Machine learning manuscript No.
(will be inserted by the editor)

Beneficial and Harmful Explanatory Machine Learning

Lun Ai · Stephen H. Muggleton · Céline Hocquette · Mark Gromowski · Ute Schmid

Received: date / Accepted: date

Abstract Given the recent successes of Deep Learning in AI there has been increased interest in the role and need for explanations in machine learned theories. A distinct notion in this context is that of Michie's definition of Ultra-Strong Machine Learning (USML). USML is demonstrated by a measurable increase in human performance of a task following provision to the human of a symbolic machine learned theory for task performance. A recent paper demonstrates the beneficial effect of a machine learned logic theory for a classification task, yet no existing work has examined the potential harmfulness of machine's involvement in human learning. This paper investigates the explanatory effects of a machine learned theory in the context of simple two person games and proposes a framework for identifying the harmfulness of machine explanations based on the Cognitive Science literature. The approach involves a cognitive window consisting of two quantifiable bounds and it is supported by empirical evidence collected from human trials. Our quantitative and qualitative results indicate that human learning aided by a symbolic machine learned theory which satisfies a *cognitive window* has achieved significantly higher performance than human self learning. Results also demonstrate that human learning aided by a symbolic machine learned theory that fails to satisfy this window leads to significantly worse performance than unaided human learning.

Lun Ai
Department of Computing, Imperial College London, London, UK
E-mail: lun.ai15@imperial.ac.uk

Stephen H. Muggleton
Department of Computing, Imperial College London, London, UK
E-mail: s.muggleton@imperial.ac.uk

Céline Hocquette
Department of Computing, Imperial College London, London, UK
E-mail: celine.hocquette16@imperial.ac.uk

Mark Gromowski
Cognitive Systems Group, University of Bamberg, Bamberg, Germany
E-mail: mark.gromowski@uni-bamberg.de

Ute Schmid
Cognitive Systems Group, University of Bamberg, Bamberg, Germany
E-mail: ute.schmid@uni-bamberg.de

1 Introduction

In a recent paper [34] the authors provided an operational definition for comprehensibility of logic programs and used this, in experiments with humans, to provide the first demonstration of Michie's *Ultra-Strong Machine Learning* (USML). The authors demonstrated USML via empirical evidence that humans improve out-of-sample performance in concept learning from a training set E when presented with a first-order logic theory which has been machine learned from E. The improvement of human performance indicates a beneficial effect of comprehensible machine learned models on human skill acquisition. The present paper investigates the explanatory effects of machine's involvement in human skill acquisition of simple games. Our results indicate that when a machine learned theory is used to teach strategies to humans, in some cases the human's out-of-sample performance is reduced. This degradation of human performance is recognised to indicate the existence of harmful explanations.

In the current paper, which extends our previous work on the phenomenon of USML, both beneficial and harmful effects of a machine learned theory are explored in the context of simple games. Our definition of explanatory effects is based on human out-of-sample performance in the presence of natural language explanation generated from a machine learned theory (Figure 1). The analogy between understanding a logic program via declarative reading and understanding a piece of natural language text allows the explanatory effects of a machine learned theory to be investigated.

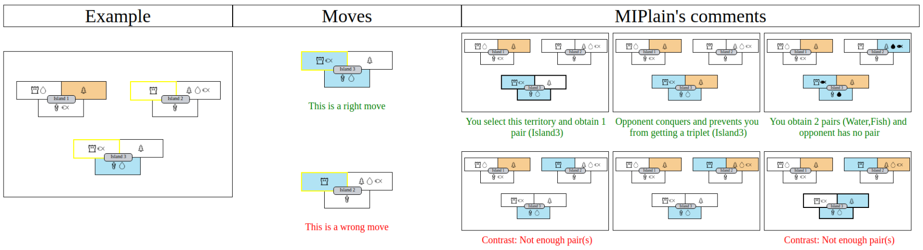

Example
Moves
MIPlain's comments
Island 1
Island 2
Island 3
This is a right move
This is a wrong move
You select this territory and obtain 1 pair (Island3)
Opponent conquers and prevents you from getting a triplet (Island3)
You obtain 2 pairs (Water,Fish) and opponent has no pair
Contrast: Not enough pair(s)
Contrast: Not enough pair(s)

Fig. 1: Textual and visual explanations are shown to treated participants along with a training example for winning a two player game isomorphic to Noughts and Crosses. Textual explanations were generated from the rules learned by our **M**eta-**I**nterpretive ex**Plain**able game learner *MIPlain*.

The results of relevant Cognitive Science literature allow the properties of a logic theory which are harmful to human comprehension to be characterised. Our approach is based on developing a framework describing a cognitive window which involves bounds with regard to 1) descriptive complexity of a theory and 2) execution stack requirements for knowledge application. We hypothesise that a machine learned theory provides a harmful explanation to humans when theory complexity is high and execution is cognitively challenging. Our proposed cognitive window model is confirmed by empirical evidence collected from multiple experiments involving human participants of various backgrounds.

We summarise our main contributions as follows:

- We define a measure to evaluate beneficial/harmful explanatory effects of machine learned theory on human comprehension.

- We develop a framework to assess a cognitive window of a machine learned theory. The approach encompasses theory complexity and the required execution stack.
- Our quantitative and qualitative analyses of the experimental results demonstrate that a machine learned theory has a harmful effect on human comprehension when its search space is too large for human knowledge acquisition and it fails to incorporate executional shortcuts.

This paper is arranged as follows. In Section 2, we discuss existing work relevant to the paper. The theoretical framework with relevant definitions is presented in Section 3. We describe our experimental framework and the experimental hypotheses in Section 4. Section 5 describes several experiments involving human participants on two simple games. We examine the impact of a cognitive window on the explanatory effects of a machine learned theory based on human performance and verbal input. In Section 6, we conclude our work and comment on our analytical results – only a short and simple-to-execute theory can have a beneficial effect on human comprehension. We discuss potential extensions to the current framework, curriculum learning and behavioural cloning, for enhancing explanatory effects of a machine learned theory.

2 Related work

This section summarises related research of game learning and familiarises the reader with the core motivations for our work. We first present a short overview of related investigations in explanatory machine learning of games. Subsequently, we cover various approaches for teaching and learning between humans and machines.

2.1 Explanatory machine learning of games

Early approaches to learning game strategies [47, 41] used the decision tree learner ID3 to classify minimax depth-of-win for positions in chess end games. These approaches used carefully selected board attributes as features. However, chess experts had difficulty understanding the learned decision tree due to its high complexity [26]. Methods for simplifying decision trees without compromising their accuracy have been investigated [42] on the basis that simpler models are more comprehensible to humans. An early Inductive Logic Programming (ILP) [35] approach learned optimal chess endgame strategies at depth 0 or 1 [5]. An informal complexity constraint was applied which limits the number of clauses used in any predicate definition to 7 ± 2 clauses. This number is based on the hypothesised limit on human short term memory capacity of 7 ± 2 chunks [29]. A different approach involving the augmentation of training data with high-level annotations was explored in [18]. Initialisation requires explanations to be provided for the target data set and the predicative accuracy of explanations is evaluated similarly to the predicative accuracy of labels.

The earliest reinforcement learning system *MENACE* (Matchbox Educable Noughts And Crosses Engine) [25] was specifically designed to learn an optimal agent policy for Noughts and Crosses. Later, Q-Learning [54] and Deep Reinforcement Learning were spawned and have led to a variety of applications including the Atari 2600 games [33] and the game of Go [50]. While these systems defeated the strongest human players, they are not human-like since they lack the ability to explain the encoded knowledge to humans. Recent approaches such as [55] have aimed to explain the policies learned by these models, but the learned strategy is implicitly encoded into the continuous parameters of the policy function which makes their operation opaque to humans. Relational Reinforcement Learning [14] and Deep Relational Reinforcement Learning [56] have attempted to address these drawbacks by incorporating the use of relational biases to ensure human understandability.

In [30,31], the author provided a survey of most relevant work in explainable AI and argued that explanatory functionalities were mostly subjective to the developer's view. While there is a general lack of demonstration on explanatory effect which should be examined by empirical trials, no existing framework accounts for the explanatory harmfulness of machine learned models.

2.2 Two-way learning between human and machine

As an emerging sub-field of AI, Machine Teaching [16] provides an algorithmic model for quantifying the teaching effort and a framework for identifying an optimized teaching set of examples to allow maximum learning efficiency for the learner. The learner is usually a machine learning model of a human in a hypothesised setting. In education, machine teaching has been applied to devise intelligent tutoring systems to select examples for teaching [59,43]. On the other hand, rule-based logic theories are important mechanisms of explanation. Rule-based knowledge representations are generalised means of concept encoding and have a structure analogous to human conception. Mechanisms of logical reasoning, induction and abduction, have long been shown to be highly related to human concept attainment and information processing [23,19]. Additionally, humans' ability to apply recursion plays a key role in understanding of relational concepts and semantics of language [17] which are important for communication.

The process of reconstructing implicit target knowledge which is easy to operate but difficult to describe via machine learning has been explored under the topic of Behavioural Cloning. The cloning of human operation sequence has been applied in various domains such as piloting [28] and crane operation [53]. The cloned human knowledge and experience are more dependable and less error-prone due to perceptual and executional inconsistency being averaged across the original behavioural trace. To our knowledge, no existing work has attempted to estimate human errors and target these mistakes in interactive teaching sessions for achieving a measurable "clean up" effect [27] from machine explanations.

3 Theoretical framework

3.1 Meta-interpretive learning of simple games

Meta-Interpretive Learning (MIL) [37,38] is a sub-field of ILP which supports predicate invention, dependent learning [24], learning of recursions and higher-order programs. Given an input $(\mathcal{B}, \mathcal{M}, \mathcal{E}+, \mathcal{E}-)$ where the background knowledge $\mathcal{B}$ is a first-order logic program, meta-rules $\mathcal{M}$ are second-order clauses, positive examples $\mathcal{E}+$ and negative examples $\mathcal{E}-$ are ground atoms, a MIL algorithm returns a logic program hypothesis $\mathcal{H}$ such that $\mathcal{M} \cup \mathcal{H} \cup \mathcal{B} \models \mathcal{E}+$ and $\mathcal{M} \cup \mathcal{H} \cup \mathcal{B} \not\models \mathcal{E}-$. The meta-rules (for examples see Figure 3) contain existentially quantified second-order variables and universally quantified first-order variables. They clarify the declarative bias employed for substitutions of second-order Skolem constants. The resulting first-order theories are thus strictly logical generalisation of the meta-rules.

Table 1: A set of win rules is learned by *MIGO*. *MIGO*'s background knowledge contains a general move generator *move/2* and a won classifier *won/1* to encode the minimum rules of the game. The program is dyadic and $win\_1/2$ can be reduced to $win\_1(A, B) : -move(A, B), won(B)$ by removing literals after unfolding.

Depth	**Rules**

1	`win_1(A,B):- win_1_1_1(A,B),won(B).`
	`win_1_1_1(A,B):-move(A,B),won(B).`

2	`win_2(A,B):-win_2_1_1(A,B),not(win_2_1_1(B,C)).`
	`win_2_1_1(A,B):-move(A,B), not(win_1(B,C)).`

3	`win_3(A,B):-win_3_1_1(A,B),not(win_3_1_1(B,C)).`
	`win_3_1_1(A,B):-win_2_1_1(A,B), not(win_2(B,C)).`

Table 2: The logic program learned by *MIPlain* represents a strategy for the first player to win at different depths of the game. The predicate $win\_3\_4/1$ can be reduced to $win\_3\_4(A) : -win\_2(A, B)$ by removing literals after unfolding.

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).`

The MIL game learning framework *MIGO* [36] is a purely symbolic system based on the adapted Prolog meta-interpreter Metagol [12]. *MIGO* learns exclusively from positive examples by playing against the optimal opponent. MIGO is provided with a set of three relational primitives, move/2, won/1, drawn/1 which are a move generator, a won and a drawn classifier respectively. These primitives represent the minimal information a human would expect to know before playing a two-person game. For Noughts and Crosses and Hexapawn, *MIGO* learns a rule-like symbolic game strategy (Table 1) that supports human understanding and was demonstrated to converge using less training data compared to Deep and classical Q-Learning. For successive values of k, MIGO learns a series of inter-related definitions for predicates $win\_k/2$. These predicates define maintenance of minimax win in k-ply.

We introduce *MIPlain*[1], a variant of *MIGO* which focuses on learning the task of winning for the game of Noughts and Crosses. In addition to learning from positive examples, *MIPlain* identifies moves which are negative examples for the task of winning. When a game is drawn or lost for the learner, the corresponding path in the game tree is saved for later backtracking following the most updated strategy. *MIPlain* performs a selection of hypotheses based on the efficiency of hypothesised programs using *Metaopt* [13].

An additional primitive $number\_of\_pairs/3$ is provided to *MIPlain* which depicts the number of pairs for a player (X or O) on a given board. A pair is the alignment of two marks of one player, the third square of this line being empty. An example of pairs is shown in Figure 2. This additional primitive serves as an executional shortcut that reduces the depth of the search when executing the learned

[1] MIPlain source is available at https://github.com/LAi1997/MIPlain

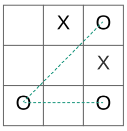


Fig. 2: O has two pairs represented in green and X has no pairs.

Meta-rule
$P(A, B) \leftarrow Q(A, B), R(B).$
$P(A) \leftarrow Q(A, B), R(B).$
$P(A) \leftarrow Q(A, S, T), R(A).$
$P(A) \leftarrow Q(A, S, T), R(A, U, V).$

Fig. 3: Letters P, Q, R, S, T, U, V denote existentially quantified second-order variables and A, B, C are universally quantified first-order variables.

strategy. Furthermore, $MIPlain$ is given the meta-rules described in Figure 3, which are two variants of the *postcon* meta-rule with monadic or dyadic head, and two variants of the *conjunction* meta-rule with currying in either the first or both body literals. Currying allows the learning of programs with higher-arity predicates where existentially quantified argument variables are bound to constants. The learned strategy presented in Table 2 describes conditions in a rule-like manner that the player's optimal move has to satisfy.

3.2 Explanatory effectiveness of a machine learned theory

We extend the machine-aided human comprehension of examples in [34] and $C(D, H, E)$ denotes the unaided human comprehension of examples where D is a target definition, H is a group of humans and E is a set of examples. Based on the analogy between declarative understanding of a logic program and understanding of a natural language explanation, we describe measures for estimating the degree to which the output of a symbolic machine learning algorithm as an explanation can aid human comprehension.

Definition 1 (Machine-explained human comprehension of examples, $C_{ex}(D, H, M(E))$): Given a definition D, a group of humans H, a theory $M(E)$ learned using machine learning algorithm M and examples E, the machine-explained human comprehension of examples E is the mean accuracy with which a human $h \in H$ after brief study of an explanation based on $M(E)$ can classify new material selected from the domain of D.

Definition 2 (Explanatory effect of a machine learned theory, $E_{ex}(D, H, M(E))$): Given a definition D, a group of humans H, a symbolic machine learning algorithm M, the explanatory effect of the theory $M(E)$ learned from examples E is

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

Definition 3 (Beneficial/harmful effect of a machine learned theory): Given a definition D, a group of humans H, a symbolic machine learning algorithm M:

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

In the scope of this work, we relate the explanatory effectiveness of a theory to performance which means that a harmful explanation provided by the machine degrades comprehension of the task and therefore reduces performance.

3.3 Cognitive window of a machine learned theory

In this section, we suggest a window of a machine learned theory that constraints its explanatory effectiveness. A basic assumption of cognitive psychology and artificial intelligence is that human information processing can be modelled in analogy to symbol manipulation of computers – respectively its formal characterisation of a Turing Machine [29, 21, 39]. More specifically, computational models of cognition share the view that intelligent action is based on manipulation of representations in working memory. In consequence, human inferential reasoning is limited by working memory capacity which corresponds to limitations of tape length and instruction complexity in Turing Machines.

Besides general restrictions of human information processing, performance can be influenced by internal or environmental disruptions such that the given competencies of a human in a specific domain are not always reflected in observable actions [11, 49]. However, it can be assumed that humans – at least in domains of higher cognition – are able to explain their actions by verbalising the rules which they applied to produce a given result [46]. Although rules in general can be classified as procedural knowledge, the ability to verbalise rules makes them part of declarative memory [3, 46]. For complex domains, the rules which govern action generation will typically be computationally complex as measured by the Kolmogorov complexity [22]. One can assume that increase in complexity can have a negative effect on performance.

In language processing and in general problem solving, hierarchisation of complex action sequences can make information processing more efficient. Typically, a general goal is broken down into sub-goals as it has been proposed in production system models [39] as well as in the context of analogical problem solving [9]. Rules which guide problem solving behaviour, for instance in puzzles such as Tower of Hanoi or games such as Noughts and Crosses, might be learned. From a declarative perspective, such learned rules correspond to explicit representations of a concept such as the win-in-two-steps move introduced above. Studies of rule-based concept acquisition suggest that human concept learning can be characterised as search in a pool of possible hypotheses which are explored in some order of preference [8]. This observation relates to the concept of version space learning introduced in machine learning [32].

Based on these different observations concerning human information processing, we propose that a) human learners are version space learners with limited hypothesis space search capability that use meta-rules to learn sub-goal structure and primitives as background knowledge. This allows us to compute a bound on the human hypothesis space size based on the MIL complexity analysis in [24]. We assume that b) rules can be represented explicitly in a declarative, verbalisable form. Finally, we postulate the existence of a cognitive window such that a machine learned theory can be an effective explanation if it satisfies two constraints: 1) a hypothesised human learning procedure which has a limited search space and 2) a knowledge application model based on the Kolmogorov complexity [22]. For the following definitions, we restrict ourselves to learning datalog programs which do not include function symbols.

Definition 4 (Cognitive bound on the hypothesis space size, $B(P, H)$): Consider a symbolic machine learned datalog program P using p predicate symbols and m meta-rules each having at most j body literals. For a group of humans H, $B(P, H)$ is a bound on the size of hypothesis space such that at most n clauses in P can be comprehended by H and $B(P, H) = m^n p^{(1+j)n}$.

When learned knowledge is cognitively challenging, execution overflows human working memory and instruction stack. We then expect decision making to be more error prone and the task performance of human learners to be less dependable. To account for the cognitive complexity of applying a machine learned theory, we define the cognitive resource of a logic term and atom.

Definition 5 (Cognitive cost of a logic term and atom, $C(T)$): Given T a logic term or atom, the cost of $C(T)$ can be computed as follows:

- $C(\top) = C(\bot) = 1$
- A variable V has cost $C(V) = 1$
- A constant c has cost $C(c)$ which is the number of digits and characters in c
- A list $[T_1, T_2, ...]$ as a data structure used by *MIGO* and *MIPlain* has cost $C([T_1, T_2, ...]) = C(T_1) + C(T_2) + \dots$
- An atom $Q(T_1, T_2, ...)$ has cost $C(Q(T_1, T_2, ...)) = 1 + C(T_1) + C(T_2) + \dots$

Example 1 The Noughts and Crosses position in Figure 2 is represented by an array $[e, x, o, e, e, x, o, e, o]$, where e is an empty field and o and x are marks on the board. It has cognitive cost $C([e, x, o, e, e, x, o, e, o]) = 9$.

Note that we compute cognitive costs of programs without redundancy since repeated literals in programs learned by *MIGO* and *MIPlain* were removed after unfolding for generating explanations which are presented to human populations. Also, a game position can be represented by different data types. We ignore the cost due to implementation and only count digits and marks.

Example 2 An atom $win\_2([e, x, o, e, e, x, o, e, o], X)$ with variable X has a cognitive cost $C(win\_2([e, x, o, e, e, x, o, e, o], X)) = 11$.

We model the inferential process of evaluating training and testing examples by the run-time execution stack of a datalog program. The resolution of a query represents a mental application of a piece of knowledge given a training or testing example. In this work, we neglect the cost of computing the sub-goals of a primitive and compute its cost as if it were a normal predicate for simplicity.

Example 3 A primitive $move(S1, S2)$ which is an atom with variables $S1$ and $S2$ has a cognitive cost $C(move(S1, S2)) = 3$.

Definition 6 (Execution stack of a datalog program, $S(P, q)$): Given a query q, the execution stack $S(P, q)$ of a datalog program P is a set of atoms or terms evaluated during the execution of P to compute q. Each exit point of the execution is replaced with the value $\top$, and each backtrack point has the value $\bot$.

Definition 7 (Cognitive cost of a datalog program, $Cog(P, q)$): Given a query q, and let St represent $S(P, q)$, the cognitive cost of a datalog program P is

$$Cog(P, q) = \min_{St} \sum_{t \in St} C(t)$$

Example 4 The primitive $move/2$ outputs a valid Noughts and Crosses state from a given input game state; the query is $move(s1, B)$. The execution stack contains $move(s1, B)$ and move(s1, s2), $Cog(P, move(s1, B))$ is 10.

S(move(A,B), move(s1, B))	move(s1, B)	move(s1, s2)	$\top$
C(T)	4	5	1

The maintenance cost of task goals in working memory affects performance of problem solving [10]. Background knowledge provides key mappings from solutions obtained in other domains or past experience [4,40] and grants shortcuts for the construction of the current solution process. We expect that when knowledge that provides executional shortcuts is comprehended, the efficiency of human problem solving could be improved due to a lower demand for cognitive resource. Contrarily, in the absence of informative knowledge, performance would be limited by human operational error and would not be better than solving the problem directly. To account for the latter case, we define the cognitive cost of a problem solution that involves the minimum amount of information from the task.

Definition 8 (Minimum primitive solution program, $\bar{M}_\phi(E)$): Given a set of primitives ϕ and examples E, a datalog program learned from examples E using a symbolic machine learning algorithm $\bar{M}$ and a set of primitives $\phi' \subseteq \phi$ is a minimum primitive solution program $\bar{M}_\phi(E)$ if and only if for all sets of primitives $\phi'' \subseteq \phi$ where $|\phi''| < |\phi'|$ and for all symbolic machine learning algorithm M' using ϕ'', there exists no machine learned program $M'(E)$ that is consistent with examples E.

Given a machine learning algorithm M using primitives ϕ and examples E, a minimum primitive solution program $\bar{M}_\phi(E)$ is learned by using the smallest subset of ϕ such that $\bar{M}_\phi(E)$ is consistent with E. A minimum primitive solution program is defined to not use more auxiliary knowledge than necessary but does not necessarily have the minimum cognitive cost over all programs learned with examples E.

Remark 1 Given that the training examples of Noughts and Crosses are winnable and $MIPlain$ uses the set of primitives $\phi = \{move/2, won/1, number\_of\_pairs/3\}$, a minimum primitive solution program is produced by $MIGO$. This is because $MIGO$ uses primitives $\{move/2,\ won/1\}$ which is a strict subset of ϕ for making a move and deciding a win when the input is winnable. Primitives $move/2$ and $won/1$ are also the necessary and sufficient primitives to win Noughts and Crosses and no theory can be learned using a subset of ϕ with the cardinality of one.

Definition 9 (Cognitive cost of a problem solution, $CogP(E, \phi, q)$): Given examples E, primitive set ϕ and a query q, the cognitive cost of a problem solution is

$$CogP(E, \phi, q) = \min_{\bar{M}} Cog(\bar{M}_\phi(E), q)$$

where $\bar{M}_\phi(E)$ is a minimum primitive solution program.

Remark 2 The program P learned by $MIPlain$ has less cognitive cost than the one learned by $MIGO$ except for queries concerning $win\_1/2$. Given sufficient examples E and primitive set used by $MIPlain$, $\phi = \{move/2, won/1, number\_of\_pairs/3\}$, based on Definition 6 to 9, we have $Cog(P, x_1) = CogP(E, \phi, x_1)$, $Cog(P, x_2) < CogP(E, \phi, x_2)$ and $Cog(P, x_3) < CogP(E, \phi, x_3)$ where $x_i = win_i(s_i, V)$ in which s_i represents a position winnable in i moves and V is a variable.

We give a definition of human cognitive window based on theory complexity during knowledge acquisition and theory execution cost during knowledge application. A machine learned theory has 1) a harmful explanatory effect when its hypothesis space size exceeds the cognitive bound and 2) no beneficial explanatory effect if its cognitive cost is not sufficiently lower than the cognitive cost of the problem solution.

Table 3: Criteria for evaluating verbal responses and examples for category $win\_2/2$.

$Q(r)$	Criteria	Exemplary r
Level 0	r does not fit into any of the categories below	“Follow the instructions.”
Level 1	One or more primitives in the machine learned theory, directly or by synonyms, are described correctly in r	“This move gives me a pair.”
Level 2	All primitives in the machine learned theory, directly or by synonyms, are described correctly in r	“I should have picked this move to prevent the opponent and get two attacks.”
Level 3	r is unambiguous and follows a matching executional order as the machine learned theory	"This move gives me two attacks and prevents the opponent from getting a pair."
Level 4	r explains one or more primitives in the machine learned theory in correct causal relations	“This is a good move because by making two pairs and blocking the opponent, the opponent cannot win in one turn and can only block one of my pairs.”

Definition 10 (Cognitive window of a machine learned theory): Given a definition D, a symbolic machine learning algorithm M, examples E, $M(E)$ is a machine learned theory using the primitive set ϕ and belongs to a program class with hypothesis space S. For a group of humans H, E_{ex} satisfies

1. $E_{ex}(D, H, M(E)) < 0$ if $|S| > B(M(E), H)$ and
2. $E_{ex}(D, H, M(E)) \leq 0$ if $Cog(M(E), x) \geq CogP(E, \phi, x)$ for queries x that $h \in H$ have to perform after study

4 Experimental framework

In the following section, we describe an experimental framework for assessing the impact of cognitive window on the explanatory effects of a machine learned theory. Our experimental framework involves 1) a set of criteria for evaluating the participants' learning quality from their own verbal descriptions of learned strategies and 2) an outline of experimental hypotheses. For game playing, we assume humans are able to explain actions by verbalising procedural rules of strategy. We expect verbal responses to provide insights about human decision making and knowledge acquisition. The quality of verbal responses can be affected by multiple factors such as motivation, familiarity with the introduced concepts and understanding of the game rules. We take into account these factors in the evaluation criteria.

Definition 11 (Primitive coverage of a verbal response): A verbal response correctly describes a primitive if the semantic meaning of the primitive is unambiguously stated in the response. The primitive coverage is the number of primitives in a symbolic machine learned theory that are described correctly in a verbal response.

Definition 12 (Quality of a verbal response, $Q(r)$): A verbal response r is checked against the specifications from Table 3 in an increasing order from criteria level 1 to level 4. $Q(r)$ is the highest level i that r can satisfy. When a response does not satisfy any of the higher levels, the quality of this response is the lowest level 0.

To illustrate, we consider the predicate $win\_2/2$ learned by $MIPlain$ (Table 2). Primitive predicates are $move/2$ and $number\_of\_pairs/3$. We present in Table 3 a number of examples of verbal responses. A high quality response reflects a high motivation and good understanding of game concepts and strategy. On the other hand, a poor quality response demonstrates a lack of motivation or poor understanding.

Definition 13 (High (HQ) / low (LQ) quality verbal response): A HQ response rh has $Q(rh) \geq 3$ and a LQ response rl has $Q(rl) < 3$.

We define the following null hypotheses to be tested in Section 5 and describe the motivations. Let M denote a symbolic machine learning algorithm. E stands for examples, D is a target definition, H is a group of participants sampled from a human population. $M(E)$ denotes a machine learned theory which belongs to a definite clause program class with hypothesis space S. First, we are interested in demonstrating whether 1) the verbal response quality of learned knowledge reflects comprehension, 2) there exist cognitive bounds for humans to provide verbal responses of higher quality and 3) the machine learned theory helps improve the quality of verbal responses.

H1: *Unaided human comprehension $C(D, H, E)$ and machine-explained human comprehension $C_{ex}(D, H, M(E))$ manifest in verbal response quality $Q(r)$.* We examine if high post-test accuracy correlates with high response quality and high primitive coverage of each question category.

H2: *Difficulty for human participants to provide verbal response increases with quality $Q(r)$.* We examine if the proportion of verbal responses reduces with respect to high response quality and high primitive coverage of each question category.

H3: *Machine learned theory $M(E)$ improves verbal response quality $Q(r)$.* We examine if machine-aided learning results in more HQ responses.

The impact of a cognitive window on explanatory effects is tested via the following hypotheses. ϕ is a set of primitives introduced to H. Let x denote the set of questions that human $h \in H$ answers after learning.

H4: *Learning a complex theory ($|S| > B(M(E), H)$) exceeding the cognitive bound leads to a harmful explanatory effect ($E_{ex}(D, H, M(E)) < 0$).* We examine if the post-test accuracy, after studying a machine learned theory that participants cannot recall fully, is worse than the accuracy following self-learning.

H5: *Applying a theory without a low cognitive cost ($Cog(M(E), x) \geq CogP(E, \phi, x)$) does not lead to a beneficial explanatory effect ($E_{ex}(D, H, M(E)) \leq 0$).* We examine if the post-test accuracy, after studying a machine learned theory that is cognitively costly, is equal to or worse than the accuracy following self-learning.

5 Experiments

This section introduces the materials and experimental procedure which we designed to examine the explanatory effects of a machine learned theory on human learners. Afterwards, we describe the experiment interface and present experimental results.

5.1 Materials

We assume that Noughts and Crosses is a widely known game a lot of participants of the experiments are familiar with. This might result in many participants already

Table 4: Summary of experiment parts. Participants played one mock game against a random computer player for the more difficult Island Game. After selecting a move in training and regardless of its correctness, participants received the labels of the two moves presented; treated participants additionally received explanations generated from $MIPlain$'s learned program. We introduced the primitive set used by $MIPlain$.

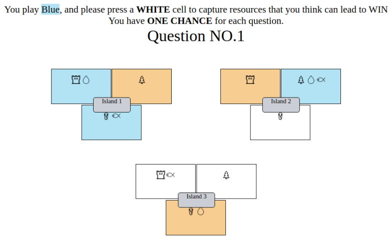
Part	Participant's assignment	No.	Question format
Intro	Understand rules to move and win	1	practice
Pre-test	Choose the optimal move	15	five canonical positions each for win_1, win_2 & win_3
Training	Understand the concept of pairs; choose the optimal move and reflect on the choice	9	two choices each for win_1, win_2 & win_3; presentation of the labels
Post-test	Choose the optimal move	15	five canonical positions each for win_1, win_2 & win_3; Rotated and flipped from pre-test questions.
Open questions	Describe the strategy of a previously made move	6	Questions requiring verbal response
Survey	Provide gender, age group & education level	3	multiple choices

playing optimally before receiving explanations, leaving no room for potential performance increase. In order to address this issue, the *Island Game* was designed as a problem isomorphic to Noughts and Crosses. [51] define isomorphic problems as "problems whose solutions and moves can be placed in one-to-one relation with the solutions and moves of the given problem". This changes the superficial presentation of a problem without modifying the underlying structure. Several findings imply that this does not impede solving the problem via *analogical inference* if the original problem is consciously recognized as an analogy; on the other hand, the prior step of initially identifying a helpful analogy via *analogical access* is highly influenced by superficial similarity [15,20,44]. Given that the Island Game presents a major re-design of the game surface, we expect that participants will less likely recall prior experience of Noughts and Crosses that would facilitate problem solving, leading to less optimal play initially and more potential for performance increase.

The Island Game (Figure 4) contains three islands, each with three territories on which one or more resources are marked. The winning condition is met when a player controls either all territories on one island or three instances of the same resource. The nine territories resemble the nine fields in Noughts and Crosses and the structure of the original game is maintained in regard to players' turns, possible moves, board states and win conditions. This isomorphism masks a number of spatial relations that represent the membership of a field to a win condition. In this way, the fields can be rearranged in an arbitrary order without changing the structure of the game.

5.2 Methods and design

We use two experiment interfaces, one for Noughts and Crosses and another one for the Island Game. For both, we adopt a two-group pre-test post-test design (Table 4). In the pre-test, performance of participants in both self learn-



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

Fig. 4: Example of pre- and post-test question for the Island Game. A board is presented to the participant who has to select the move that he or she thinks is optimal.

ing and machine-aided learning groups are measured in an identical way. During training, we introduce to participants the concept of pairs and they are able to see correct answers of some game positions. In the post-test, performance of both self-learning and machine-aided groups are evaluated in the exact same way as in the pre-test. This experiment setting allows to evaluate the degree of change in performance as the result of explanations. Each question in pre- and post-test is the presentation of a board for which it is the participant's turn to play. They are asked to select what they consider to be the optimal move. A question category of win_i denotes a game position winnable in i moves of the human player. An exemplary question is shown in the Figure 4. The post-test questions are rotated and flipped from pre-test questions. In each test, only 15 questions are given to limit experiment duration to one hour. The response time of participants was recorded for each pre-test and post-test question.

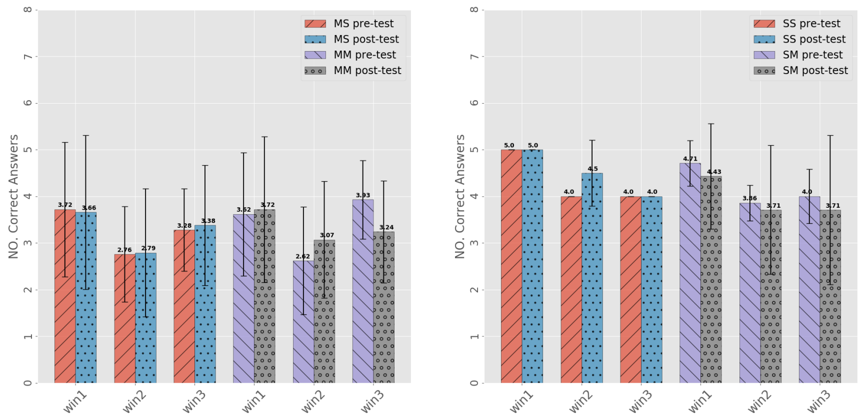
The treatment was applied to the machine-aided group. In the interest of experimentation, during treatment, we present both visual and textual explanations to avoid unnecessary effort of participants to associate textual explanations to game positions and concepts. This is based on the consideration that direct association between textual explanations and game states which can be abstract for participants who are not familiar with the designed game domain. Learned first-order theories have been translated with manual adjustments based on primitives provided to all participants and to *MIPlain*. An exemplary explanation is shown in Figure 1. Both visual and textual explanations preserve the structure of hypotheses to account for the reasons that make a move right and the other move wrong. Conversely, during training, the self-learning group was presented with similar game position without the corresponding visual and textual explanations. For the Island Game experiments, we recorded an English description of the strategy they used for each of the selected post-test questions. Participants are presented previously submitted answers, one at a time along with a text input box for written answers. Moves for these open questions are selected from post-test with a preference order from wrong and hesitant moves to consistently correct moves. We associate hesitant answers with higher response times. A total of six questions are selected based on individual performance during the post-test.

5.3 Experiment results

We conducted three experiments[2] using the interface with Noughts and Crosses questions and explanations. These experiments were carried out on three samples: an undergraduate student group from Imperial College London, a junior student group from a German middle school and a mixed background group from Amazon Mechanical Turk[3] (AMT). No consistent explanatory effects could be observed for any of the mentioned samples. The problem solving strategy that humans apply can be affected by factors such as task familiarity, problem difficulty, and motivation. For

[2] raw data are available upon request from the authors

[3] AMT is a online crowdsourcing platform which we used to recruit experiment samples



(a) Mixed background self learning and machine-aided learning.

(b) Student self learning and machine-aided learning.

Fig. 5: Number of correct answers in pre- and post-test with respect to question categories.

instance, [45] suggested that a rather superficial analogical transfer of a strategy is applied when a problem is too difficult or when there is no reason to gain a more general understanding of a problem. Given that the majority of subjects achieved reasonable initial performance, we ascribe the reason of such results to experience with the game and complexity of explanations. The game familiarity of adult groups led to less potential for performance improvement. Early middle school students had limited attention and were overwhelmed by information intake. Alternatively, we focused on specially designed experiment materials in the following experiments.

5.3.1 Island Game with open questions

A sample from Amazon Mechanical Turk and a student sample from the University of Bamberg participated in experiments[2] that used the interface with Island Game questions and explanations. To test hypotheses **H1** to **H5**, we employed a quantitative analysis on test performance and a qualitative analysis on verbal responses. A sub-sample with a mediocre initial performance within one standard deviation of the mean was selected for the performance analysis. This aims to discount the ceiling effect (initial performance too high) and outliers (e.g. struggling to use the interface).

From AMT sample, we had 90 participants who were 18 to above 65 years old. A sub-sample of 58 participants with a mediocre initial performance was randomly partitioned into two groups, **MS** (Mixed background Self learning $n = 29$) and **MM** (Mixed background Machine-aided learning, $n = 29$). A different sub-sample of 30 participants completed open questions and was randomly split into two groups, **MSR** (Mixed background Self learning and strategy Recall, $n = 15$) and **MMR** (Mixed background Machine-aided learning and strategy Recall, $n = 15$). As shown in Figure 5a, in category win_2, **MM** post-test had a better comprehension ($p = 0.028$) than **MS** post-test while **MM** and **MS** had similar pre-test performance ($p > 0.1$) in this category. Results in category win_2 indicate that explanations have a beneficial effect on **MM**. However, **MM** did not have a better comprehension on win_1 than **MS** given the same initial performance ($p > 0.1$). In addition, **MM** had the same initial performance as **MS** on *win*_3 ($p > 0.1$) but **MM**'s performance reduced after receiving explanations of *win*_3 ($p = 0.005$).

From a group of students involved in a Cognitive Systems course at the University of Bamberg, we had 13 participants who were 18 to 24 years old and a few outliers between 25 and 54 years. All participants were asked to complete open questions and were randomly split into two groups, **SSR** (Student Self learning and strategy Recall, $n = 4$) and **SMR** (Student Machine-aided learning and strategy Recall, $n = 9$). A sub-sample of 9 with a mediocre initial performance was randomly divided into **SS** (Student Self learning, $n = 2$) and **SM** (Student Machine-aided learning, $n = 7$). The imbalance in the student sample was caused by a number of participants leaving during the experiment. The machine-aided learning results show large performance variances in post-test as evidence for insignificant levels of performance degradation.

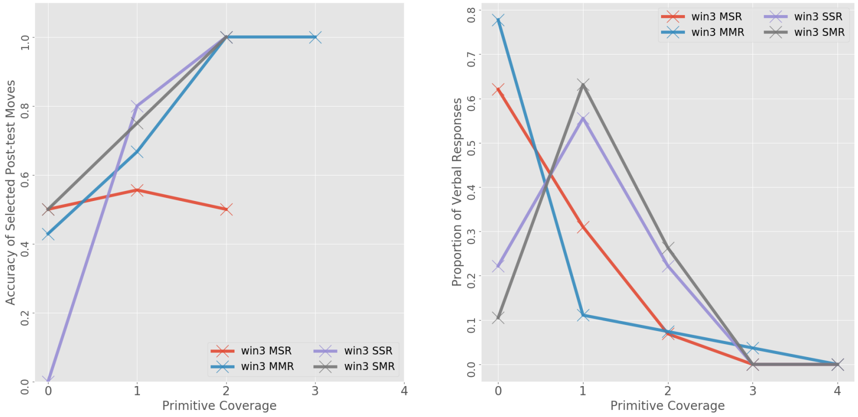
In Table 5, we identified that participants who were able to provide high quality responses for their test answers scored higher on these questions. This is not the case for *win*_3, however, due to the high difficulty of providing good description of strategy for *win*_3 category. Additionally, in the *win*_2 category, both machine-aided groups (**MMR**: 2/(2+35), **SMR**: 9/(9+14)) have greater proportions of high quality responses than self learning groups (**MSR**: 1/(1+32), **SSR**: 1/(1+8)). Also, we observed a pattern in which there are less HQ responses than LQ responses in *win*_1 and *win*_2 categories. This pattern is more significant in *win*_2 category.

Figure 6 illustrates the difficulty of providing good quality verbal response for the non-trivial category *win*_3. Since *win*_1 contains only two predicates, we examined primitive coverage of non-trivial categories *win*_2 and *win*_3. However, for clarity of presentation, we only show category *win*_3 which has more remarkable trends. When counting primitives based on Definition 11, we only consider the constraint *number_of_pairs*/3 and ignore the move generator *move*/2 as participants were required to make a move when they answered a question.

In Figure 6a, we plotted primitive coverage against the accuracy of post-test answers that were selected as open questions. We observed a major *monotonically increasing trend* in accuracy with respect to primitive coverage. This indicates that high matching between verbal responses and the machine learned theory correlates with high performance. In Figure 6b, we observed *downward curves* for **MSR** and **MMR** in the number of verbal responses from the lower to the higher primitive coverage. More responses were provided by **SSR** and **SMR** covering *one primitive* than **MSR** and **MMR**. Participants gave very few responses that cover *more than two* primitives. Based on the learned theory in Table 2, the results suggest an

Table 5: The number and accuracy of HQ and LQ responses for groups **MSR**, **MMR**, **SSR**, **SMR** and each question category. For *win*_3, the most mentally challenging category of all three, no HQ response was given.

		*win*_1	*win*_2	*win*_3
MSR	No. HQ / post-train accuracy	9 / 0.889	1 / 1.0	-
	No. LQ / post-train accuracy	19 / 0.421	32 / 0.406	29 / 0.517
MMR	No. HQ / post-train accuracy	8 / 1.00	2 / 1.00	-
	No. LQ / post-train accuracy	16 / 0.250	35 / 0.486	29 / 0.483
SSR	No. HQ / post-train accuracy	6 / 1.00	1 / 1.00	-
	No. LQ / post-train accuracy	0 / 0.00	8 / 0.750	9 / 0.667
SMR	No. HQ / post-train accuracy	9 / 1.00	9 / 0.778	-
	No. LQ / post-train accuracy	3 / 0.00	14 / 0.571	19 / 0.737



(a) The accuracy of verbal responses increases with respect to the number of primitives covered.

(b) The proportion of quality verbal responses decreases with respect to the number of primitives covered.

Fig. 6: *win*_3 reuses *win*_2 and uses four *number*_*of*_*pairs*/3 when unfolded. In Figure 6b, both mixed background groups (**MSR** and **MMR**) had lower proportions of responses covering one predicate than student groups (**SSR** and **SMR**). Mixed background and student groups could not provide a significant proportion of response covering more than one and two primitives respectively (Figure 6a).

Table 6: Hypotheses concerning quality of verbal responses and comprehension. C stands for **c**onfirmed, N denotes **n**ot confirmed, H stands for **h**ypothesis. Test outcomes are presented for *win*_1, *win*_2 and *win*_3 categories.

H		*win*_1	*win*_2	*win*_3
H1	Human comprehensions manifest in verbal response quality	C	C	C
H2	Difficulty for human participants to provide verbal response increases with verbal response quality	C	C	C
H3	Machine learned theory improves verbal response quality	N	C	N

increasing difficulty to provide more complete strategy descriptions *beyond two (mixed background groups) and four (student groups) clauses* of *win*_3.

5.4 Discussion

Results concerning null hypotheses **H1** to **H5** are summarised in Table 6 and 7. First, we assume that (H1 Null) comprehension does not correlate with verbal response quality. Results of HQ responses in two categories (Table 5) suggest that being able to provide better verbal responses of strategy corresponds to a high comprehension. We also examined the coverage of primitives (specifically for LQ responses of *win*_3) in verbal responses (Figure 6a). Evidence in all categories shows a correlation between comprehension and the degree of verbal response matching with explanations. We reject the null hypothesis in all categories which implies the confirmation of H1.

In addition, we assume that (H2 Null) the difficulty for human participants to provide verbal response is not affected by verbal response quality. Since high response quality is difficult to achieve (Table 5) and it is challenging to correctly describe all primitives (Figure 6b), we reject this null hypothesis for all categories and confirm H2

Table 7: Hypotheses concerning cognitive window and explanatory effects. C stands for **c**onfirmed, H stands for **h**ypothesis, T stands for **t**est outcome.

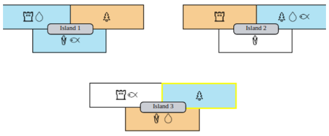
H		T
H4	Learning a complex theory exceeding the cognitive bound leads to a harmful explanatory effect	C
H5	Applying a learned theory without a low cognitive cost does not lead to a beneficial explanatory effect	C

as it is increasingly difficult for participants to provide higher quality verbal response. Hence, two additional trends we observed from the same figure suggest two mental barriers of learning. As we assume a human sample is a collection of version space learners, the search space of participants is limited to programs of size two (mixed background groups) and four (student groups). When H is taken as the student sample and P to be the machine learned theory on winning the Island Game, the cognitive bound $B(P,H) = m^4 * p^{4(j+1)} = 4^4 * 2^{12}$ corresponds to the hypothesis space size for programs with four clauses (four metarules are used with at most two body literals in each clause, primitives are $move/2$ and $number\_of\_pairs/3$).

Furthermore, we assume that (H3 Null) machine learned theory does not improve verbal response quality. Results (Table 5) show higher proportion of HQ responses for machine-aided learning than self-learning in category win_2. Thus, for win_2, we reject this null hypothesis which means H3 is confirmed in category win_2 where the machine explanations result in more high quality verbal responses being provided.

We assume that (H4 Null) learning a descriptively complex theory does not affect comprehension harmfully. When P is the program learned by $MIPlain$, $B(P,H)$ for two samples correspond to program class with size no larger than 4. Only win_3 which has a larger size of seven after unfolding exceeds these cognitive bounds. As harmful effects (Figure 5a and 5b) have been observed in category win_3, this null hypothesis is rejected and H4 is confirmed as learning a complex machine learned theory has a harmful effect on comprehension. We also assume that (H5 Null) applying a theory without a sufficiently low cognitive cost has a beneficial effect on comprehension. Given that the predicate win_1 in $MIPlain$'s learned theory does not have a low cognitive cost, we reject this null hypothesis since no significant beneficial effect has been observed. This null hypothesis is therefore rejected and we confirm H5 – knowledge application requiring much cognitive resource does not result in better comprehension.

The performance analysis (Figure 5a) demonstrates a comprehension difference between self learning and machine-aided learning in category win_2. An explanatory effect has not been observed for the student sample. While the conflicting results suggest that a larger sample size would likely ensure consistency of statistical evidence, the patterns in results suggest more significant results in category win_2 than win_1 and win_3. The predicate win_2 in the program learned by $MIPlain$ satisfies both constraints on hypothesis space bound for knowledge acquisition and cognitive cost for knowledge application. In addition, the cognitive window explains the lack of beneficial effects of predicates win_1 and win_3. The former does not have a lower cognitive cost for execution so that operational errors cannot be reduced, thus there has been no observable effects. The latter is a complex rule with a larger hypothesis space for human participants to search from and harmful effects have been observed due to partial knowledge being learned.



```
win_1(A,B):-move(A,B),
            win_1_1(B).
win_1_1(A):-number_of_pairs(A,x,1),
            number_of_pairs(A,x,1).
```

Fig. 7: Left: participant's chosen move from the initial position in Figure 4. Right: *Metagol* one-shot learns from participant's move a program representing his strategy. The learned program denotes strategy of finding a pair rather than going for a direct win, which is a mismatch between taught and learned knowledge.

6 Conclusions and further work

While the focus of explainable AI approaches has been on explanations of classifications [1], we have investigated explanations in the context of game strategy learning. In addition, we have explored both beneficial and harmful sides of the AI's explanatory effect on human comprehension. Our theoretical framework involves a cognitive window to account for the properties of a machine learned theory that lead to improvement or degradation of human performance. The presented empirical studies have shown that explanations are not helpful in general but only if they are of appropriate complexity – being neither informatively overwhelming nor more cognitively expensive than the solution to a problem itself. It would appear that complex machine learning models and models which cannot provide abstract descriptions of internal decisions are difficult to be explained effectively. However, we acknowledge the limitation of our empirical studies in terms of consistency of statistical evidence as groups vary greatly in sample size which might be addressed with further experimentation.

To explain a strategy, typically goals or sub-goals must be related to actions which can fulfill these goals. If the strategy involves to keep in mind a stack of open sub-goals – as for example the Tower of Hanoi [2,46] – explanations might become more complex than figuring out the action sequence. Based on [8], knowledge is learned by humans in an incremental way, which was recently emphasized by [58] on human category learning. A potential approach to improve explanatory effectiveness of a machine learned theory is to process complex concepts into smaller chunks by initially providing simple-to-execute and short sub-goal explanations. Mapping input to another sub-goal output thus consumes lower cognitive resources and improvement in performance is more likely. It is worth investigating for future work a teaching procedure involving a sequence of teaching sessions that issues increasingly difficult tasks and explanations. Abstract descriptions might be generated in the form of invented predicates as it has been shown in previous work on ILP as an approach to USML [34]. An example for such an abstract description for the investigated game is the predicate *number_of_pairs*/3. Therefore, learning might be organised incrementally, guided by a curriculum [6,52].

In addition, the current teaching procedure, which only specifies humans as learners, could be augmented to enable two-way learning between human and machine. Human decisions might be machine learned and explanations would be provided based on estimation of human errors during the course of training. A simple demonstration of this idea is presented in Figure 7. We would like to explore, in the future, an interactive procedure in which a machine iteratively re-teaches human learners by targeting human learning errors via specially tailored explanations. [7] suggested it is crucial for machine produced clones to be able to represent goal-oriented knowledge which is in a form that is similar to human conceptual structure. Hence, MIL is an appropriate candidate for cloning since it is able to iteratively learn complex concepts by inventing sub-goal predicates. We hope to incorporate cloning to predict and target mistakes in human learned knowledge from answers in a sequence of re-training. We expect a "clean up" on operation errors of human behaviours from empirical experiments by presenting appropriate explanations in re-training. Such corrections and improvements guided by identified errors in a human strategy are also helpful in the context of intelligent tutoring [57] where classic strategies such as algorithmic debugging [48] can be applied to make humans and machines learn from each other.

Acknowledgements

The contribution of the authors from University of Bamberg is part of a project funded by the Deutsche Forschungsgemeinschaft (DFG, German Research Foundation) – 405630557 (PainFaceReader). The second author acknowledges support from the UK's EPSRC Human-Like Computing Network, for which he acts as director.

References

1. A. Adadi and M. Berrada. Peeking inside the black-box: A survey on explainable artificial intelligence (XAI). *IEEE Access*, 6:52138–52160, 2018.
2. E. Altmann and J. G. Trafton. Memory for goals: An activation-based model. *Cognitive Science*, 26:39–83, 2002.
3. J. R. Anderson, N. Kushmerick, and C. Lebiere. *Rules of the Mind*, chapter The Tower of Hanoi and goal structures, pages 121–142. Hillsdale, NJ: L. Erlbaum, 1993.
4. J. R. Anderson and R. Thompson. *Use of Analogy in a Production System Architecture*, page 267–297. Cambridge University Press, USA, 1989.
5. M. Bain and S. H. Muggleton. *Learning Optimal Chess Strategies*, pages 291–309. Oxford University Press, Inc., New York, NY, USA, 1995.
6. Y. Bengio, J. Louradour, R. Collobert, and J. Weston. Curriculum learning. In *Proceedings of the 26th Annual International Conference on Machine Learning*, pages 41–48, 2009.
7. I. Bratko, T. Urbančič, and C. Sammut. Behavioural cloning: Phenomena, results and problems. *IFAC Proceedings Volumes*, 28(21):143–149, 1995.
8. J. S. Bruner, J. J. Goodnow, and G. A. Austin. *A study of thinking: With an appendix on language by Roger W. Brown.* New York, NY: Wiley, 1956.
9. J. Carbonell. Derivational analogy: A theory of reconstructive problem solving and expertise acquisition. *Machine Learning*, 11:26, 1985.
10. P. Carpenter, M. Just, and P. Shell. What one intelligence test measures: A theoretical account of the processing in the raven progressive matrices test. *Psychological review*, 97:404–431, 1990.
11. N. Chomsky. *Aspects of the theory of syntax.* Cambridge: M.I.T. Press, 1965.
12. A. Cropper and S. H. Muggleton. Metagol system. https://github.com/metagol/metagol, 2016.
13. A. Cropper and S. H. Muggleton. Learning efficient logic programs. *Machine Learning*, 108:1063–1083, 2019.
14. S. Džeroski, L. De Raedt, and K. Driessens. Relational reinforcement learning. *Machine Learning*, 43:7–52, 2001.
15. D. Gentner and R. Landers. Analogical reminding: A good match is hard to find. *Proceedings of the International Conference on Systems, Man and Cybernetics*, 1985.
16. S. A. Goldman and M. J. Kearns. On the complexity of teaching. *Journal of Computer and System Sciences*, 50:20–31, 1995.
17. M. D. Hauser, N. Chomsky, and W. T. Fitch. The faculty of language: what is it, who has it, and how did it evolve? *Science*, 298:1569–1579, 2002.

18. M. Hind, D. Wei, M. Campbell, N. Codella, A. Dhurandhar, and A. e. a. Mojsilovic. Ted: Teaching ai to explain its decisions. *Proceedings of the 2019 AAAI/ACM Conference on AI, Ethics, and Society*, 2019.
19. J. R. Hobbs. Abduction in natural language understanding. *In The Handbook of Pragmatics (eds L.R. Horn and G. Ward)*, 2008.
20. K. J. Holyoak and K. Koh. Surface and structural similarity in analogical transfer. *Memory & Cognition 15(4)*, pages 332–340, 1987.
21. P. N. Johnson-Laird. *Mental Models: Towards a Cognitive Science of Language, Inference, and Consciousness.* Harvard University Press, USA, 1986.
22. A. N. Kolmogorov. On tables of random numbers. *Sankhya: The Indian Journal of Statistics, Series A,*, 207(25):369–375, 1963.
23. E. Lemke, H. Klausmeier, and C. Harris. Relationship of selected cognitive abilities to concept attainment and information processing. *Journal of educational psychology*, 58:27–35, 1967.
24. D. Lin, E. Dechter, K. Ellis, J. Tenenbaum, and S. H. Muggleton. Bias reformulation for one-shot function induction. *In Proceedings of the 23rd European Conference on Artificial Intelligence (ECAI 2014)*, pages 525–530, 2014.
25. D. Michie. Experiments on the mechanization of game-learning part i. characterization of the model and its parameters. *The Computer Journal, Volume 6, Issue 3*, pages 232–236, 1963.
26. D. Michie. Inductive rule generation in the context of the fifth generation. *Machine Learning Workshop*, page 65, 1983.
27. D. Michie, M. Bain, and J. Hayes-Michie. Cognitive models from sub cognitive skills. *Knowledge-Based Systems in Industrial Control*, pages 71–99, 1990.
28. D. Michie and R. Camacho. Building symbolic representations of intuitive real-time skills from performance data. In *Machine Intelligence*, 1992.
29. G. A. Miller. The magical number seven, plus or minus two: Some limits on our capacity for processing information. *The Psychological Review*, 63:81–97, 1956.
30. T. Miller. Explanation in artificial intelligence: Insights from the social sciences. *Artificial Intelligence*, 267:1–38, 2019.
31. T. Miller, P. Howe, and L. Sonenberg. Explainable ai: Beware of inmates running the asylum or: How i learnt to stop worrying and love the social and behavioural sciences. *Proc. IJCAI Workshop Explainable Artif. Intell. Melbourne, Australia.*, 2017.
32. T. M. Mitchell. Generalization as search. *Artificial Intelligence*, 18:203–226, 1982.
33. V. Mnih, K. Kavukcuoglu, and D. e. a. Silver. Human-level control through deep reinforcement learning. *Nature*, 518:529–533, 2015.
34. S. Muggleton, U. Schmid, C. Zeller, A. Tamaddoni-Nezhad, and T. Besold. Ultra-strong machine learning: comprehensibility of programs learned with ilp. *Machine Learning*, 2018.
35. S. H. Muggleton. Inductive logic programming. *New Gen. Comput.*, 8:295–318, 1991.
36. S. H. Muggleton and C. Hocquette. Machine discovery of comprehensible strategies for simple games using meta-interpretive learning. *New Generation Computing*, 37:203–217, 2019.
37. S. H. Muggleton and D. Lin. Meta-interpretive learning of higher-order dyadic datalog: Predicate invention revisited. *In Proceedings of the 23rd International Joint Conference Artificial Intelligence*, pages 1551–1557, 2013.
38. S. H. Muggleton, D. Lin, N. Pahlavi, and A. Tamaddoni-Nezhad. Meta-interpretive learning: application to grammatical inference. *Machine Learning*, pages 25–49, 2014.
39. A. Newell. *Unified Theories of Cognition.* Harvard University Press, USA, 1990.
40. L. Novick and K. Holyoak. Mathematical problem solving by analogy. *Journal of experimental psychology. Learning, memory, and cognition*, 17:398–415, 1991.
41. J. Quinlan. *Learning Efficient Classification Procedures and Their Application to Chess End Games*, pages 463–482. Springer Berlin Heidelberg, Berlin, Heidelberg, 1983.
42. J. R. Quinlan. Simplifying decision trees. *International Journal of Man-Machine Studies*, 27:221–234, 1987.
43. A. N. Rafferty, E. Brunskill, T. L. Griffiths, and P. Shafto. Faster teaching via pomdp planning. *Cognitive science*, pages 1290–1332, 2016.
44. S. K. Reed, C. C. Ackinclose, and A. A. Voss. Selecting analogous problems: Similarity versus inclusiveness. *Memory & Cognition 18(1)*, pages 83–98, 1990.
45. U. Schmid and J. Carbonell. Empirical evidence for derivational analogy. *Proceedings of the 21st annual conference of the cognitive science society*, 2000.

46. U. Schmid and E. Kitzelmann. Inductive rule learning on the knowledge level. *Cognitive Systems Research*, 12:237–248, 2011.
47. A. Shapiro and T. Niblett. Automatic induction of classification rules for a chess endgame. In M. Clarke, editor, *Advances in Computer Chess*, volume 3, pages 73–91. Pergammon, Oxford, 1982.
48. E. Y. Shapiro. Algorithmic program debugging. acm distinguished dissertation, 1982.
49. E. Shohamy. Performance and competence in second language acquisition. *Competence and performance in language testing*, pages 138–151, 1996.
50. D. Silver, A. Huang, C. J. Maddison, A. Guez, L. Sifre, and G. e. a. van den Driessche. Mastering the game of go with deep neural networks and tree search. *Nature*, 529(7587):484–489, 2016.
51. H. A. Simon and J. R. Hayes. The understanding process: Problem isomorphs. *Cognitive Psychology 8*, pages 165–190, 1976.
52. J. A. Telle, J. Hernández-Orallo, and C. Ferri. The teaching size: computable teachers and learners for universal languages. *Machine Learning*, 108:1653–1675, 2019.
53. T. Urbančič and I. Bratko. Reconstructing human skill with machine learning. *Proceedings of the 11th European Conference on Artificial Intelligence*, pages 498–502, 1994.
54. C. Watkins. *Learning from Delayed Rewards*. PhD thesis, 1989.
55. T. Zahavy, N. B. Zrihem, and S. Mannor. Graying the black box: Understanding dqns. *Proceedings of the 33rd International Conference on Machine Learning*, 2016.
56. V. F. Zambaldi, D. C. Raposo, A. Santoro, V. Bapst, Y. Li, and I. e. a. Babuschkin. Deep reinforcement learning with relational inductive biases. In *ICLR*, 2019.
57. C. Zeller and U. Schmid. Automatic generation of analogous problems to help resolving misconceptions in an intelligent tutor system for written subtraction. In *Workshops Proceedings for the Twenty-fourth International Conference on Case-Based Reasoning*, volume 1815, pages 108–117, 2016.
58. C. Zeller and U. Schmid. A human like incremental decision tree algorithm: Combining rule learning, pattern induction, and storing examples. In *LWDA*, 2017.
59. X. Zhu. Machine teaching: An inverse problem to machine learning and an approach toward optimal education. In *Proceedings of the Twenty-Ninth AAAI Conference on Artificial Intelligence*, page 4083–4087. AAAI Press, 2015.