Autonomous Explainable Agents

Venkatsampath Raja Gogineni

Wright State University, Dayton OH 45435, USA
gogineni.14@wright.edu

Abstract. Current trends in Artificial Intelligence are leading to the development of autonomous agents to perform critical operations in the real world. Events in real-world can endanger a wide range of discrepancies and the user should trust the agent to handle them. To achieve this the agent should be able to smartly adapt its behavior to handle the discrepancies and explain it to the human user. This thesis proposes a three-phase approach to address the above-mentioned problem. In the first phase, the agent uses case-based explanations and behavior adaptation in response to a discrepancy. This phase will not only help the agent build its knowledge about the discrepancy, but also forms a basis for its adapted behavior. In the second phase, the agent transforms the knowledge attained from the first phase to explain its behavior to the human operator. This knowledge includes both the causal understanding of the discrepancy and the reasoning behind its adapted behavior. In the final phase, the agent uses the feedback from the human counterpart to adapt its causal knowledge as well as its reasoning behind the behavior adaptation. Finally, this approach will be evaluated through the performance of the agent an underwater mine clearance domain, which is a surveillance mission to create a safe passage for ships.

Keywords: Case selection, case-base explanation, explanation patterns.

1 Introduction

Artificial Intelligence technologies made substantial progress in developing autonomous agents. Although, these agents are designed for very specific applications like driving vehicles or medical diagnosis, they are not completely trusted by their users. To bridge this gap of trust between a human user and the autonomous agent, a branch of AI called explainable Artificial Intelligence (XAI) has gained research traction. XAI focuses on explaining the behavior or decisions of the autonomous agent to the human user. Such an explainable system should develop a rich knowledge base over the time. I propose to acquire this knowledge when there is a discrepancy and transform it to explain to their human operators. Let us look at an example from an underwater mine clearance domain, If the agent finds a mine field at a location where it is not expected to be, then the agent retrieves the hypothetical causal knowledge that an enemy laid the mine. Such causal knowledge can help the agent take a smart decision to apprehend the enemy and resume its survey. Later after the mission the agent can provide the causal knowledge as a reason for its behavior to the human counterpart. Furthermore, the feedback from the human counterpart helps the agent adapt its behavior as well as its causal knowledge. In this example a feedback can help the agent delegate the goal of apprehending the enemy to its counterparts and complete its survey on time.

In conclusion to the approach described earlier there are three phases involved in this process. In the first phase, when a discrepancy is identified the agent uses its causal knowledge to explain the discrepancy and adapt its behavior while in the second phase it uses the causal knowledge along with its behavior adaptation to explain it to the human operators. Finally, in the third phase it uses the feedback to adapt its causal knowledge, reasoning behind the adapted behavior or both.

Section 2 describes a representation of the explanatory cases, their retrieval, behavior adaptation and a possible approach towards explaining the agent's behavior adaptation. Section 3 describes the underwater mine clearance domain and possible discrepancies that may occur in the domain. Related work is illustrated in Section 4 followed by Research plan in Section 5.

2 Case representation, retrieval and behavior adaptation

In this approach, we use case-based explanations [1, 2, 3] to explain a discrepancy. Each case in the case-base is an abstract *explanation pattern (XP)* [4, 5] engineered for a specific domain (see Figure 1). An XP is a data structure that represents a causal relationship between two states and/or actions; each action/state is abstractly defined with variables to be adapted during or after case retrieval. An action or state is referred to as a *node* and different types of nodes are described based on their role in an XP.

- *Explains node*: A discrepancy/unknown state that is observed;
- *Pre-XP node*: Action/state that is observed along with the explains node;
- *XP-asserted node*: Action, state or XP contributing to the explanation's cause.

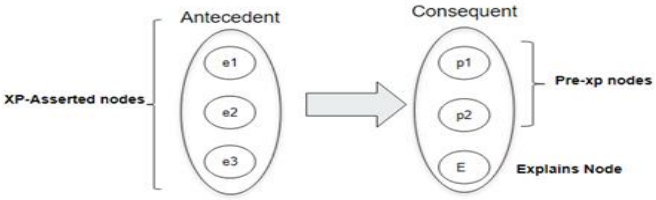

Figure 1: The explanation pattern (XP) causal structure in which XP-asserted nodes (e1, e2, e3) form an antecedent, and a consequent is made up of pre-XP nodes (p1, p2, p3) and an explains node (E); XP-asserted nodes thus cause the associated explains and pre-XP nodes.

2.1 Retrieving, Reusing and Revising Explanation Patterns from a Case Base

Case-based reasoning follows a four-step process to retrieve, reuse, revise and retain cases [6,7]. The following describes how XPs are retrieved, reused and revised.

A set of abstract XPs is retrieved when an unpredicted state or action observed by the agent unifies with each explains node of an XP in the case base. If the unification turns out to be successful then the pre-XP nodes of the corresponding case are unified with the observations of the corresponding states or actions, if they turn out to be successful then the specific XP is retrieved. The retrieved abstract XP is reused by binding variables in the antecedent to values found during unification of the consequent. However, if the XP-asserted nodes in the reused XP contain hypothetical information they can be revised when the new knowledge is obtained from further observations. Although, retention of the revised XP is helpful for improving the case-base it is not the scope of this paper.

In case of multiple XP's leading to a discrepancy, weights can be associated to an XP. These weights can be based on the frequency of its retrieval in the domain or can be based on the number of evidences obtained. However, the method to calculate weights is not in the scope of this paper.

2.2 Behavior adaptation and Incorporating Human Feedback

Behavior adaptation is essential for an intelligent agent to respond to discrepancies [8, 9]; in this approach, we formulate goals as a process of behavior adaptation. Goals are formulated by preventing the recurrence of one or more explanation antecedent nodes. Antecedent nodes may include actions and/or states; therefore, when the agent wishes to prevent an undesired consequent from recurring, it considers the elimination of antecedent actors or objects that participate in antecedent states as potential goals.

The agent's explanation to the human operator increases trust between them. As discussed in the previous sections, an XP is a data structure with the causal representation of antecedents leading to a consequent (discrepancy). A template created with a discrepancy, antecedents of the retrieved XP and the newly formulated goal will explain the agent's adapted behavior to the human operator. Moreover, feedback from the human operator can assist the agent in giving weights to the explanations in the case base. This can be beneficial to the agent in retrieving the appropriate causal knowledge.

3 Underwater Mine Clearance Domain

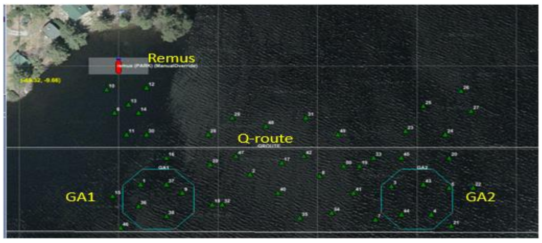

Figure 2: Underwater Mine Clearance domain with two clearance areas in the Q-route.

Our approach will be implemented in a limited mine clearance domain [10], which is simulated using MOOS-IVP [11], software that provides complete autonomy for marine vehicles. Figure 2 shows the simulation of the mine clearance domain with the agent as a Remus unmanned underwater vehicle. The Q-route is a safe passage for ships to enter and leave the port and is represented as a rectangular area in Figure 2. GA1 and

GA2 are the two octagonal areas where mines are expected to exist, while the triangular objects are the mines. The goals of the agent are to survey and clear mines in GA1 and GA2. These goals are given to the agent after a reconnaissance mission performed by a different agent across the whole sea route.

In the underwater mine clearance domain, several events often co-occur simultaneously, and many events cannot be predicted based on knowledge available to an agent. These events might affect the agent itself or the mission of the agent. Explanations help the agent to recognize these events and respond to them. We will look at several uncertain events that might happen.

Events in this domain include minelaying, sensor failure, and reconnaissance failure. Minelaying events occur when an enemy ship, aerial vehicle, or fishing vessel lays traps to hurt friendly ships. Sensor failure event indicates that the agent's faulty sensor is responsible for a misclassification of mine, and the failure of proper reconnaissance mission indicates that an agent prior to the agent did not identify mines which in turn failed its mission.

4 Related Research

Generating causal knowledge to explain a discrepancy is not novel in this approach. However, reasoning about the causal knowledge to adapt agent's behavior is novel. Schank [4] introduced Explanation Patterns (XP) as a knowledge structure to handle the causal knowledge about a discrepancy. Later Ram [12] provided an approach to learn these XP's. Our recent work [13] demonstrates the use of a case base of explanations to respond to a discrepancy and adapt the agent's behavior in the underwater mine clearance domain.

Roth-Berghofer et al's [14] work on classifying explanations and their use-cases according to the user's intentions is one of the theoretical research directions towards explanations in case-based reasoning (see also [15]). This paper introduces the concept of "explanation goals" that are used to decide when and what the system should explain to users based on their expectations. We will investigate application of these techniques to prevent the system from repeatedly explaining the same type of unexpected events to a user who is already familiar with them.

Floyd et al. [16] demonstrated that the behavior adaptation from the human feedback increases the trust as well as the efficiency of the agent to perform in teams. However, this work doesn't consider the role of explanations in the human feedback.

5 Research Plan

Table 1 shows the research plan towards implementing the proposed idea. Each action item in the plan is preceded by a literature review on the topic and followed by evaluation. Evaluation is performed as an overall performance of the agent in the underwater mine clearance domain. The overall performance of the agent is calculated as the number of ships that traverse through the Q-route in Figure 2.

Table 1: Proposed Research Plan

Plan	Date
Discrepancy Explanation and behavior adaptation	August 2018
Selecting an explanation case from multiple cases	May 2019
Explanation to external agents	August 2019
Behavior adaptation from human feedback	March 2020
Case adaptation from human feedback	August 2020
Evaluation in Underwater Mine Clearance domain	December 2020

References

1. Schank, R. C., Kass, A., Riesbeck, C. K.: Inside case-based explanation. Psychology Press (2014).
2. Cox, M. T., Burstein, M. H.: Case-based explanations and the integrated learning of demonstrations. Künstliche Intelligenz (Artificial Intelligence), 22(2), 35-38 (2008).
3. Ram, A.: Indexing, elaboration and refinement: Incremental learning of explanatory cases. Machine Learning, 10, 201-248 (1993).
4. Schank, R. C.: Explanation patterns: Understanding mechanically and creatively. Psychology Press (2013).
5. Ram, A.: A theory of questions and question asking. Journal of the Learning Sciences, 1, (3 and 4), 273-318 (1991).
6. Aamodt, A., Plaza, E.: Case-Based Reasoning: Foundational Issues, Methodological Variations, and System Approaches. AI Communications 7(1), 39–52 (1994).
7. Kolodner, J.: Case-Based Reasoning. Morgan Kaufmann, San Francisco (1993).
8. Maynord, M., Cox, M. T., Paisner, M., Perlis, D.: Data-driven goal generation for integrated cognitive systems. In 2013 AAAI Fall Symposium Series, (2013).
9. Hanheide, M., Hawes, N., Wyatt, J., Göbelbecker, M., Brenner, M., Sjöö, K., and Kruijff, G. J.: A framework for goal generation and management. In Proceedings of the AAAI workshop on goal-directed autonomy, (2010).
10. Kondrakunta, S., Gogineni, V., Molineaux, M., Munoz-Avila, H., Oxenham, M., Cox, M. T.: Toward Problem Recognition, Explanation and Goal Formulation. In: Proceedings of the 6th Goal Reasoning Workshop at IJCAI/FAIM-2018. Stockholm, Sweden (2018).
11. Benjamin, M. R., Schmidt, H., Newman, P. M., Leonard, J. J.: (2010). Nested autonomy for unmanned marine vehicles with MOOS-IvP. Journal of Field Robotics, 27(6), 834-875 (2010).
12. Ram, A.: Indexing, elaboration and refinement: Incremental learning of explanatory cases. Machine Learning, 10, 201-248 (1993).
13. Gogineni, V., Kondrakunta, S., Molineaux, M., Cox, M. T.: Application of Case-based Explanations to Formulate Goals in an Unpredictable Mine Clearance Domain. In: Proceedings of the ICCBR-2018 workshop on Case-Based Reasoning for the Explanation of Intelligent Systems, pp. 42-51. Stockholm, Sweden (2018).
14. Roth-Berghofer, T. R., Jörg Cassens.: Mapping goals and kinds of explanations to the knowledge containers of case-based reasoning systems. In: International Conference on Case-Based Reasoning. Springer, Berlin, Heidelberg (2005).
15. Aamodt, A. (1994). Explanation-Driven Case-Based Reasoning. In S.Wess, K.Althoff, M.Richter (eds.): Topics in case-based reasoning (pp 274-288).Berlin: Springer.
16. Floyd, M.W., Drinkwater, M., & Aha, D.W. (2015). Trust-guided behavior adaptation using case-based reasoning. Proceedings of the Twenty-Fourth International Joint Conference on Artificial Intelligence (pp. 4261-4267). Buenos Aires, Argentina: AAAI Press.