

BENEFICIAL AND HARMFUL EXPLANATORY MACHINE LEARNING

Lun Ai¹, Stephen H. Muggleton¹, Céline Hocquette¹, Mark Gromowski² and Ute Schmid²

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

XAI Background

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

- Under-specified and ambiguous definitions
- A lack of empirical data to support claims
- Limited references to valuable social science literature
- No or little accounting for humans' perspective
- Not enough emphasis recently on the harmful side

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

MIL and predicate invention

Meta-Interpretive Learning (MIL) is a sub-field of Inductive Logic Programming (ILP). Given higher-order clauses \mathcal{M} MIL uses logic programming to represent examples by a program \mathcal{H} and background knowledge \mathcal{B} ,

$$\begin{aligned} \forall e+ \in E \quad \mathcal{H} \cup \mathcal{B} \cup \mathcal{M} &\models e+ \\ \forall e- \in E \quad \mathcal{H} \cup \mathcal{B} \cup \mathcal{M} &\not\models e- \end{aligned}$$

MIL supports predicate invention, dependent learning, learning of recursions and higher-order programs.

Human comprehension

Given a definition D , a group of humans H , a symbolic machine learning algorithm M , **explanatory effect** $E_{ex}(D, H, M(E))$ of the theory $M(E)$ learned from examples E is

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

Machine-explained human comprehension $C_{ex}(D, H, M(E))$,

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

is the mean accuracy of H after brief study of an explanation based on $M(E)$ can classify new material selected from the domain of D . $C(D, H, E)$ is the unaided human comprehension of examples E .

Cognitive window

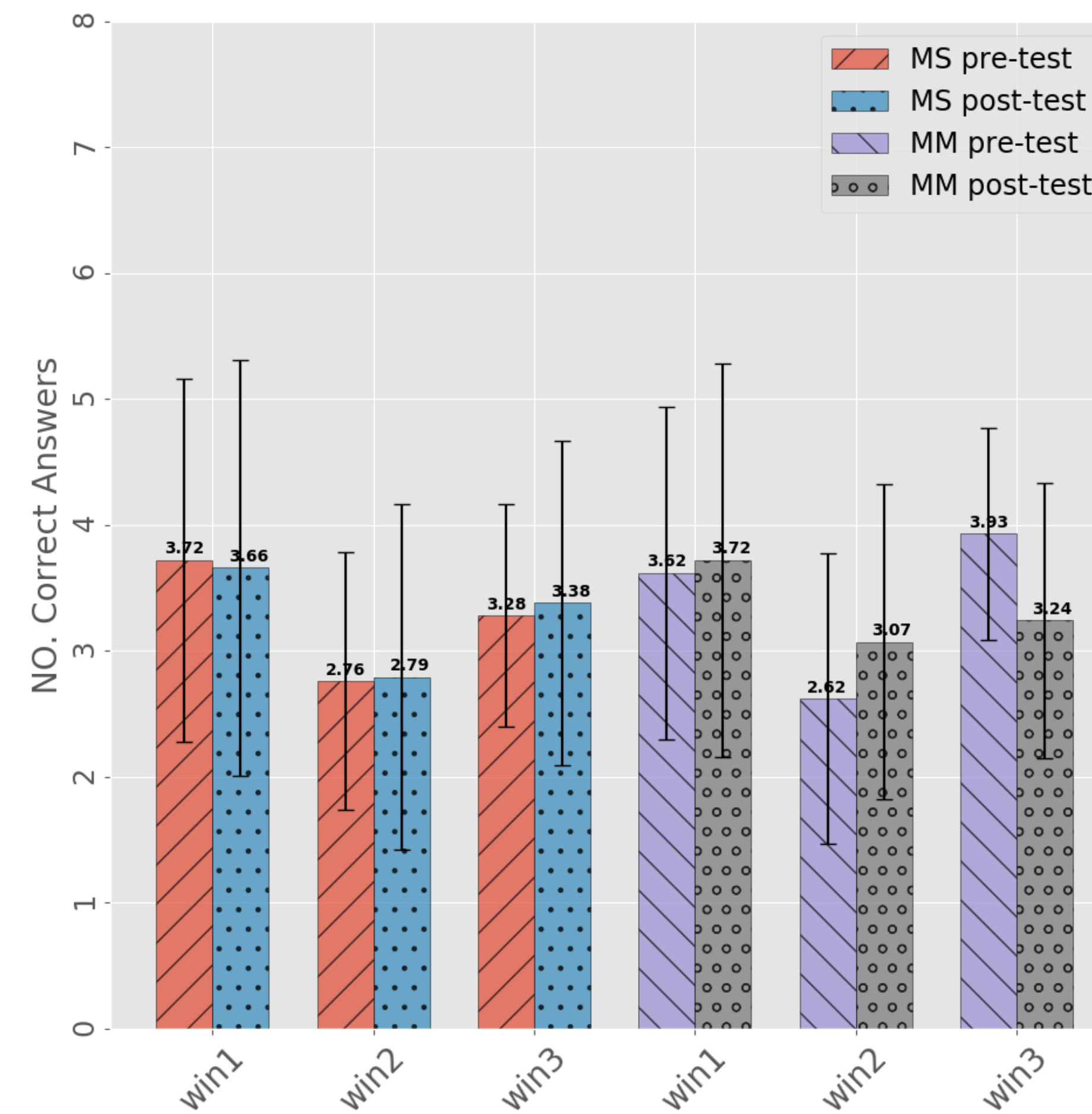
We estimate the mental execution complexity of a query by new variants of Kolmogorov complexity, **cognitive cost** of datalog program Cog and problem solution $CogP$. We hypothesise a bound on human hypothesis space size B and postulated a **cognitive window** which includes two constraints:

1. $E_{ex}(D, H, M(E)) < 0$ if $|S| > B(M(E), H)$
2. $E_{ex}(D, H, M(E)) \leq 0$ if $Cog(M(E), x) \geq CogP(E, \bar{M}, \phi, x)$

Rule verbalisation utilises declarative memory, and an increase in computational complexity and working memory corresponds to a negative effect on human performance.

Results

MS and MM denote human self-learning and machine-aided learning.



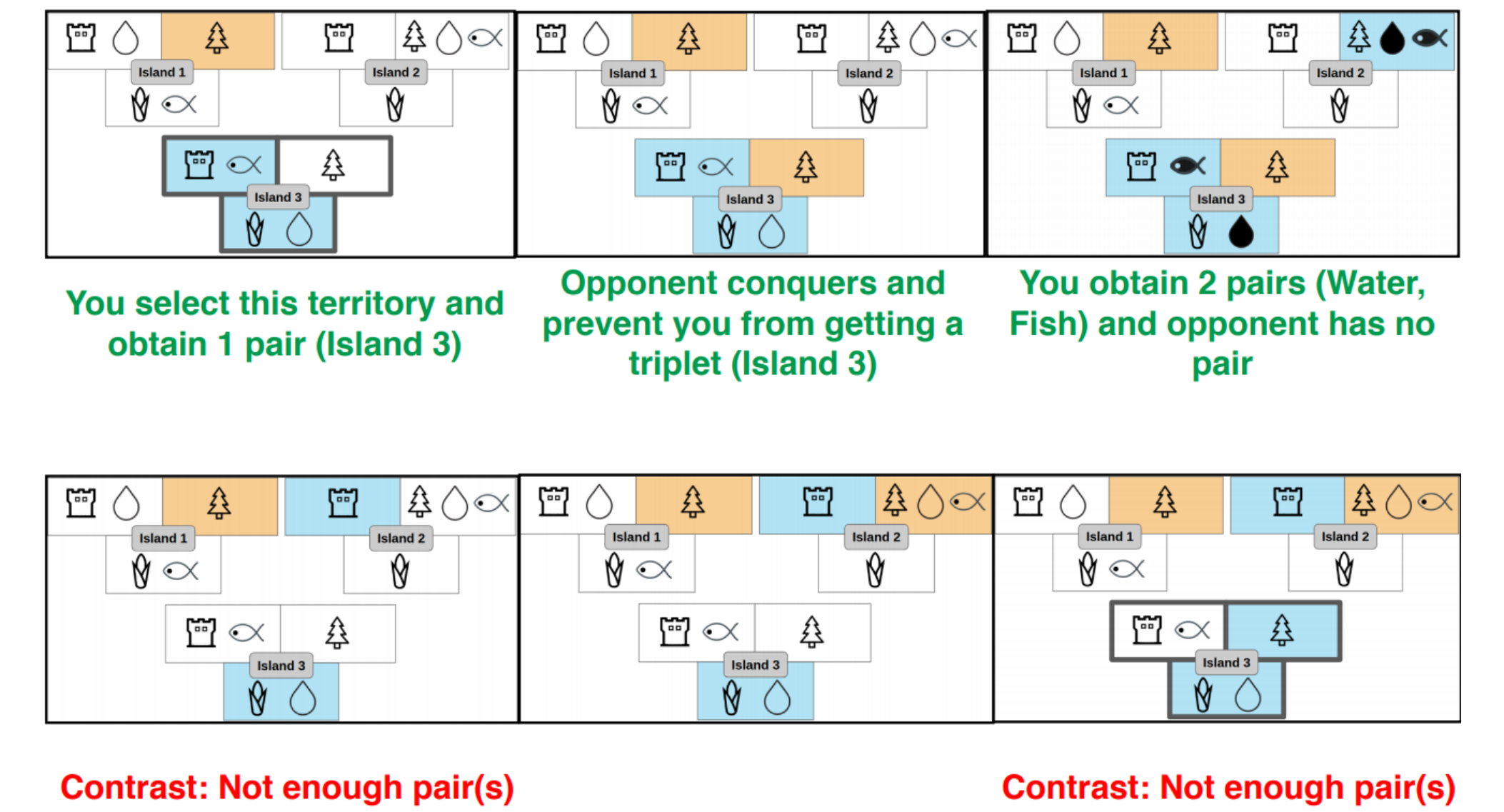
- **win1**: violates the cognitive cost constraint and $E_{ex} = 0$
- **win2**: does not violate cognitive window constraints and $E_{ex} > 0$
- **win3**: violates the hypothesis space size constraint, could only learn a maximum of four clauses and $E_{ex} < 0$

Materials

A MIL system **MIPlain** learns a complete and consistent logic program (below) for tasks win_1 , win_2 and win_3 which are Noughts and Crosses positions with increasing minimax search depth.

Depth	Rules
1	win_1(A,B) :- move(A,B), won(B)
2	win_2(A,B) :- move(A,B), win_2_1(B)
	win_2_1(A) :- number_of_pairs(A,x,2), number_of_pairs(A,o,0)
3	win_3(A,B) :- move(A,B), win_3_1(B)
	win_3_1(A) :- number_of_pairs(A,x,1), win_3_2(A)
	win_3_2(A) :- move(A,B), win_3_3(B)
	win_3_3(A) :- number_of_pairs(A,x,0), win_3_4(A)
	win_3_4(A) :- win_2(A,B), win_2_1(B)

MIPlain which is a variant of MIL game learning framework **MIGO** learns a winning Noughts and Crosses strategy. Textual explanations are translated from the program above. An example of visual and textual explanations is presented below.



Future and ongoing works

- More interactive human-machine explanatory teaching
- Teaching explanations from stochastic logic programs
- Improving explanatory beneficiality via sequential teaching
- Behavioural debugging of human errors by ILP