Dagstuhl Seminar Approaches and Applications of Inductive Programming 2023

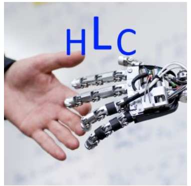
Effects of explaining machine-learned logic programs for human comprehension and discovery

Lun Ai
Department of Computing
Imperial College London

Lun Ai

Imperial College London

Doctoral researcher at Imperial College London, UK



Interests:

- **Effects of AI explanations** in human-AI collaboration
- **Sustainable** and **user-friendly** AI to drive science

TAILOR

AI-4-EB Consortium

Supervisor: Stephen Muggleton

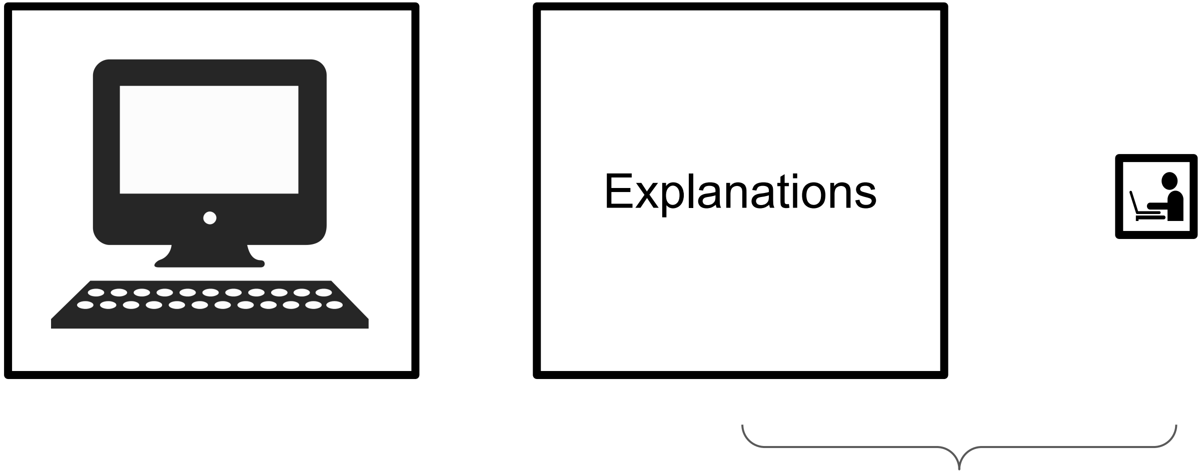
- Inductive & Abductive Logic Programming
- Explainable AI
- Computational Scientific Discovery (Biology)

Explainable AI (XAI) is necessary

[Gunning and Ada, 2019; Miller, 2019; Markus et al. 2021; Minh et al., 2022; Krenn et al. 2022; Schmid and Wrede, 2022; Adamson, 2022]

… for our interactions with AI

Do users actually understand AI explanations?



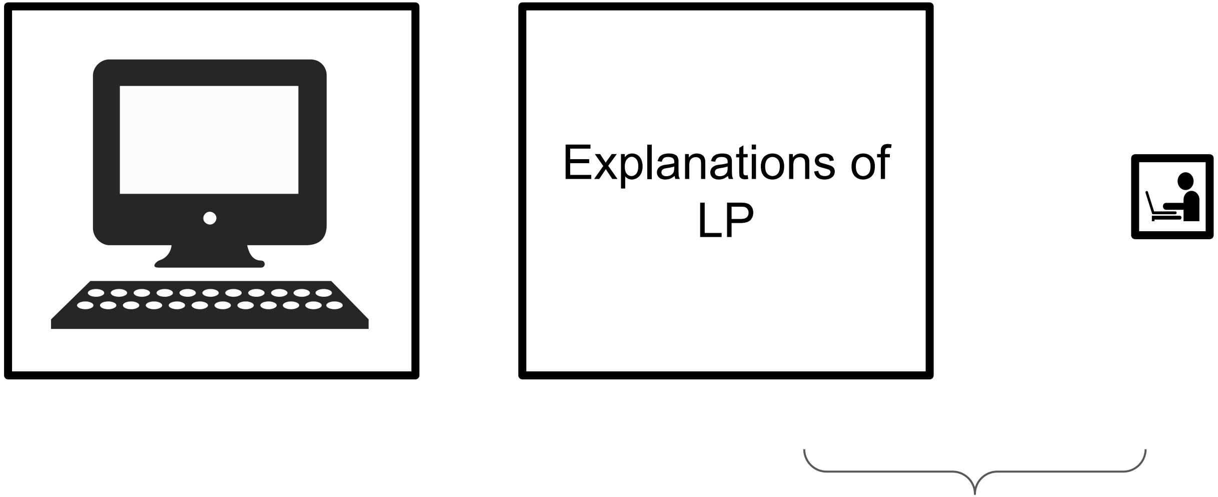
Explanations

Not quantifiable, e.g. interpretability [Lipton, 2018]

Comprehensibility = model complexity [Guidotti et al., 2018] ?

“Logic programs are human-understandable”

Problematic assumption



Explanations of LP

There are **very few** attempts to understand effects

Ultra-strong ML -> (beneficial) human behavioural change

Explanatory effect =

machine-aided task performance - self-learning task performance

Machine-aided: learning with explanations (e.g. generated from LP)

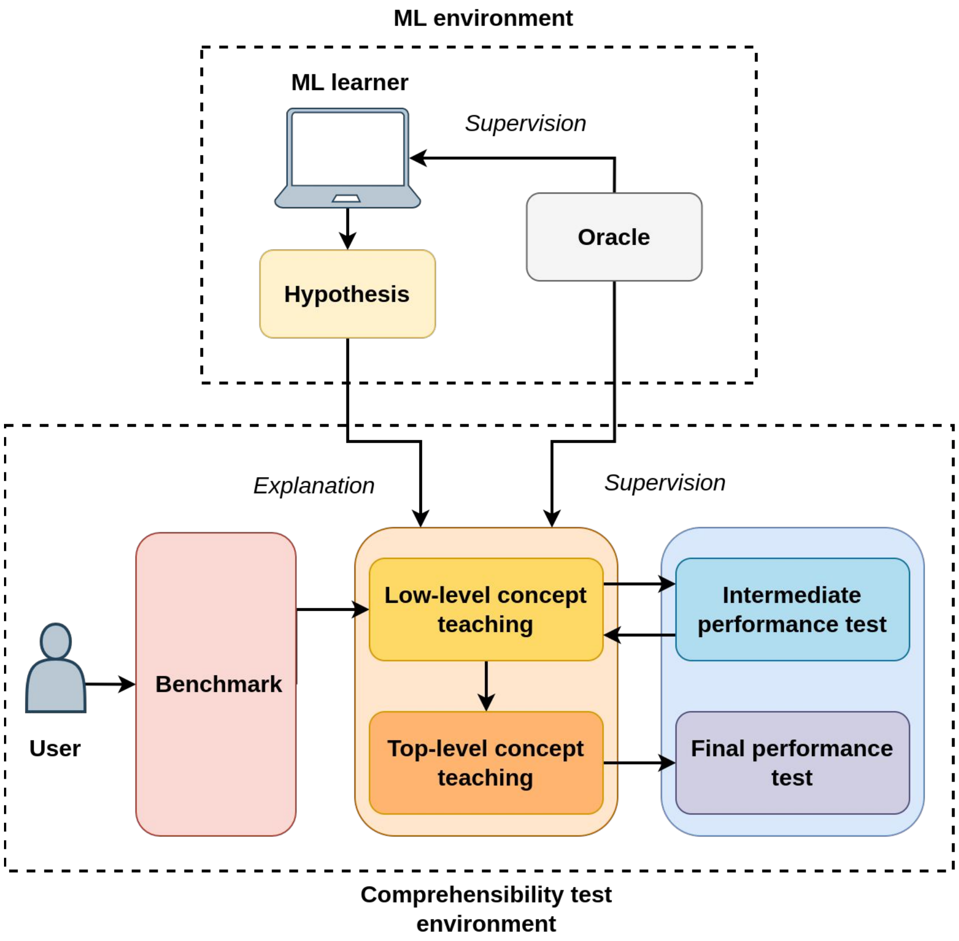
Self-learning: learning with only training examples

Performance: predictive accuracy on unseen test data

[Ai et al., 2021; Muggleton et al., 2018]

Teaching curriculum

ML:	teacher
Human:	student
Interactions:	curriculum



ML environment
ML learner
Supervision
Oracle
Hypothesis
Explanation
Supervision
User
Benchmark
Low-level concept teaching
Top-level concept teaching
Intermediate performance test
Final performance test
Comprehensibility test environment

Humans are symbol manipulation systems

Cognitive window for a machine-learned logic program P**:**

Axiom 1: Hypothesis space to necessarily learn P must be small

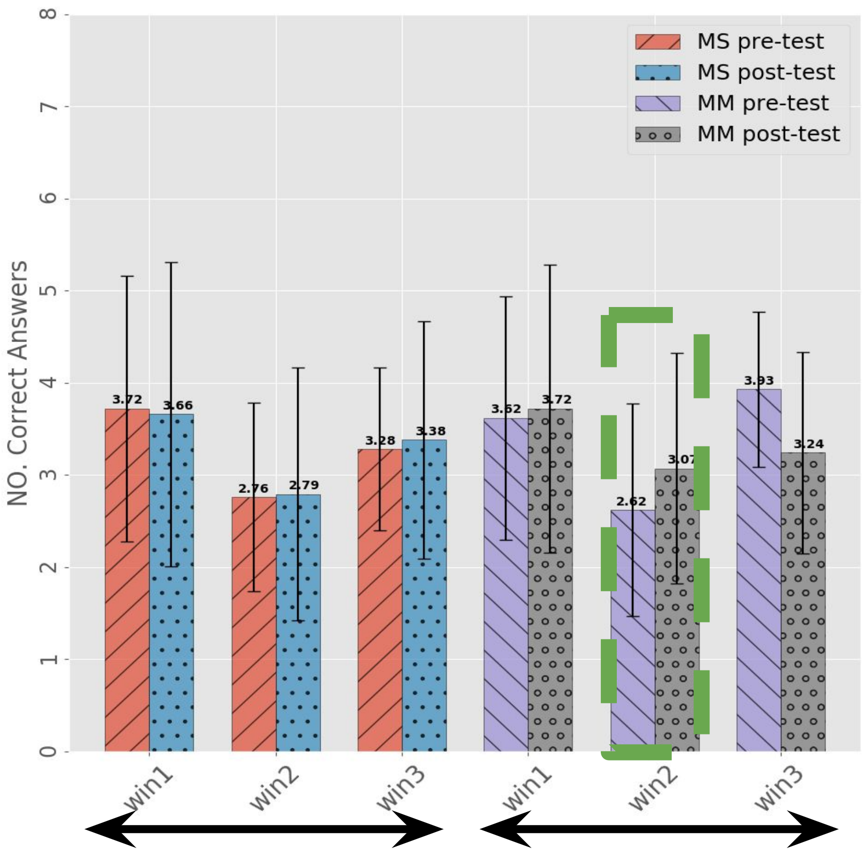
Humans have limited search ability in the hypothesis space

Axiom 2: Shortcuts in P to reduce grounding cost (cognitive cost)

Humans have limited capacity for mental computations

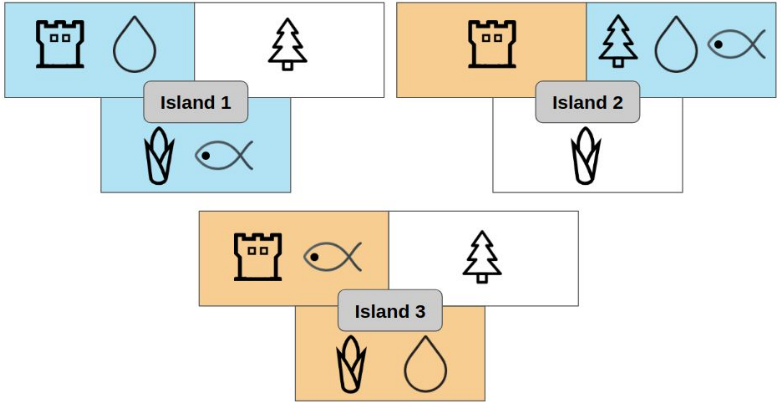
[Ai et al., 2021;Ai et al., 2023]

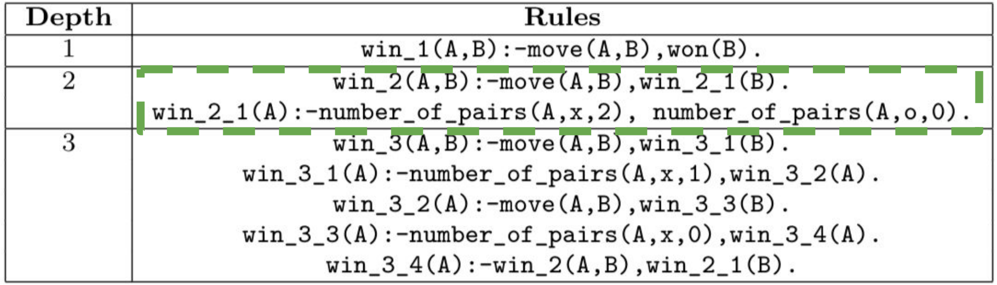
Cognitive window **satisfaction** = **beneficial** effect



Without explanations

With explanations





Depth	Rules
1	`win_1(A,B):-move(A,B),won(B).`

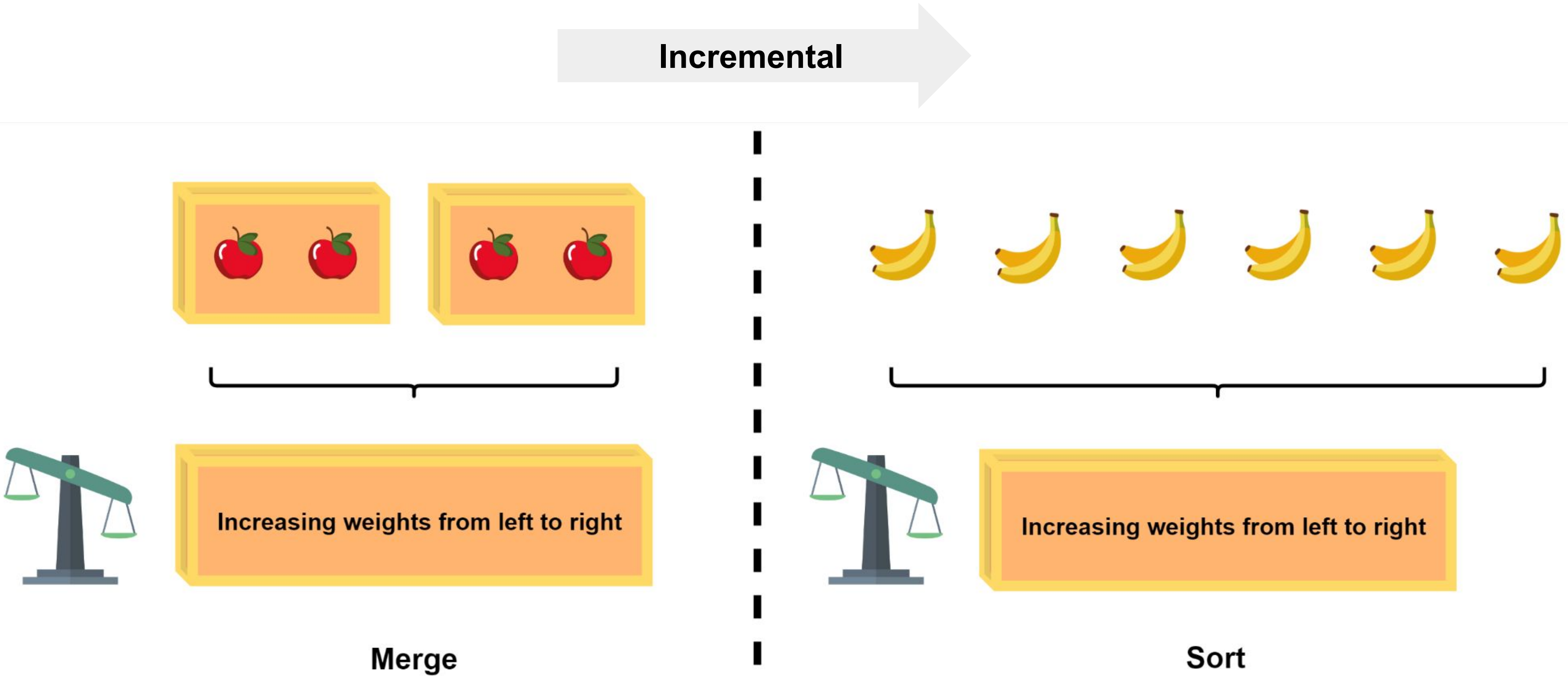
2	`win_2(A,B):-move(A,B),win_2_1(B).`
	`win_2_1(A):-number_of_pairs(A,x,2), number_of_pairs(A,o,0).`

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).`

Learned by Metagol

[Ai et al., 2021]

Teach **Merge Sort** to human novices



Incremental

Increasing weights from left to right

Merge

Increasing weights from left to right

Sort

[Ai et al., 2023]

A variant of **bottom-up** merge sort [Goldstine & Neumann, 1963]

Input:

[4, 6, 5, 2, 3, 1]

After Iteration 1

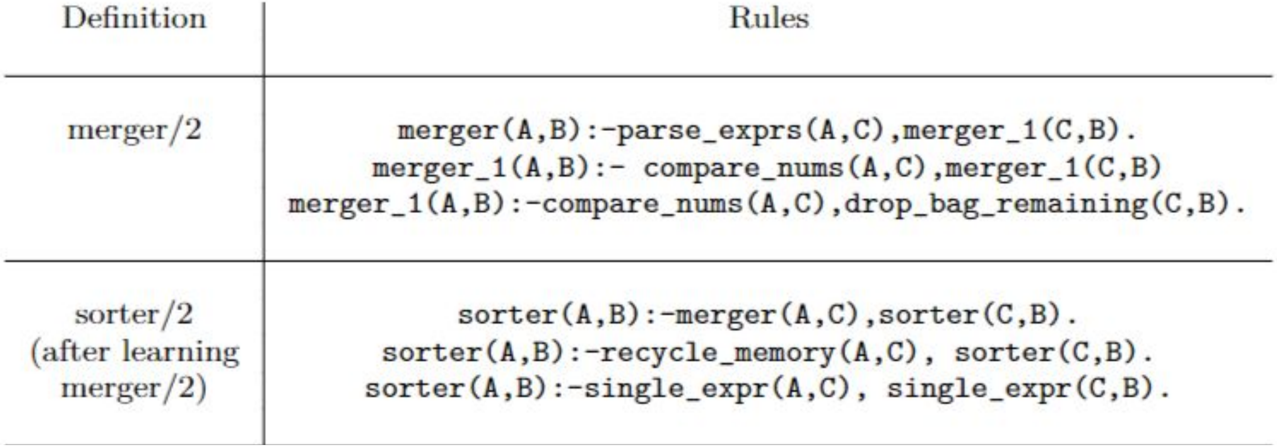
[4 < 6, 2 < 5, 1 < 3]

After Iteration 2

[2 < 4 < 5 < 6, 1 < 3]

After Iteration 3

[1 < 2 < 3 < 4 < 5 < 6]

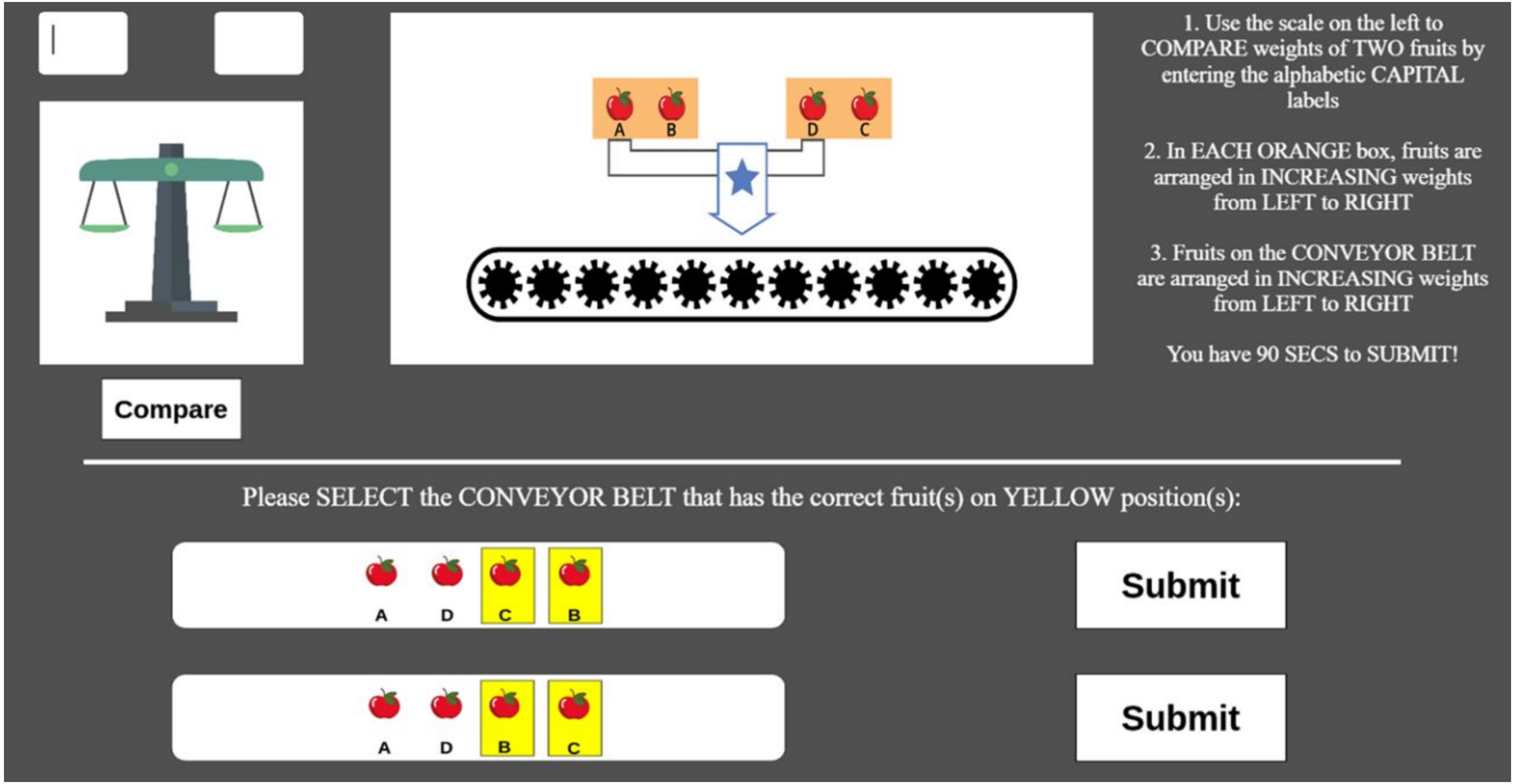


Definition	Rules
merger/2	merger(A,B):-parse_exprs(A,C),merger_1(C,B). merger_1(A,B):- compare_nums(A,C),merger_1(C,B) merger_1(A,B):-compare_nums(A,C),drop_bag_remaining(C,B).
sorter/2 (after learning merger/2)	sorter(A,B):-merger(A,C),sorter(C,B). sorter(A,B):-recycle_memory(A,C), sorter(C,B). sorter(A,B):-single_expr(A,C), single_expr(C,B).

Learned by Metagol

[Ai et al., 2023]

Learn to merge



1. Use the scale on the left to COMPARE weights of TWO fruits by entering the alphabetic CAPITAL labels

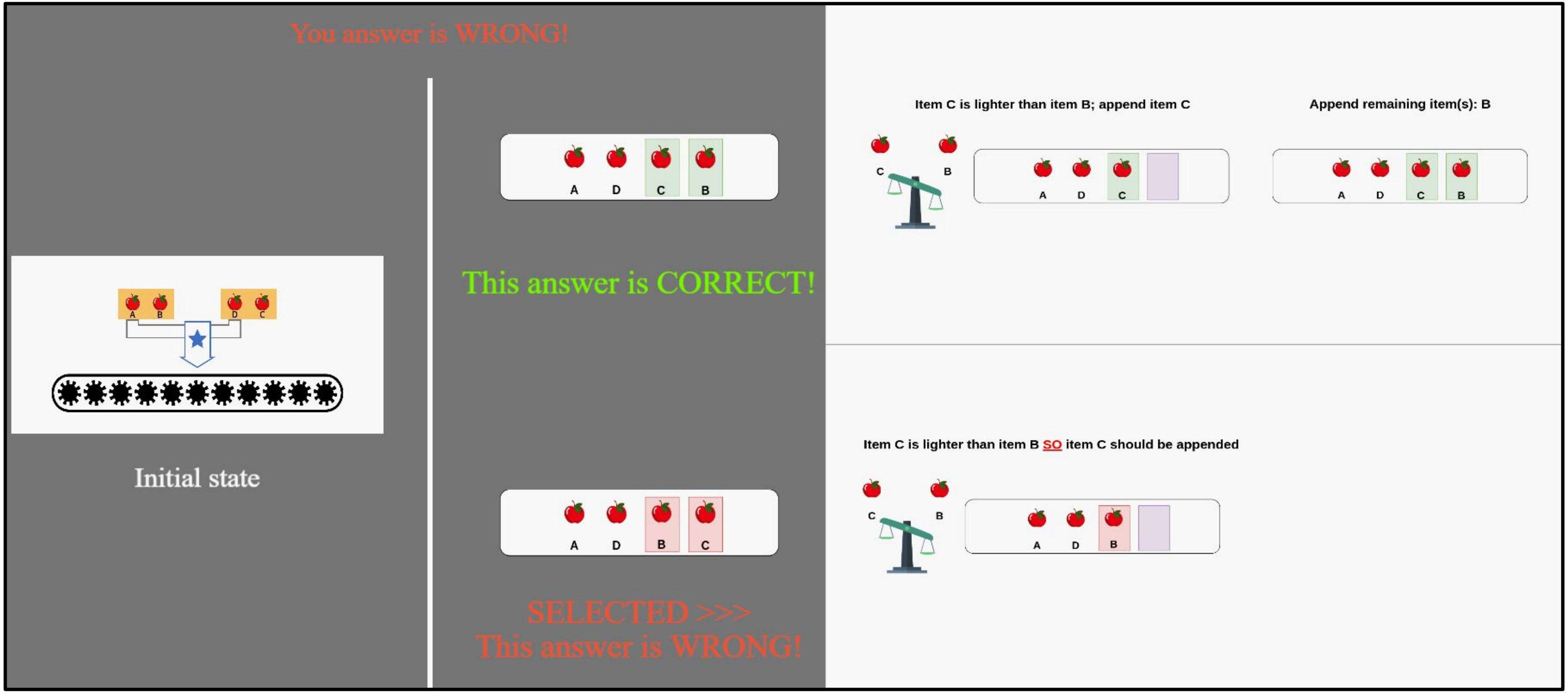
2. In EACH ORANGE box, fruits are arranged in INCREASING weights from LEFT to RIGHT

3. Fruits on the CONVEYOR BELT are arranged in INCREASING weights from LEFT to RIGHT

You have 90 SECS to SUBMIT!

Please SELECT the CONVEYOR BELT that has the correct fruit(s) on YELLOW position(s):

Why is/isn’t an action optimal?



You answer is WRONG!

This answer is CORRECT!

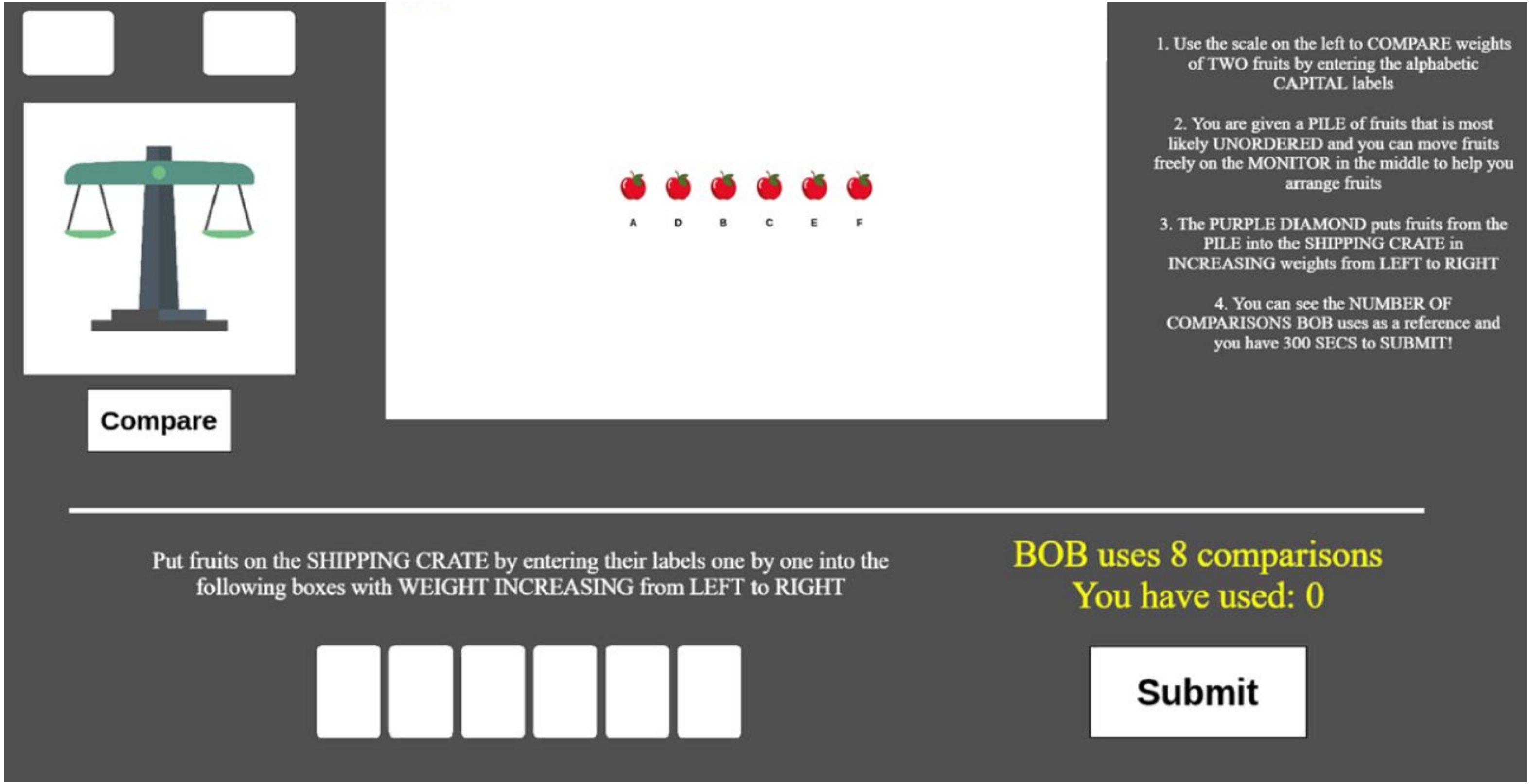
SELECTED >>>
This answer is WRONG!

Item C is lighter than item B; append item C

Append remaining item(s): B

Item C is lighter than item B SO item C should be appended

Learn to sort



1. Use the scale on the left to COMPARE weights of TWO fruits by entering the alphabetic CAPITAL labels

2. You are given a PILE of fruits that is most likely UNORDERED and you can move fruits freely on the MONITOR in the middle to help you arrange fruits

3. The PURPLE DIAMOND puts fruits from the PILE into the SHIPPING CRATE in INCREASING weights from LEFT to RIGHT

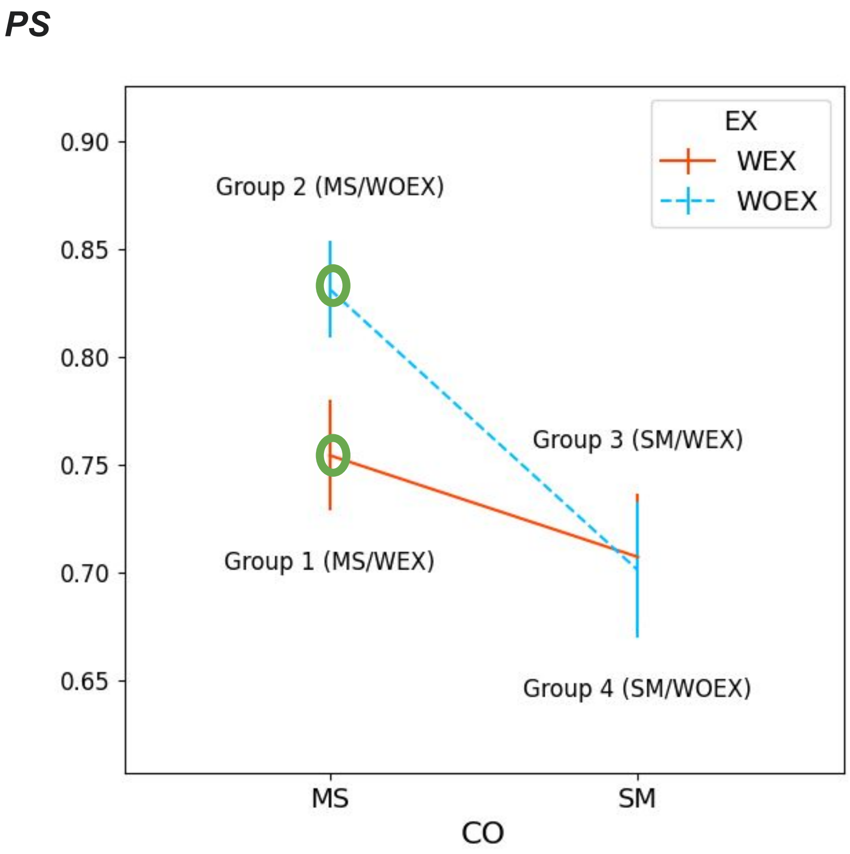
4. You can see the NUMBER OF COMPARISONS BOB uses as a reference and you have 300 SECS to SUBMIT!

Put fruits on the SHIPPING CRATE by entering their labels one by one into the following boxes with WEIGHT INCREASING from LEFT to RIGHT

BOB uses 8 comparisons
You have used: 0

Incremental curriculum: improved human performance

Average sorting performance ***PS***:
Monotonic correlation of target vs. human answers (Spearman rank)



An alternative evaluation? An example.

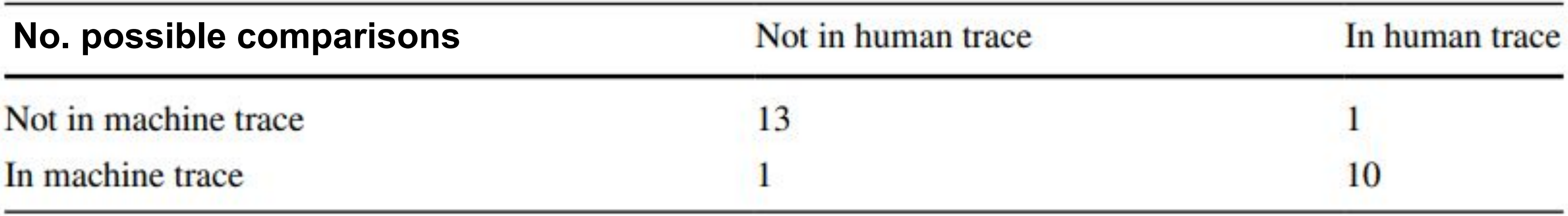
Sequence [4, 6, 5, 2, 3, 1]

Human trace

[(6, 4), (5, 2), (3, 1), (4, 2), (5, 4), (6, 5), (2, 1), (3, 2), (4, 3)]

Machine trace (24 algorithms, 6 categories)

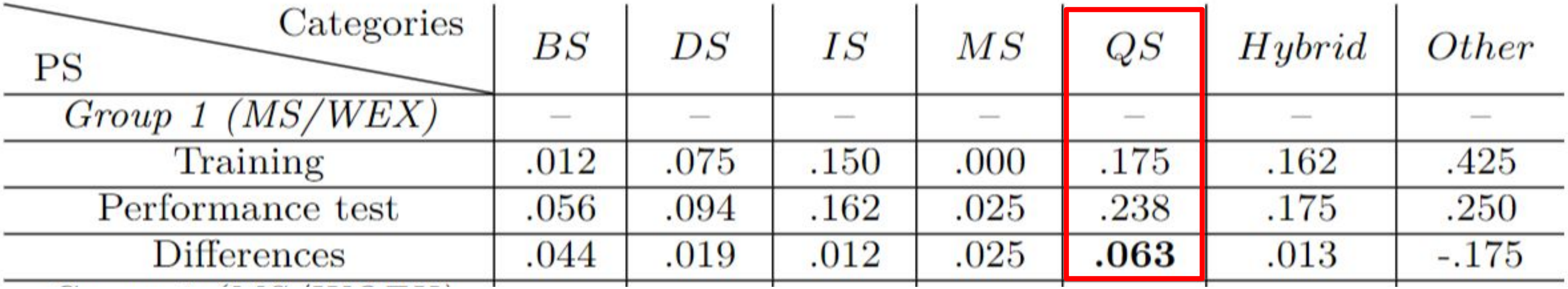
[(4, 6), (5, 2), (2, 4), (4, 5), (5, 6), (3, 1), (1, 2), (2, 3), (3, 4)]



No. possible comparisons	Not in human trace	In human trace
Not in machine trace	13	1
In machine trace	1	10

$\boldsymbol{\chi}^2$ = 14.3 with $p < .001$ and Spearman rank correlation $\boldsymbol{\rho}$ =.9 and $p < .001$

Novel strategy adaptation: quick sort



Categories PS	*BS*	*DS*	*IS*	*MS*	*QS*	*Hybrid*	*Other*
Group 1 (MS/WEX)	–	–	–	–	–	–	–
Training	.012	.075	.150	.000	.175	.162	.425
Performance test	.056	.094	.162	.025	.238	.175	.250
Differences	.044	.019	.012	.025	**.063**	.013	-.175

Unexpected efficient strategy with **better** performance
(incremental learning with explanations)

Messages & Open questions

1. LPs are not always human-comprehensible
    - How do we optimise **comprehensibility**
    - Is possible to formulate a **theory** of incomprehensibility?
2. We can learn a lot by studying effects of LP explanations
    - What insights can we get from **human trace** and **ILP learner trace**?
    - How can we **design curricula** to enable human discovery?
3. There are **limitations** to performance-based evaluations
    - How should we evaluate **strategy adaptations**?

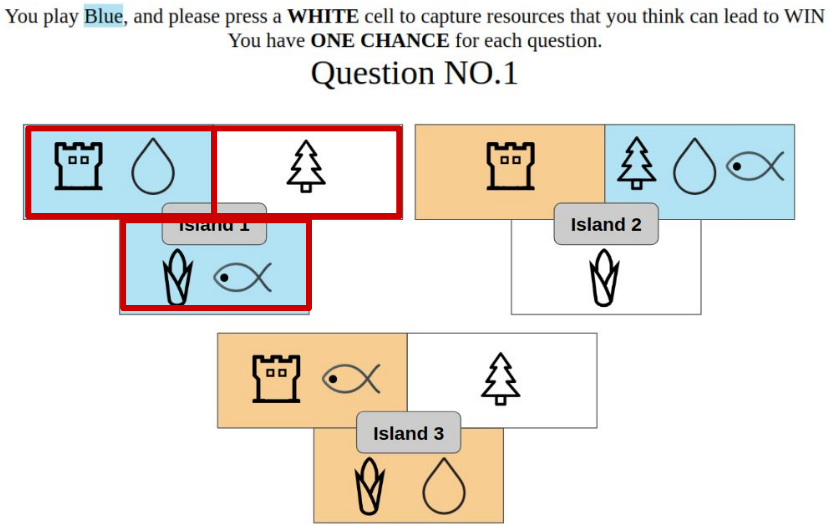
Lun Ai

Email: lun.ai15@imperial.ac.uk

Website: https://lai1997.github.io/

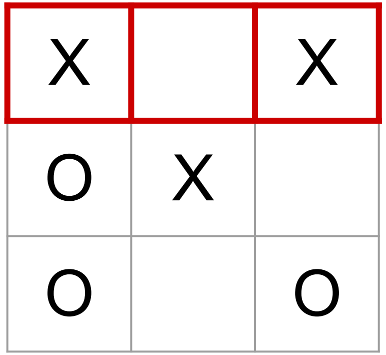
Linkedin: https://www.linkedin.com/in/lun-ai-46481a128/

Isomorphism of Noughts and Crosses



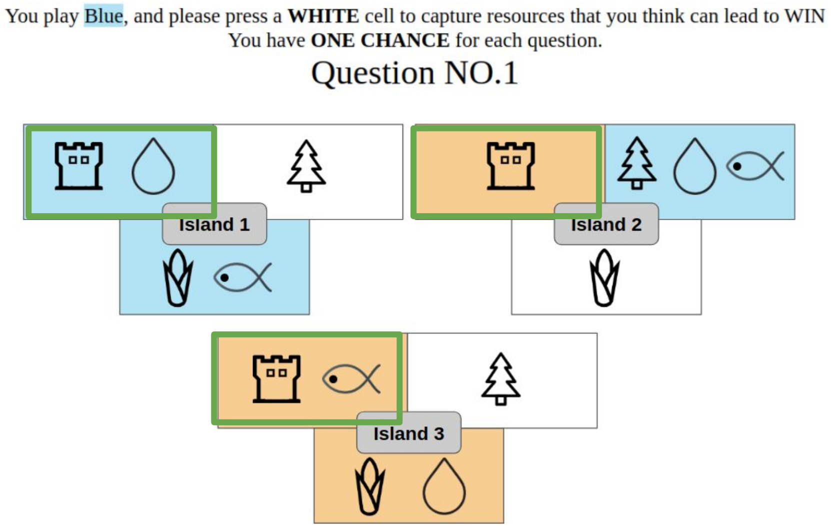
You play Blue, and please press a **WHITE** cell to capture resources that you think can lead to WIN
You have **ONE CHANCE** for each question.

Question NO.1



X		X
O	X	
O		O

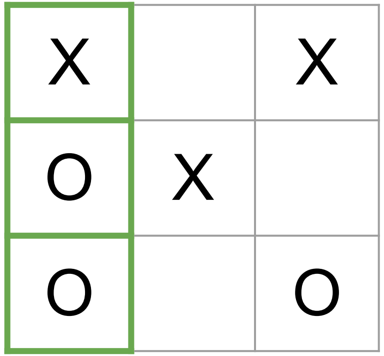
Isomorphism of Noughts and Crosses



You play Blue, and please press a **WHITE** cell to capture resources that you think can lead to WIN

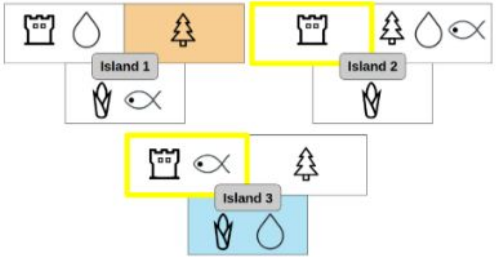
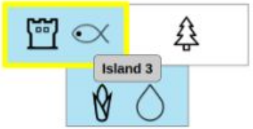
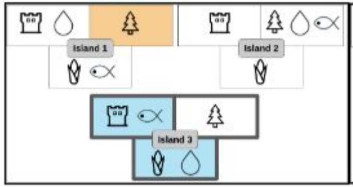
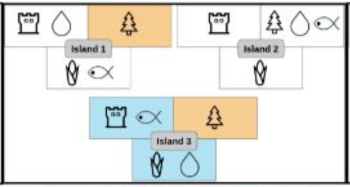
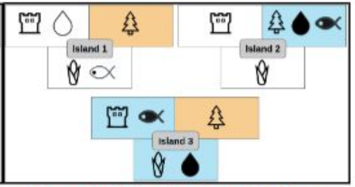
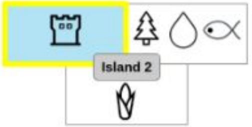
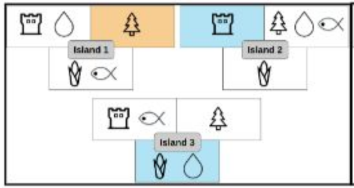
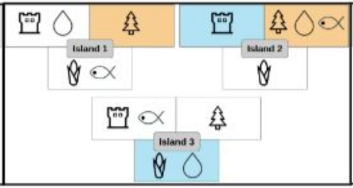
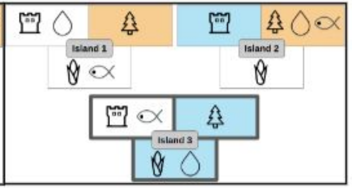
You have **ONE CHANCE** for each question.

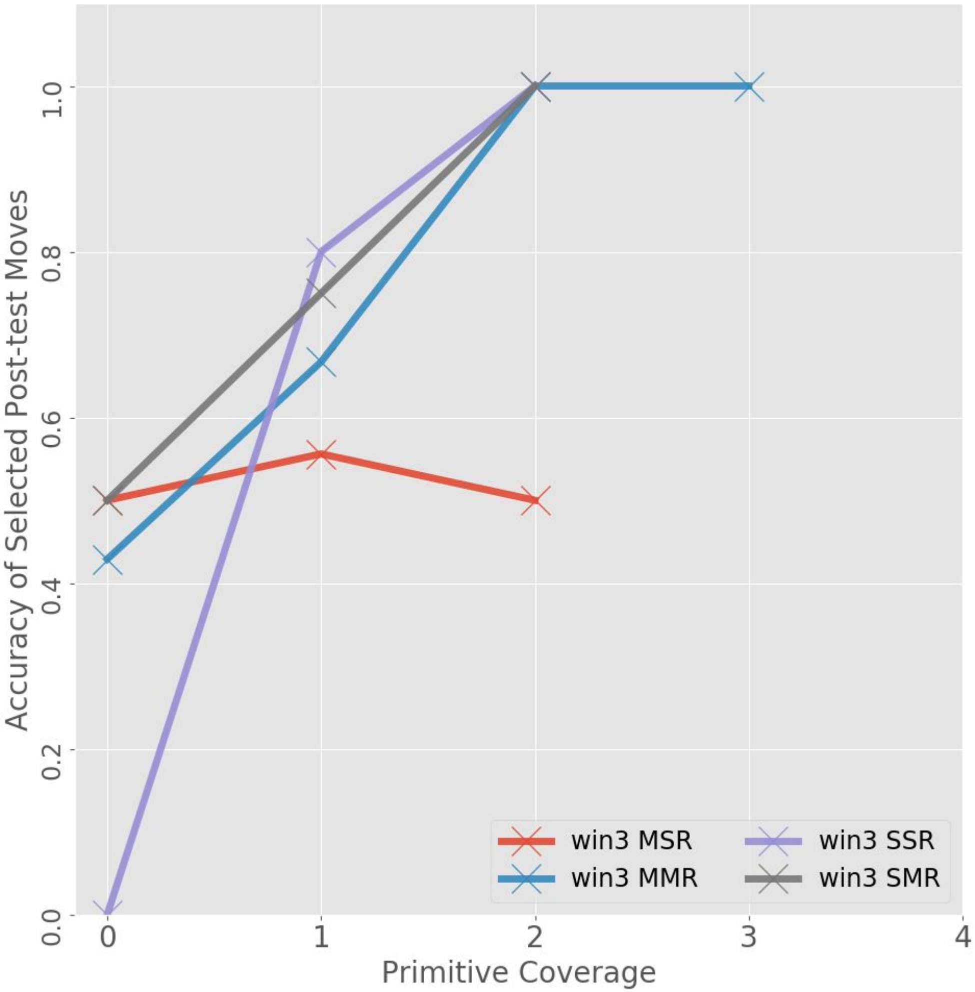
Question NO.1



X		X
O	X	
O		O

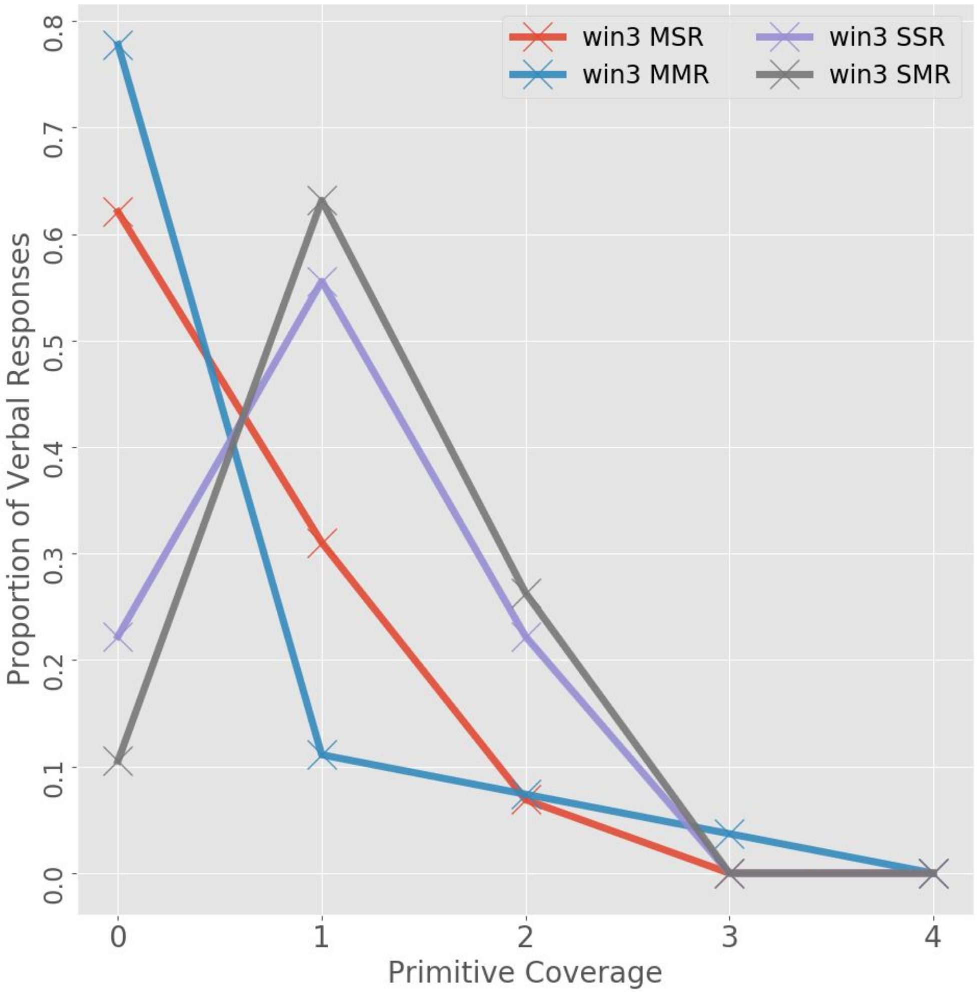
Learn the Island Game (Noughts and Crosses isomorphism)

Example	Moves	MIPlain's comments		
	 This is a right move	 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
	 This is a wrong move	 Contrast: Not enough pair(s)		 Contrast: Not enough pair(s)



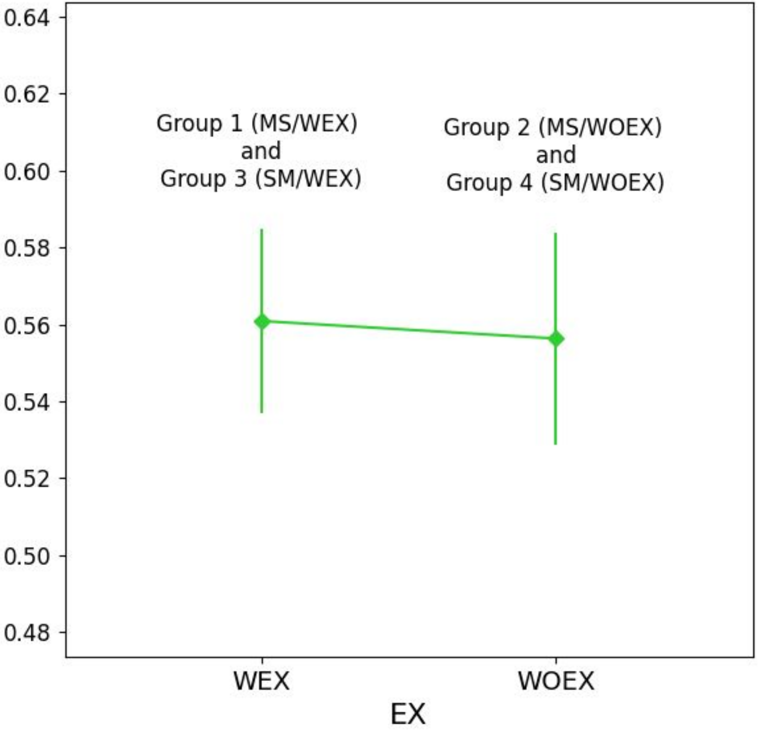
Primitive coverage: No. descriptions of primitives in textual responses

High correlation with performance

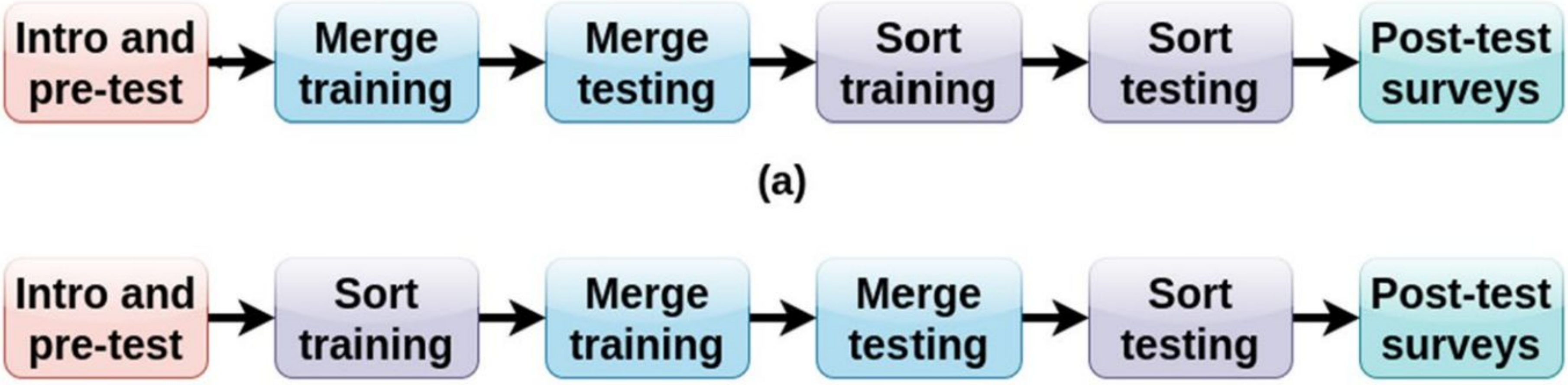


Low frequency of high coverage (key predicates) responses

No performance change by explaining merge



Curriculum arrangements



Intro and pre-test → Merge training → Merge testing → Sort training → Sort testing → Post-test surveys

(a)

Intro and pre-test → Sort training → Merge training → Merge testing → Sort testing → Post-test surveys

(b)