understanding. *Annual review of psychology*, 57:227–54.

W. B. Knox, P. Stone, and C. Breazeal. 2013. Training a robot via human feedback: A case study. In *ICSR*.

S. Krening, B. Harrison, K. M. Feigh, C. L. Isbell, M. Riedl, and A. Thomaz. 2017. Learning from explanations using sentiment and advice in rl. *IEEE Transactions on Cognitive and Developmental Systems*, 9(1):44–55.

Isaac Lage, Daphna Lifschitz, Finale Doshi-Velez, and Ofra Amir. 2019. Exploring computational user models for agent policy summarization. *CoRR*, abs/1905.13271.

Toby Jia-Jun Li, Tom Mitchell, and Brad Myers. 2020. Interactive task learning from GUI-grounded natural language instructions and demonstrations. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics: System Demonstrations*, pages 215–223, Online. Association for Computational Linguistics.

Alan Lindsay. 2019. Towards exploiting generic problem structures in explanations for automated planning. In *Proceedings of the 10th International Conference on Knowledge Capture*, K-CAP '19, page 235–238, New York, NY, USA. Association for Computing Machinery.

Rui Liu and X. Zhang. 2017. A review of methodologies for natural-language-facilitated human–robot cooperation. *International Journal of Advanced Robotic Systems*, 16.

T. Lombrozo and Nicholas Z. Gwynne. 2014. Explanation and inference: mechanistic and functional explanations guide property generalization. *Frontiers in Human Neuroscience*, 8.

Jelena Luketina, Nantas Nardelli, Gregory Farquhar, Jakob N. Foerster, Jacob Andreas, E. Grefenstette, S. Whiteson, and Tim Rocktäschel. 2019. A survey of reinforcement learning informed by natural language. *ArXiv*, abs/1906.03926.

Prashan Madumal, Tim Miller, Liz Sonenberg, and Frank Vetere. 2019. Explainable reinforcement learning through a causal lens. *ArXiv*, abs/1905.10958.

T. Miller. 2019. Explanation in artificial intelligence: Insights from the social sciences. *ArXiv*, abs/1706.07269.

A. Najar, Olivier Sigaud, and M. Chetouani. 2020. Interactively shaping robot behaviour with unlabeled human instructions. *Autonomous Agents and Multi-Agent Systems*, 34:1–35.

Anis Najar and Mohamed Chetouani. 2020. Reinforcement learning with human advice. a survey.

E. Schwartz, Guy Tennenholtz, Chen Tessler, and S. Mannor. 2020. Language is power: Representing states using natural language in reinforcement learning. *arXiv: Computation and Language*.

Felipe Leno Da Silva, Garrett Warnell, Anna Helena Reali Costa, and Peter Stone. 2019. Agents teaching agents: a survey on inter-agent transfer learning. *Autonomous Agents and Multi-Agent Systems*, 34:1–17.

Nisan Stiennon, Long Ouyang, Jeff Wu, Daniel M. Ziegler, Ryan Lowe, Chelsea Voss, Alec Radford, Dario Amodei, and Paul Christiano. 2020. Learning to summarize from human feedback.

Theodore R. Sumers, M. Ho, R. D. Hawkins, K. Narasimhan, and T. Griffiths. 2020. Learning rewards from linguistic feedback. *ArXiv*, abs/2009.14715.

R. Sutton and A. Barto. 2005. Reinforcement learning: An introduction. *IEEE Transactions on Neural Networks*, 16:285–286.

M. Taylor. 2018. Improving reinforcement learning with human input. In *IJCAI*.

A. Thomaz, Guy Hoffman, and C. Breazeal. 2005. Real-time interactive reinforcement learning for robots. In *American Association for Artificial Intelligence*.

Silvia Tulli, Marta Couto, Miguel Vasco, Elmira Yadollahi, Francisco Melo, and Ana Paiva. 2020. Explainable agency by revealing suboptimality in child-robot learning scenarios. In *Social Robotics*, pages 23–35, Cham. Springer International Publishing.

Sebastian Wallkotter, Silvia Tulli, Ginevra Castellano, Ana Paiva, and Mohamed Chetouani. 2020. Explainable agents through social cues: A review.