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

Ikram Chraibi Kaadoud[a], Adrien Bennetot[b,c,d], Barbara Mawhin[e], Vicky Charisi[f], Natalia Díaz-Rodríguez[g]

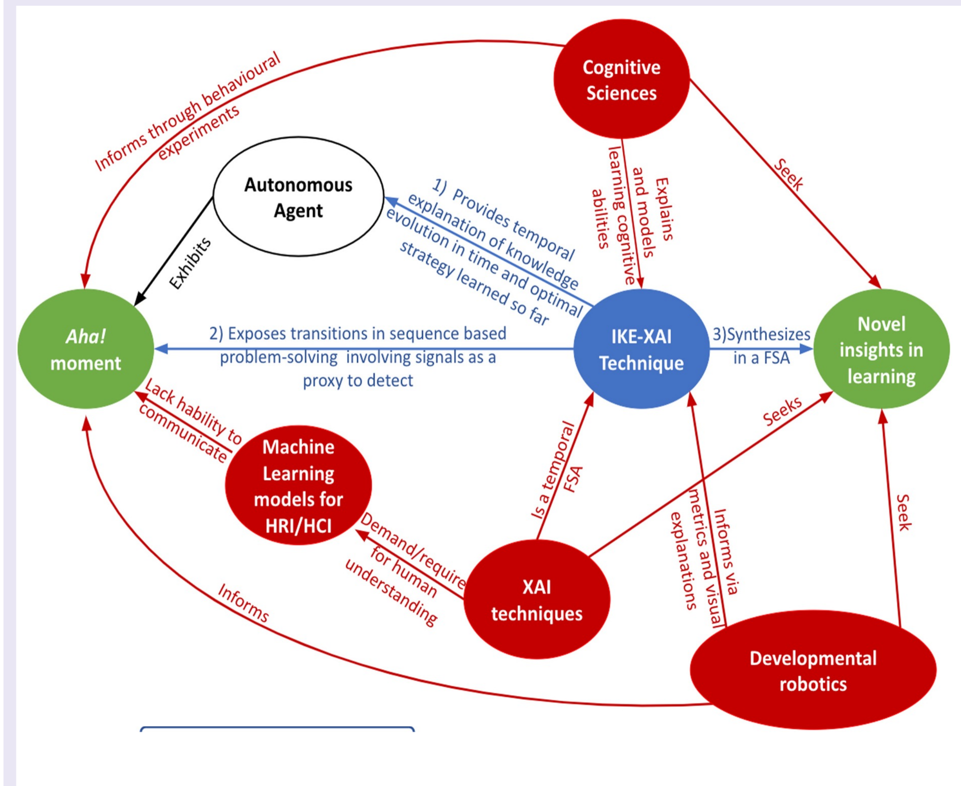
[a] IMT Atlantique, Lab-STICC, UMR CNRS 6285, F-29238 Brest, France,
[b] U2IS Dept., ENSTA, Institut Polytechnique Paris, Inria Flowers Team, 828, Boulevard des Maréchaux 91762 Palaiseau Cedex, France, [c] Segula Technologies, Parc d'activité de Pissaloup, Trappes, France,
[d] Institut des Systèmes Intelligents et de Robotique, Sorbonne Université, Paris, France
[e] Human Factors Department, EBT-Salient Aero Foundation, Spain, [f] European Commission, Joint Research Center (JRC), Seville, Spain,
[g] DaSCI Andalusian Institute in Data Science and Computational Intelligence, AI Lab, CITIC, University of Granada, 18071 Granada, Spain

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

## Abstract

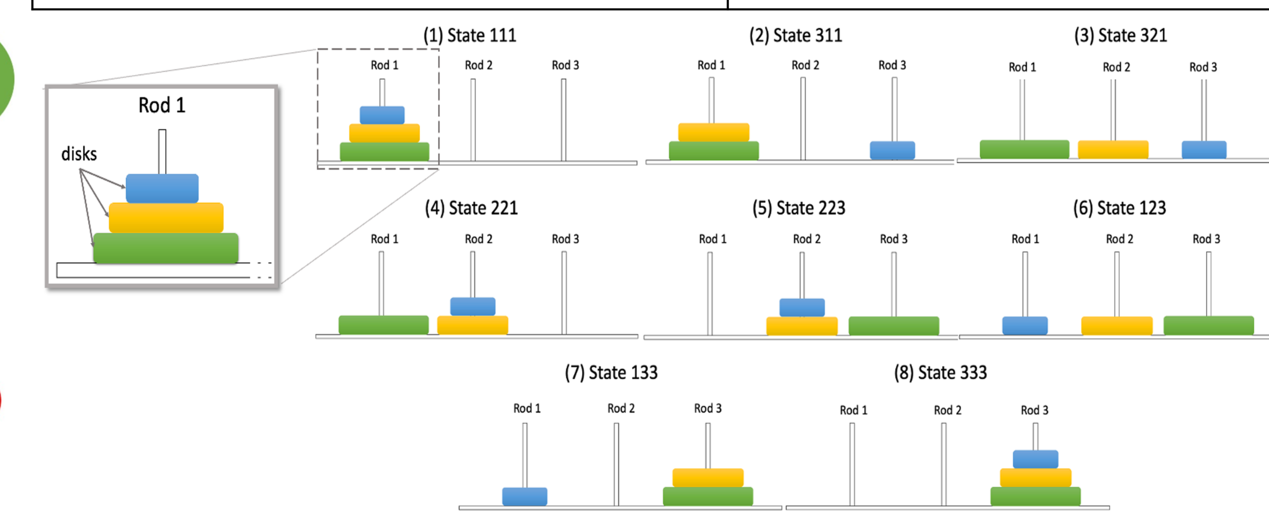
During the learning process, a child develops a mental representation of the task he or she is learning. A Machine Learning algorithm develops a latent representation of the task it learns. We investigate the development of the knowledge construction of an artificial agent (AA) by getting inspiration 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 Hanoï (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.

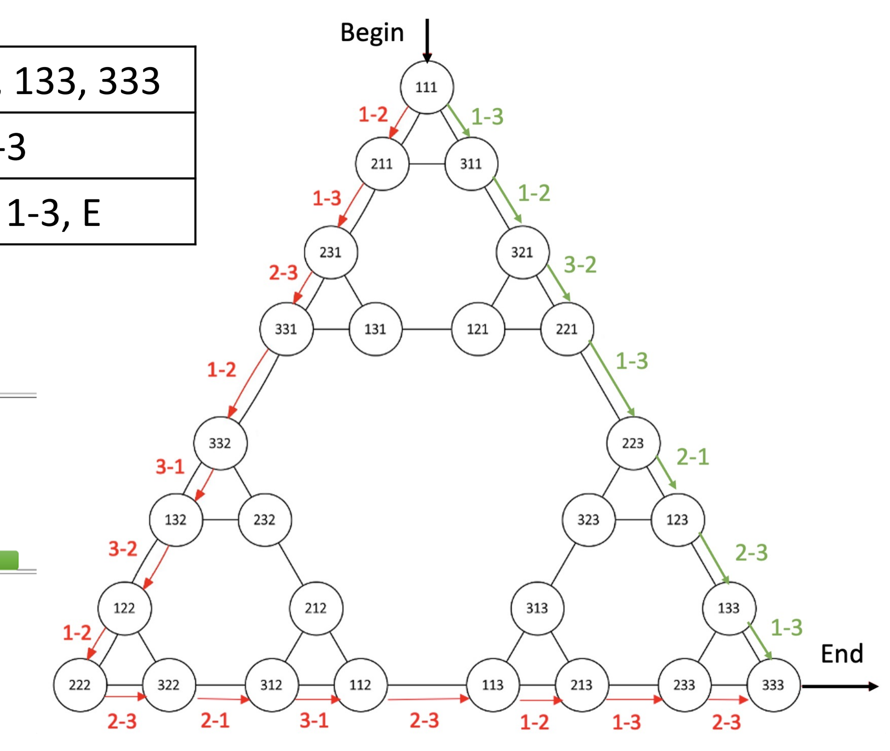
## Context




**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 |

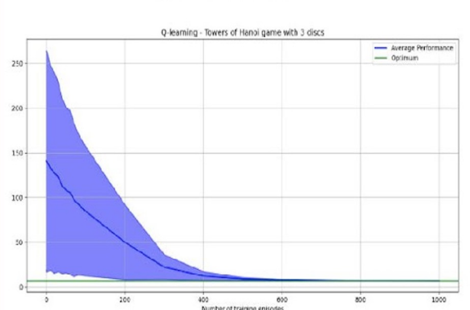


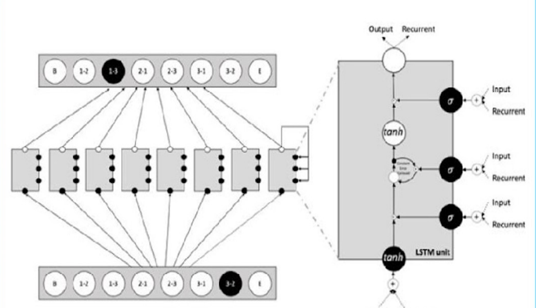



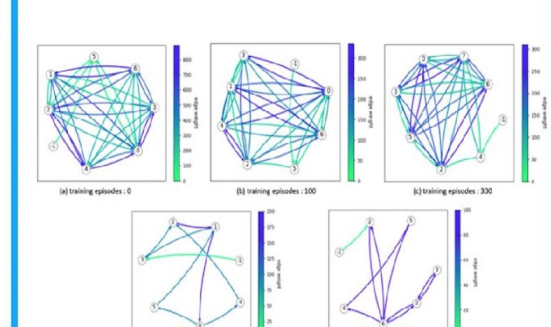

**Knowledge:** A set of facts, information, and skills acquired through experience by the AA that contribute to gaining a theoretical or practical understanding of a subject or the world.

## IKE-XAI methodology : Implicit Knowledge Extraction for eXplainable AI

| | - STEP 1 - | - STEP 2 - | - STEP 3 - |
|---|---|---|---|
| FIELDS of RESEARCH | Developmental Robotics : Machine learning | Cognitive developmental psychology : Piaget's concrete operational stage | EXplainable AI: Sequence Models Graphical Summary Representation. |
| USE CASE | Towers of Hanoï - Concrete Problem Solving - Means-ends heuristics (General Problem Solver - Newell & Simon, 1972) | | |
| EXPERIMENTAL PROTOCOL | Reinforcement learning phase : Exploration and Training | Sequence learning phase: Implicit rule encoding | XAI phase : Understanding the AA learning process |
| | AA – QLearning Agent<br>Recording of number of movements | Trained LSTM model*<br>AA sequences of movements | Post-hoc Rule extraction process and graph visualisation techniques<br>Data set of recorded hidden patterns |
| RAW DATA obtained | Extraction of AA sequences of movements | Data set of recorded hidden patterns | Graphs of AA behavior at different stage of training |
| RESULTS | **An AA trained to perform TOH** whose behavior is observable through its movements sequences | **A trained LSTM model having developed a latent representation of the TOH** including all possible actions and their resulting states | **Accurate and specific graphical representations** for a targeted audience = AI researchers (Dev. robotics community) |







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

**Summary of what IXE-XAI provides:**

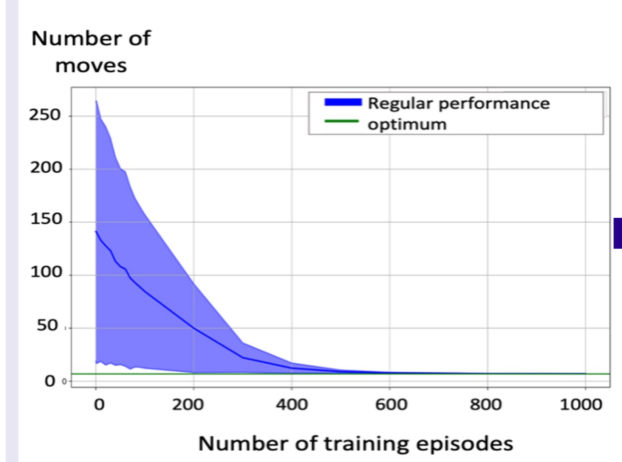
-1- Optimal strategy: key action to perform the task

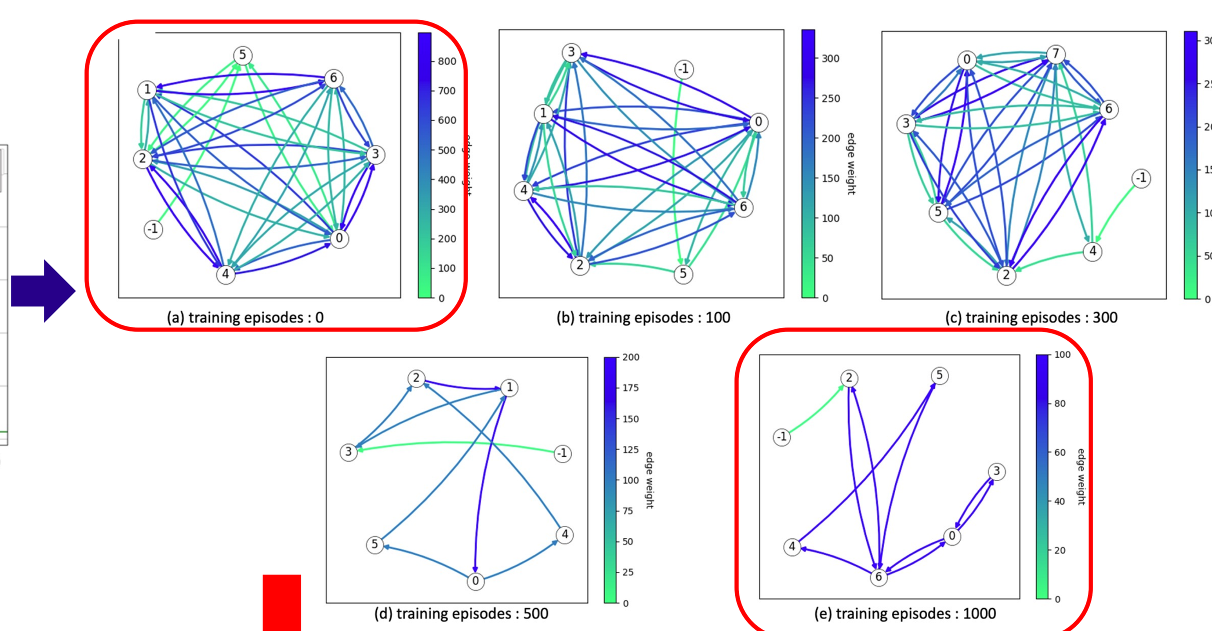
-2- Temporal explanation of acquired knowledge evolution towards *Aha!* moment

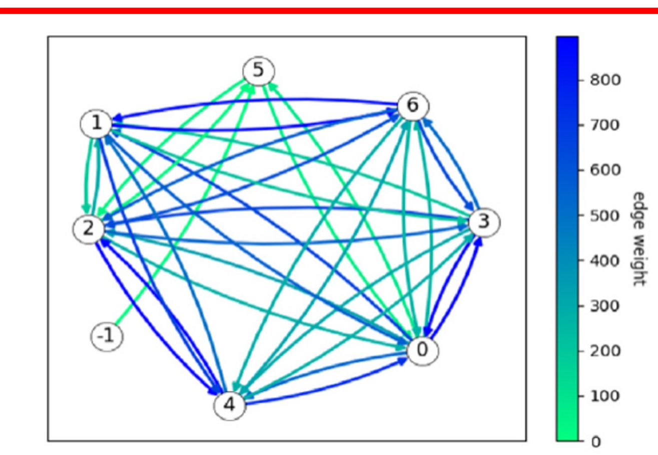
-3- Novel insight in learning

## Experiments & Results

**The TOH with N = 3 disks:**

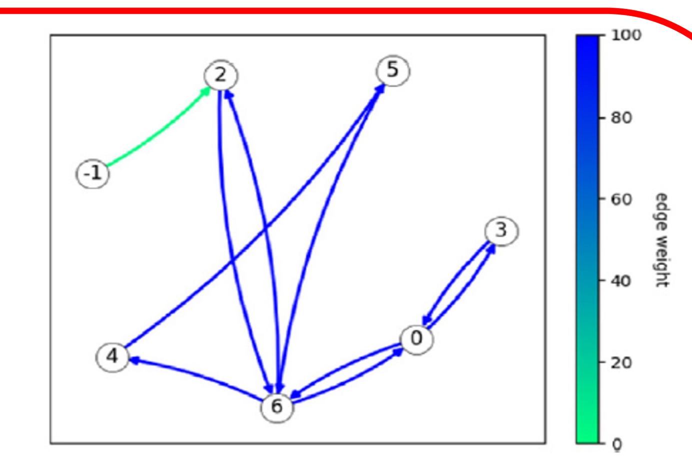


| FSA Edge (source node - target node) | Associated action (move) | Computed weight |
|---|---|---|
| "-1" - 5 | B | 1 |
| 5 - 0 | "1-3" | 58 |
| 5 - 2 | "1-2" | 42 |
| 0 - 3 | "3-1" | 882 |
| 0 - 2 | "1-2" | 335 |
| 0 - 6 | "2-3" | 261 |
| 0 - 4 | "2-1" | 538 |
| 0 - 1 | "3-2" | 658 |
| 0 - 5 | B |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

(a) training episode:0 (truncated table)



| FSA Edge (source node - target node) | Associated action (move) | Computed weight |
|---|---|---|
| "-1" - 111 | B | 1 |
| 2 - 6 | "1-3" | 100 |
| 6 - 4 | "1-2" | 100 |
| 4 - 5 | "3-2" | 100 |
| 5 - 6 | "1-3" | 100 |
| 6 - 0 | "2-3" | 100 |
| 0 - 3 | "3-1" | 100 |
| 3 - 0 | "2-3" | 100 |
| 0 - 6 | "1-3" | 100 |
| 6 - 2 | B | 99 |

(e) training episode:1000

**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 *Aha!* 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

WIML @Neurips 2022, Lousiana, USA, December 2022