



HAL
open science

Explaining Aha! moments in artificial agents through IKE-XAI: Implicit Knowledge Extraction for eXplainable AI

Ikram Chraibi Kaadoud, Adrien Bennetot, Barbara Mawhin, Vicky Charisi,
Natalia Díaz-Rodríguez

► To cite this version:

Ikram Chraibi Kaadoud, Adrien Bennetot, Barbara Mawhin, Vicky Charisi, Natalia Díaz-Rodríguez. Explaining Aha! moments in artificial agents through IKE-XAI: Implicit Knowledge Extraction for eXplainable AI. Women in Machine Learning WIML @ Neurips 2022, Nov 2022, Louisiane, United States. , 2022. hal-03875691

HAL Id: hal-03875691

<https://hal.science/hal-03875691>

Submitted on 28 Nov 2022

HAL is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers.

L'archive ouverte pluridisciplinaire **HAL**, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d'enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.



Distributed under a Creative Commons Attribution 4.0 International License

Explaining *Aha!* moments in artificial agents through IKE-XAI: Implicit Knowledge Extraction for eXplainable AI

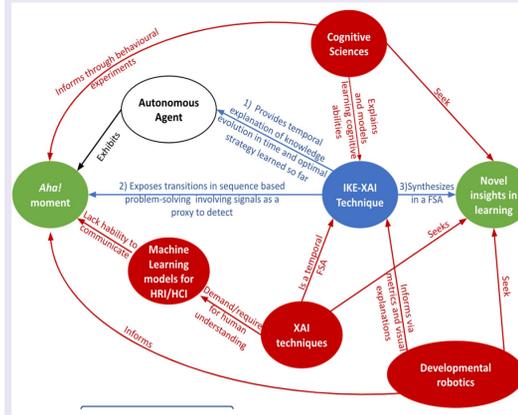
Ikram Chraïbi Kaadoud^a, Adrien Bennetot^{b,c,d}, Barbara Mawhin^e, Vicky Charisif, Natalia Díaz-Rodríguez^g

Paper in *Neural Networks*
 DOI=10.1016/j.neunet.2022.08.002

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:

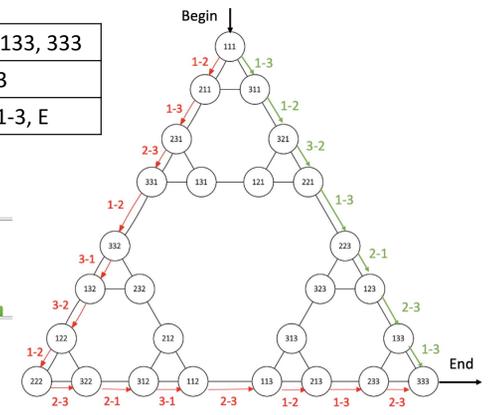
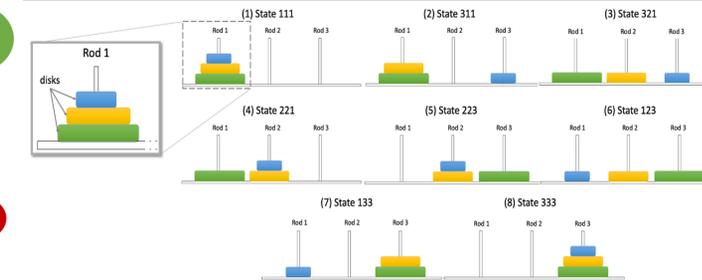
- 1) a Q-learning agent that learns to perform the TOH task;
 - 2) a trained LSTM recurrent neural network that encodes an implicit representation of the TOH task; and
 - 3) an XAI process using a post-hoc implicit rule extraction algorithm to extract finite state automata.
- We propose using **graph 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.



Context

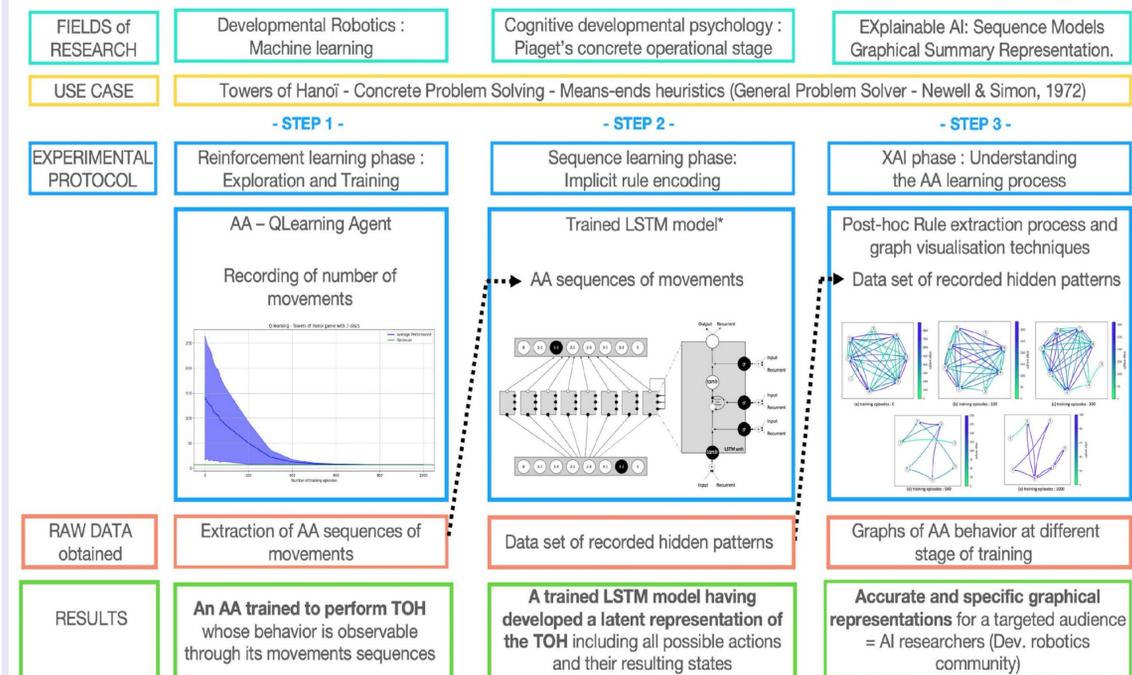
Use case: the TOH with N = 3 disks

(a) Sequence of visited states	111, 311, 321, 221, 223, 123, 133, 333
(b) Sequence of moves	1-3, 1-2, 3-2, 1-3, 2-1, 2-3, 1-3
(c) Sequence of moves encapsulated	B, 1-3, 1-2, 3-2, 1-3, 2-1, 2-3, 1-3, E



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.

IKE-XAI methodology : Implicit Knowledge Extraction for eXplainable AI



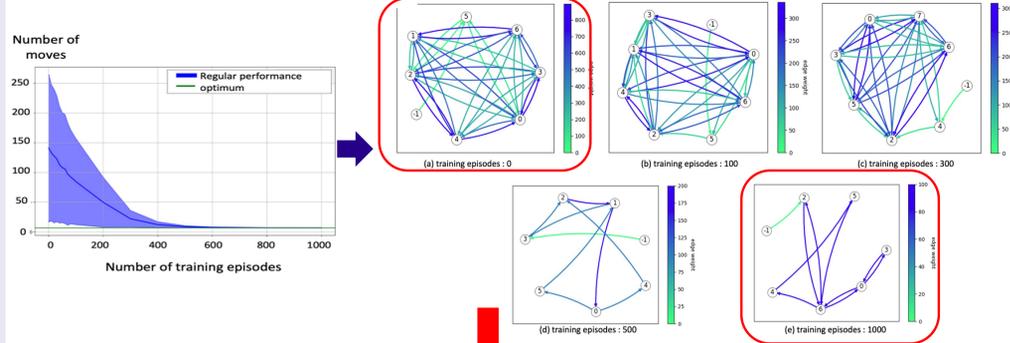
* Trained LSTM model : a RNN with LSTM units that learned sequences of moves generated from Towers of Hanoi to predict the next move according to the past and current ones

Summary of what IXE-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

Experiments & Results

The TOH with N = 3 disks:



Experiments on TOH with variable N disks:

	N = 3	N = 4	N = 6
Optimal number of moves $2^N - 1$	7	15	63
Number of nodes	27	81	729
Number of edges	39	120	1092
Aha! moment (average number of training episodes)	500	3000	100000
Average length of sequences at the beginning of training	159	800	21500
Average length of sequences after the <i>Aha!</i> moment	9	15	63

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 exploration to exploitation i.e., *Aha!* moment for the agent and the *Aha!* moment for the researcher when he/she understands when it happens

FSA Edge (source node - target node)	Associated action (move)	Computed weight	FSA Edge (source node - target node)	Associated action (move)	Computed weight
"-1"-5	B	1	"-1"-111	B	1
5 - 0	"1-3"	58	2 - 6	"1-3"	100
5 - 2	"1-2"	42	6 - 4	"1-2"	100
0 - 3	"3-1"	882	4 - 5	"3-2"	100
0 - 2	"1-2"	335	5 - 6	"1-3"	100
0 - 6	"1-3"	261	6 - 0	"2-3"	100
0 - 4	"2-1"	538	0 - 3	"3-1"	100
0 - 1	"3-2"	658	3 - 0	"2-3"	100
0 - 5	B	82	0 - 6	"1-3"	100
0 - 0	"1-3"	20	3 - 2	B	99
3 - 2	"1-2"	650			

(a) training episode:0 (truncated table)

(e) training episode:1000