EXPLANATORY EFFECTS OF MACHINE LEARNED LOGIC THEORIES

Lun Ai[1],**Stephen H. Muggleton**[1], **Ute Schmid**[2], **Céline Hocquette**[1], **Mark Gromowski**[2] and **Johannes Langer**[2]

[1]Department of Computing, Imperial College London; [2]Cognitive Systems Group, University of Bamberg

XAI Background

Common drawbacks in relevant studies were identified and discussed in recent surveys which are summarised as follows:

- **Ambiguous**
- **Subjective**
- **Disconnection with social science literature**
- **Neglect the human perspective**
- **Focused on only beneficial effects**

Objective - Explore comprehensibility of machine learned logic programs in interactive machine-human teaching contexts.

Learning logic theories

Inductive Logic Programming (ILP) uses logic programs to

- derive declarative logic rules
- learn programs from small data
- perform abduction and induction

Meta-Interpretive Learning (MIL) is a variant of ILP that

- supports predicate invention and dependent learning
- can learn recursive and higher-order programs

Human comprehension

Challenge: Difficult to operationally measure comprehension

Method: Human comprehension measured by human out-of-sample predictive accuracy in human trials

Given a definition D, a group of humans H, a symbolic machine learning algorithm M, examples E, $C_{ex}(D, H, M(E))$ denotes **machine-explained human comprehension** after studying explanations and $C(D, H, E)$ is the **unaided human comprehension**.

We define the **explanatory effect** $E_{ex}(D, H, M(E))$ of the theory $M(E)$ learned from examples E as

$$E_{ex}(D, H, M(E)) = C_{ex}(D, H, M(E)) - C(D, H, E)$$

Cognitive window

The **explanatory effectiveness** of a theory is defined as

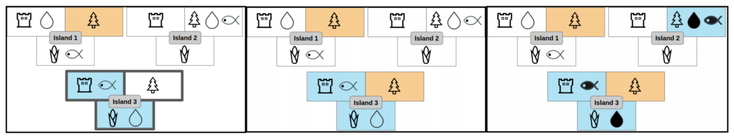
- $M(E)$ is *beneficial* to H if $E_{ex}(D, H, M(E)) > 0$
- $M(E)$ is *harmful* to H if $E_{ex}(D, H, M(E)) < 0$
- Otherwise, $M(E)$ does not have an observable effect on H

Framework: We hypothesise a bound on human hypothesis space size and estimate the cognitive complexity by a new variant of Kolmogorov complexity. The two constraints on human comprehension are summarised in the **cognitive window** (CW) conjecture:

1. the search space of the theory cannot be too large
2. the theory provides mental execution "shortcuts"

Beneficial and harmful explanatory machine learning [1]

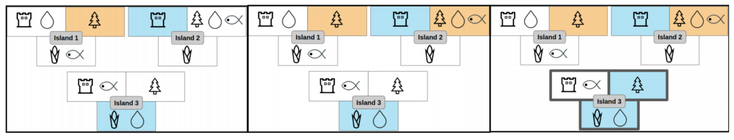
Materials: Humans were asked to learn a masked isomorphic problem of Noughts and Crosses. Explanations are translated from a machine learned logic program by a MIL system. An example of visual and textual explanations used in our two-group human experiment is presented below.



You select this territory and obtain 1 pair (Island 3)

Opponent conquers and prevent you from getting a triplet (Island 3)

You obtain 2 pairs (Water, Fish) and opponent has no pair



Contrast: Not enough pair(s)

Contrast: Not enough pair(s)

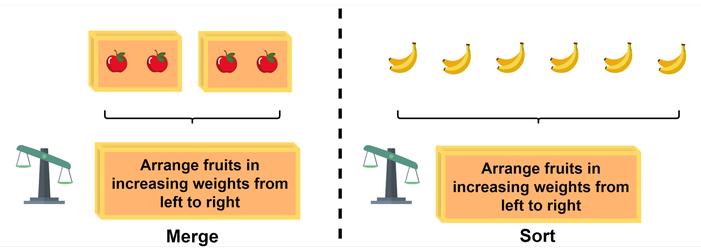
Results summary:

- A clear boundary of human short-term comprehension
- A **beneficial effect** on human comprehension when CW is satisfied
- A **harmful effect** on human comprehension when CW is not satisfied

Explanatory machine learning for sequential human teaching [2]

Framework: Extensions of frameworks of explanatory effects to account for the impacts of reducing the size of the hypothesis space when learning with increasing concept complexity

Materials: A human trial of **sequential teaching** (ST) was conducted for teaching efficient sorting strategies. We examine the effects of concept ordering in curricula and explanations learned by a MIL system. An example of merging and sorting materials used for teaching novices merge sort is presented below.



Results summary:

- a **beneficial effect** on human comprehension when learning with increasing concept complexity
- **re-discovery of advanced and optimised algorithms** when learning with increasing concept complexity

Contributions

- Provided operational frameworks of explanatory effects
- Demonstrated beneficial and harmful explanatory effects
- Showed ST led to better human comprehension and innovative re-discovery to benefit machine-human teaching

References

[1] L. Ai et al. "Beneficial and harmful explanatory machine learning." In: *Machine Learning* 110 (2021), pp. 695–721. DOI: https://doi.org/10.1007/s10994-020-05941-0.

[2] L. Ai et al. "Explanatory machine learning for sequential human teaching". In: *arXiv* (2022). DOI: 10.48550/ARXIV.2205.10250.