Explaining Aha! moments in artificial agents through IKE-XAI: Implicit Knowledge Extraction for eXplainable AI
Ikrarn Chraïbi Kaadouda*, Adrien Bennetoc, Barbara Mawhinb, Vicky Charis, Natalia Diaz-Rodriguezc

*IMT Atlantic, Lab-STICC, UMR CNRS 6285, F-29238 Brest, France.
†Univ. Rennes, INSA-Orléans, INSA-Orléans, UMR 6397, FR2954, F-49000
‡University of Bordeaux, INSA Bordeaux, UMR CNRS 5294, L2S, F-33405
§Université Paris-Diderot, Sorbonne Université, CREST, Inria, F-75014
||ENS Paris-Saclay, CEA-List, Pole Universitaire de Grenoble, F-38000
&Université Grenoble Alpes, CNRS, LJK, UMR 5240, GREYC, F-38000

Abstract
During the learning process, a child develops a mental representation of the task he or she is learning. A Machine Learning algorithm develops a latent representation of the task it learns. We investigate the development of the knowledge construction of an artificial agent (AA) by getting inspiration from the one of children. Our main contribution is a 3-step methodology named Implicit Knowledge Extraction with eXplainable Artificial Intelligence (IKE-XAI) to extract the implicit knowledge, in form of an automaton, encoded by an artificial agent (AA) during its learning. We showcase this technique to solve and explain the Tower of Hanoi (TOH) task when researchers have only access to sequences of moves that represent observational behavior as in human-machine interaction. Our approach combines: 1) a Q-learning agent that learns to perform the TOH task; 2) a trained LSTM recurrent neural network that encodes an implicit representation of the TOH task; and 3) an XAI procedure that builds an hierarchical rule-extraction algorithm to extract finite state automata.

We propose using graph representations as visual and explicit explanations of the behavior of the Q-learning agent. Our experiments show that the IKE-XAI approach helps understanding the development of the Q-learning agent behavior by providing a global explanation of its knowledge evolution during learning. IKE-XAI also allows researchers to identify the agent’s Aha! moment by determining from what moment the knowledge representation stabilizes and the agent no longer learns. This work is published in Neural Network journal (DOI=10.1016/j.neunet.2022.08.002) available at the QR code above.

IKE-XAI methodology: Implicit Knowledge Extraction for eXplainable AI

- Developmental Robotics: Machine learning
- Cognitive developmental psychology: Sequence/Models
- Explaining AI: Sequence/Models
- Experimental protocol

RESULTS
- Explicit automaton forecasting
- Predictions and outcomes

Main findings
IKE-XAI, a post-hoc explainable methodology that provides a visual model-agnostic explanation based on the observational behavior of an AA, allows to:
- Extract the vision of the AA of a task (simple and complex one) using a sequence learning model
- Extract knowledge, in the form of FSA that represents AA’s problem-solving strategies, even not optimal ones, for their explainability.
- Make explicit the behavioral changes of an AA due to the analysis of the edge weights of the extracted automata, i.e., the transformation of its expertise in solving the task.
- Identify the shift in the AA’s behavior from exploitation i.e., Aha! moment for the agent and the Aha! moment for the researcher when he/she understands when it happens

Experiments on TOH with variable N disks:

<table>
<thead>
<tr>
<th>N</th>
<th>2</th>
<th>3</th>
<th>4</th>
</tr>
</thead>
<tbody>
<tr>
<td>N 2</td>
<td>15</td>
<td>63</td>
<td>15</td>
</tr>
<tr>
<td>N 3</td>
<td>27</td>
<td>72</td>
<td>72</td>
</tr>
<tr>
<td>N 4</td>
<td>39</td>
<td>1092</td>
<td>1092</td>
</tr>
<tr>
<td>N 5</td>
<td>500</td>
<td>100000</td>
<td>100000</td>
</tr>
<tr>
<td>N 6</td>
<td>159</td>
<td>21500</td>
<td>21500</td>
</tr>
</tbody>
</table>

Average length of sequences at the beginning of training

Average length of sequences after the Aha! moment

Experiments on TOH with variable N disks:

<table>
<thead>
<tr>
<th>N</th>
<th>2</th>
<th>3</th>
<th>4</th>
</tr>
</thead>
<tbody>
<tr>
<td>N 2</td>
<td>15</td>
<td>63</td>
<td>15</td>
</tr>
<tr>
<td>N 3</td>
<td>27</td>
<td>72</td>
<td>72</td>
</tr>
<tr>
<td>N 4</td>
<td>39</td>
<td>1092</td>
<td>1092</td>
</tr>
<tr>
<td>N 5</td>
<td>500</td>
<td>100000</td>
<td>100000</td>
</tr>
<tr>
<td>N 6</td>
<td>159</td>
<td>21500</td>
<td>21500</td>
</tr>
</tbody>
</table>

Average length of sequences at the beginning of training

Average length of sequences after the Aha! moment

Main findings
IKE-XAI, a post-hoc explainable methodology that provides a visual model-agnostic explanation based on the observational behavior of an AA, allows to:
- Extract the vision of the AA of a task (simple and complex one) using a sequence learning model
- Extract knowledge, in the form of FSA that represents AA’s problem-solving strategies, even not optimal ones, for their explainability.
- Make explicit the behavioral changes of an AA due to the analysis of the edge weights of the extracted automata, i.e., the transformation of its expertise in solving the task.
- Identify the shift in the AA’s behavior from exploitation i.e., Aha! moment for the agent and the Aha! moment for the researcher when he/she understands when it happens

Summary of what IKE-XAI provides:
1. Optimal strategy: key action to perform the task
2. Temporal explanation of acquired knowledge evolution towards Aha! moment
3. Novel insight in learning

Use case: the TOH with N = 3 disks
(a) Sequence of visited states
(b) Sequence of moves
(c) Sequence of moves encapsulated

Knowledge: A set of facts, information, and skills acquired through experience by the AA that contribute to gaining a theoretical or practical understanding of a subject or the world.

Experiments and Results

The TOH with N = 3 disks:

- Number of nodes
- Number of edges
- Aha! moment (average number of training episodes)
- Average length of sequences at the beginning of training
- Average length of sequences after the Aha! moment

Meanings of nodes and edges:

1. Node: a state of the problem, corresponding to the position of the disks.
2. Edge: an action that moves a disk from one state to another, respecting the rules of the game.

Main findings
IKE-XAI, a post-hoc explainable methodology that provides a visual model-agnostic explanation based on the observational behavior of an AA, allows to:
- Extract the vision of the AA of a task (simple and complex one) using a sequence learning model
- Extract knowledge, in the form of FSA that represents AA’s problem-solving strategies, even not optimal ones, for their explainability.
- Make explicit the behavioral changes of an AA due to the analysis of the edge weights of the extracted automata, i.e., the transformation of its expertise in solving the task.
- Identify the shift in the AA’s behavior from exploitation i.e., Aha! moment for the agent and the Aha! moment for the researcher when he/she understands when it happens