Causal Generalization via Goal-Driven Analogy

Arash Sheikhlar[1]
Kristinn R. Thórisson[1,2]

Proceedings of Artificial General intelligence, 2024, 165-175

[1] Center for Analysis & Design of Intelligent Agents
Reykjavik University

[1,2] Icelandic Institute for Intelligent Machines
Reykjavik, Iceland

OCTOBER 23 2024

Causal Generalization via Goal-Driven Analogy

Arash Sheikhlar[1 [0000−0002−0568−075X]] &
Kristinn R. Thórisson[1,2 [0000−0003−3842−0564]]

[1] Center for Analysis & Design of Intelligent Agents
Department of Computer Science, Reykjavik U. arash19@ru.is
[2] Icelandic Institute for Intelligent Machines, Reykjavík, Iceland

Abstract. Causal knowledge and reasoning allow cognitive agents to predict the outcome of their actions and infer the likely reasons behind observed events, enabling them to interact with their surroundings effectively. Causality has been the subject of some research in artificial intelligence (AI) over the past decade due to its potential for task-independent knowledge representation and generalization. Yet, the question of how the agents can autonomously generalize their causal knowledge while seeking their active goals still needs to be answered. This work introduces an analogy-based learning mechanism that enables causality-based agents to autonomously generalize their existing knowledge once the generalization aligns with the agents' goal achievement. The methodology is centered on constructivism, causality, and analogy-making. The introduced mechanism is integrated with a general-purpose cognitive architecture, Autocatalytic Endogenous Reflective Architecture (AERA), and evaluated in a robotic experiment in a 3D simulation environment. Both empirical and analytical results show the effectiveness of this mechanism.

Keywords: Analogy · Generalization · Causality · Reasoning

1 Introduction

Generalization is a process wherein information from a source is transferred to a target situation, enabling agents to extend their knowledge to circumstances they have not encountered before [12]. Generalization calls for assessing the similarities and differences between known and newly observed patterns. Comparisons can be used to estimate the familiarity of situations and tasks, enabling an agent to generate new hypotheses about how to solve tasks based on their overlap with existing knowledge. These hypotheses, however, may need to be verified by testing them in action, and possibly modified based on the results. In other words, an effective generalization mechanism calls for interaction with environments, involving taking actions, processing outcomes, and adjusting knowledge to meet the requirements of active goals and subsequent actions. For an agent to interact effectively with its surroundings, it must be able to predict the outcomes of actions and infer their underlying reasons, commonly called *causal reasoning*. A key step for realizing such a system lies in using task-independent knowledge representations based on causal information.

We present a goal-driven analogy-based induction mechanism that hypothesizes new causal-relational models (CRMs) in novel situations. Implemented in the causality-based AERA cognitive architecture (Autocatalytic Endogenous Reflective Architecture) [7], the mechanism relies on the computation of an agent's familiarity with situations, using a goal-driven comparison process. The result is a system that can autonomously learn the significance of compared properties of phenomena, using backward reasoning, and verify the hypothesized CRMs via direct interaction with environments. We evaluate the mechanism within a motor skills learning task, where an AERA agent controls a robot arm in a simulation environment to learn and generalize its own object manipulation skills.

2 Related Work

Autonomous generalization of knowledge calls for explicit, goal-driven reasoning, allowing a cognitive agent to make analogies whenever needed. The related studies in supervised machine learning, i.e., transfer learning [14] and analogy-making [6] in stand-alone artificial neural networks, rely on human programmers to choose the training and test domains, making them insufficient for interactive agents that autonomously make analogies and generalize whenever required. Even reinforcement learners (RLs) have knowledge transfer issues due to their knowledge representations and assumptions [3]. On the one hand, model-free RLs learn policies that are not transferable to tasks with different goal structures, as the policies are reward-entangled inherently. On the other hand, model-based RLs rely on learning invariant dynamics, whereas agents seeking general intelligence need to learn falsifiable and, thus, non-invariant knowledge under the *assumption of insufficient knowledge and resources* (AIKR) [13].

Symbolic approaches to generalization and analogy-making are traditionally based on identifying mappings between propositional descriptions, such as in the structure mapping engine (SME) ([4]). SME only takes a system of relations into account in analogy-making. Another approach to behavior-based analogies is case-based reasoning (CBR), which allows an agent to retrieve similar cases/situations to the current case, compute an action, predict the outcome, and then store the results ([1]). Although SME and CBR are useful for interactive agents, the standard systems based on these approaches do not learn to prune pieces of knowledge that are no longer accurate. Working under AIKR, Non-axiomatic logic system (NARS) can learn to modify the conditions under which its actions are taken [13]. Yet, the analogy-making in SME, CBR, and NARS is similarity-driven rather than goal-driven and is not necessarily involved in the systems' generalization process.

3 Theoretical & Methodological Framework

This section describes the theoretical framework of this work, emphasizing the constructivist AI principles [11] in shaping the design and development of cognitive agents. Here, we briefly describe some of the constructivist AI principles.

- **Learning controllers and feedback loops**: Constructivist AI calls for a cognitive agent, an embodied learning controller, that interacts with its environment and learns from performing experiments, allowing the generation of new pieces of knowledge or refining the existing ones.
- **Causality**: According to Pearl's causal hierarchy [8], intelligence has three levels: association, intervention, and counterfactuals. Agents with higher levels of causal understanding can not only predict interventions' outcomes but use counterfactuals to hypothesize relations about imaginative scenarios.
- **Assumption of insufficient knowledge and resources (AIKR)**: Reasoning in complex environments must be non-axiomatic because there is no ultimate guarantee that anything is as it seems, and thus, knowledge must be formed and used by taking AIKR into account [13]. According to this assumption, knowledge has degrees of truth that can change by experience.
- **Ampliative Reasoning**: For cognitive agents with causal models working under AIKR, information processing mechanisms must allow for learning to make predictions and plans using a unified reasoning mechanism occurring via deduction, abduction, and induction [13].
- **Temporal compositional representations**: A time-dependent knowledge representation allows for the temporal relationship between situations to be established [11]. The knowledge must also be able to represent complex tasks at different levels of detail, facilitating task-independent reasoning.

Constructivist AI [11, 2] lays a proper foundation for studying how agents can achieve cognitive growth, learn from their environments, and perform reasoning, ultimately moving toward the realization of artificial general intelligence. The next section formulates the generalization problems our work aims to address.

4 Causal Generalization

Causal-relational models (CRM) [7] are a representation of causal knowledge based on the principles of constructivist AI [11]. A CRM holds left-hand-side (LHS) and right-hand-side (RHS) patterns - knowledge constructs that incorporate data patterns, timing, transition functions, operations, and conditions. A CRM, denoted as $M : [A \quad B]$, represents a transition from A to B, meaning that, if A (the LHS) takes place at time t_1, then B (the RHS) will occur later at t_2. A specific case of A is (P, cmd) where P is the precondition set representing observed and internal facts, and cmd is an internal command the agent applies. Therefore, a CRM can be represented as follows

$$M : [(P(t_1), cmd(t_1)) \quad B(t_2)] \qquad cfd : [0, 1] \tag{1}$$

The precondition set P is the aggregation of a set of simultaneous, related facts representing properties, relations, and conditions contextualizing the transition from cmd to B. The CRMs are different from Drescher's *schemas* [2] in that CRMs are falsifiable hypotheses and have degrees of truth, called *confidence* cfd, representing the ratio of positive evidence pe to the total evidence available

te, $cfd = \frac{pe}{te}$, where $0 \leq cfd \leq 1$. The CRMs support ampliative reasoning. If some input facts match their LHS patterns, they deduce predictions based on their RHS, whereas abduction occurs if a goal fact matches the CRM's RHS.

Generalization of CRMs can occur at different levels of detail, from simple abstraction to complex knowledge pruning.

- **1. Abstraction**: The data the agent observes or infers can be represented by information patterns whose generalization covers a range of similar situations. Abstraction occurs by replacing values and task entities with variables.
- **2. Selective attention**: Selective attention is the process of aggregating a set of simultaneous, abstracted, relevant patterns involved in a transition. More precisely, let S be a situation representing all observed/predicted patterns, and T be a state transition. Then selective attention A chooses a subset of S deemed relevant to T. In other words, $A(S,T) \rightarrow P$ where $P \subseteq S$.
- **3. Induction**: Induction refers to the process of creating CRMs from an instance of a state change or a set of observed states. The induction process uses *selective attention* to choose a subset of relevant patterns and performs *abstraction* to generalize the selected patterns.
- **4. Knowledge pruning**: A learning controller under AIKR must be able to revise its known CRMs whenever needed, one aspect of which is to generalize CRMs by pruning constraints, allowing them to match even more situations. If the original preconditions set of a CRM is P, then the pruned preconditions set is $G(P) = P' \subseteq P$, where G selectively removes the conditions in P.

An ampliative reasoning-based controller working under AIKR calls for a mechanism that autonomously generalizes its CRMs considering the above aspects.

5 Goal-Driven Analogy: A Mechanism

In this work, a top-down analogy process is introduced that provides both induction and knowledge pruning by removing novel patterns from model preconditions and keeping the familiarities. It identifies the overlaps between the patterns of its known models and the patterns of situations via a familiarity computation.

Familiarity and comparison
We base our introduced mechanism on the theory of autonomous cumulative transfer learning [9], stating that a cognitive agent can make a prediction about a phenomenon if and only if the phenomenon is familiar to the agent. This theory implies that the *degree of familiarity* determines the *confidence of the prediction* and, more precisely, the *confidence of the model* utilized for prediction-making.

Familiarity computation is based on a similarity function, Ψ, that identifies the overlaps within observed and learned patterns and uses the overlaps to solve new tasks. More precisely, to compute a situation's familiarity Φ_{fam}, Ψ takes in a set of perceived or imaginatively created patterns φ, and the *relevant* patterns k retrieved from knowledge base KB, where $k \subset KB$, meaning that $\Phi_{fam} = \Psi(k, \varphi)$. The similarity between k and φ is found by the ratio of

patterns' intersection ($k \cap \varphi$) to all patterns involved in the comparison ($k \cup \varphi$).

$$\Phi_{fam} = \Psi(k, \varphi) = \frac{|k \cap \varphi|}{|k \cup \varphi|} \tag{2}$$

Equation (2) ensures that Φ_{fam} has a value between 0 and 1, where 1 means completely familiar ($k \cap \varphi = k \cup \varphi$) and 0 represents maximal difference ($k \cap \varphi = \varnothing$). The agent must know how important the comparisons are, as the patterns being compared must be relevant to the agent's goal achievement. The top-down reasoning via backward chaining initially yields *important patterns* P_I, where P_I is a subset of patterns in the current or predicted situations.

Analogy, induction, and knowledge pruning

Analogy process estimates the familiarity with situations and induces new CRMs accordingly [10]. It also performs knowledge pruning by gradually focusing on smaller subsets of patterns in known CRMs, leading to improved selective attention. It is an inductive process that relies on abduction for goal-driven comparison and deduction for model verification. The induction follows the common sense rule that *under similar conditions, similar actions/events lead to similar outcomes*. Assuming an agent knows M (equation (1)) upfront, it hypothesizes that command cmd^* under similar conditions (P^*) leads to a similar outcome (B^*).

$$M^* : [(P^*(t_1), cmd^*(t_1)) \quad B^*(t_2)] \qquad cfd^* < cfd \tag{3}$$

where M^* is hypothesized only if it is required for goal achievement. For hypothesizing M^*, the analogy process compares the LHS and RHS patterns of M with the patterns of the current/predicted situations and goals/subgoals, respectively, as shown in Fig.1. Left. The process starts with comparing the goal state with $B(t_1)$ (the RHS of M). If the goal pattern matches pattern B, the second comparison occurs by finding the matches between P (the LHS of M) and the current or predicted situation S. If and only if P partially matches S, a new model M^* with similar patterns, P^*, cmd^*, and B^*, is created. Creating M^* leads to adjusting the known model M patterns so they match the situations, a hypothesis worth testing through direct intervention. As illustrated in Fig. 1.Left, M^*'s prediction is verified via forward chaining before issuing *cmd**.

Confidence Computation

The familiarity equation (2) determines the induced M^*'s confidence, as M^* will be utilized for predictions about future states. The function takes in important known patterns P_I and patterns of the situation S, leading to the computation of Φ_{fam} and subsequently cfd^*. In other words, the more similar the patterns of P to S are, the higher the M^*'s confidence, cfd^*, will be. The cfd^*'s value also depends on the original model's confidence (cfd), as M^* is derived from M. Therefore, the confidence cfd^* is calculated as follows:

$$cfd^* = \Psi(S, P_I) \cdot cfd \tag{4}$$

meaning that $cfd^* \leq cfd$ holds. Issuing the command cmd^* collects positive or negative evidence for M^*, which can increase or decrease cfd^*.

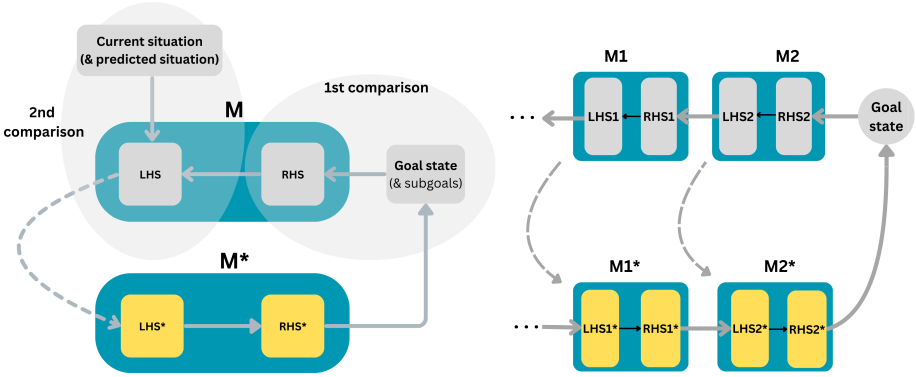

Fig. 1. Left: The analogy starts by comparing the RHS of M with the goal state. If they match, it triggers the second comparison: comparing LHS and situation patterns. As a result, M^* will be created and tested via deduction, verifying if LHS^* matches the situation and RHS^* matches the goal. **Right**: $M1$ and $M1^*$ are chained to $M2$ and $M2^*$ respectively. The requirement for chaining is that $LHS2$ and $LHS2^*$ match $RHS1$ and $RHS1^*$, respectively. The process first generates $M2^*$ as it moves backward and then generates $M1^*$. The same comparisons occur when generating $M1^*$ and $M2^*$.

Generalizing chains of models
Generalizing chains of CRMs is based on the theory that a network of relations between processes must be kept in analogy [4]. In our work, a network of relations is CRM chains generated through planning, i.e., backward and forward chaining. The analogy process, other than pattern matching with the current situation, must also identify the matches with *predicted* situations (a.k.a., subgoals). Situations can be predicted by known CRMs whose RHS patterns match the important patterns of the current situation. For example, as shown in Fig. 1. Right, model $M1 : [LHS1 \quad RHS1]$ can only be chained to $M2 : [LHS2 \quad RHS2)]$, if and only if, $RHS1$ matches $LHS2$. If $LHS2$ represents patterns holding in the current situation, then M_1 as well as M_2 can be pruned and accordingly ${M_1}^*$ and ${M_2}^*$ be induced. Due to such a generalization process, new hypothesized CRMs are combined to create more solutions, providing higher flexibility in planning.

6 AERA & Analogy

The analogy mechanism is integrated into the current implementation of autocatalytic endogenous reflective architecture (AERA) [7], called OpenAERA,[3].

Knowledge Representation & Reasoning
OpenAERA's knowledge representation relies on the introduced notion of CRMs and has the following components [7, 10]:

[3] See http://www.openaera.org — *accessed Apr. 2nd, 2024.*

- ***Entities & ontologies*** are symbols showing task elements and properties.
- ***Predicates*** are I/O facts, goals, and predictions.
  E.g. *((h essence hand) 1 t_0)* represents a fact with entity h having ontology property *essence* of *hand* and confidence 1 at time interval t_0.
- ***Causal models (CMs)*** are transformations from current to future states.
  E.g. *M:[cmd grab(h, t_0) h holding X(t_1)]* represents the causal influence from the *grab* command at t_0 on the state of h at t_1 (h will be holding X).
- ***Composite states (CSTs)*** a set of abstracted, simultaneous, related facts determining the conditions under which a CM holds true. CSTs are instantiated in forward chaining when inputs match CST patterns or during backward chaining from goals. E.g. *CST_1:[((h position P) 1 t_0), ((c position P) 1 t_0)]*, representing the entities h and c both being at identical position P.
- ***Requirement models (M_{req}s)*** connects CSTs and CMs via their instantiation. E.g. *M_{req}:[icst CST_1(c,h,p_0) imdl M(c)]*, where *icst* and *imdl* instantiate CST_1 on its LHS and M on its RHS, respectively.

In OpenAERA, ***forward chaining (deduction)*** occurs when input facts instantiate a CST on the LHS of a M_{req}, leading to the instantiation of the CM on its RHS and making a prediction. ***Backward chaining (abduction)*** happens if a goal matches the RHS of a M_{req}, causing the CST on the M_{req}'s LHS being instantiated, creating subgoals. If a command causes a state change, a triad of CST, M_{req}, and CM is learned, called ***induction*** in OpenAERA.

Extension via analogy
In OpenAERA, reasoning only occurs when situations align perfectly with known model preconditions, allowing the system to create plans for goal achievement. However, expecting a complete match between situations and known model preconditions is not always a practical assumption. OpenAERA uses the introduced analogy process to generalize the restrictive preconditions (CSTs and M_{req}s) of its causal models (CMs) and apply them to a broader range of situations. It uses the existing backward chaining mechanism from the top-level goal to the subgoals and then from subgoals to other subgoals, generalizing chains of involved CMs by generating new CSTs and M_{req}s for them. The newly generated CST-M_{req}s are immediately injected into the OpenAERA's model base and utilized in planning based on confidence values determined via familiarity computations.

7 Results & Evaluation

To evaluate the analogy mechanism, we design a pick-and-place task for a robot arm that uses the extended OpenAERA as its controller. The task's purpose is to analyze how the OpenAERA agent autonomously extends its learned cause-effect models to grasp and pick partially familiar objects. The task is performed within Webots simulation environment [5], as illustrated in Fig. 2. In this task, the OpenAERA agent, after inducing new preconditions (CST-M_{req}s) of known CMs via analogy and testing them by action-taking, learns the size of objects

is significant in determining the correct way of grasping them. The link to the demo can be found in [4]. The details of the experiment are shown in Fig. 2.

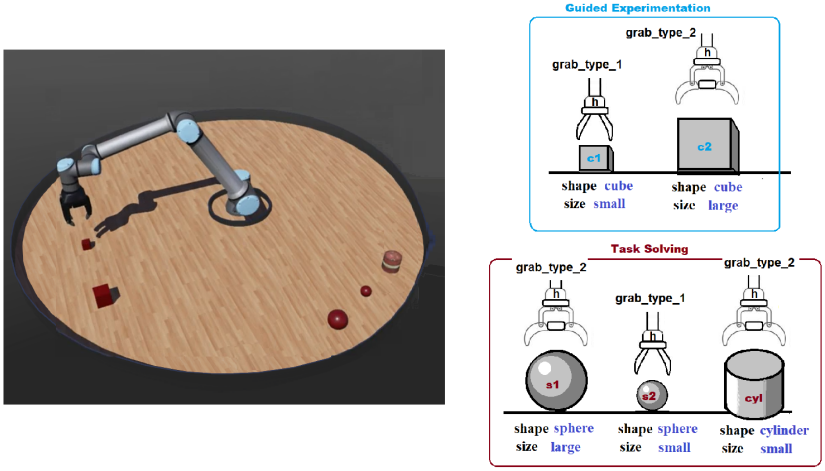

Fig. 2. The objects have shape, size, and position properties. The OpenAERA agent learns that the size of objects is significant in determining the correct way to grasp them. Positions are not shown as they are not involved in the analogy process.

The task has two phases: *guided experimentation* and *task solving*. In its *guided experimentation* phase, two cubes of different sizes, $c1$ and $c2$, are grasped and released by a robot hand h taught to apply two distinct grasping commands suited to the objects' sizes. For the small cube (c1), it applies the command $grab\_type\_1$, which involves a slight opening of the gripper's fingers. For the larger cube (c2), it uses the command $grab\_type\_2$, which requires a wider opening of the gripper fingers. Issuing the commands and successful grasping events makes OpenAERA learn triads of causal models (CMs), composite states (CSTs), and requirement models (M_{req}s) for each of the mentioned commands, one of which is shown in Fig. 3. Left. During the *task-solving* phase, h must learn to grasp new objects it has never grasped before, namely $s1$, $s2$, and cyl, which are novel in shape but familiar in size, allowing the analogy process to induce new grasping models based on the objects' familiarity. More precisely, the analogy process induces new preconditions CSTs and M_{req}s shown in Fig. 3. Right for the CMs learned in the guided experimentation phase e.g., mdl_514.

The analogy process provides multiple improvements for OpenAERA's induction, abduction, and selective attention by systematic knowledge pruning.
Knowledge pruning via analogy. In analogy-induced cst_581 and mdl_582, the shape (i.e., being *cube*) is pruned, but the sizes (e.g., being *large*) are kept; as in the task-solving phase, shapes are novel, but sizes are familiar.

[4] https://youtu.be/JXgdSjU-7OI — *accessed on Mar. 29th, 2024.*

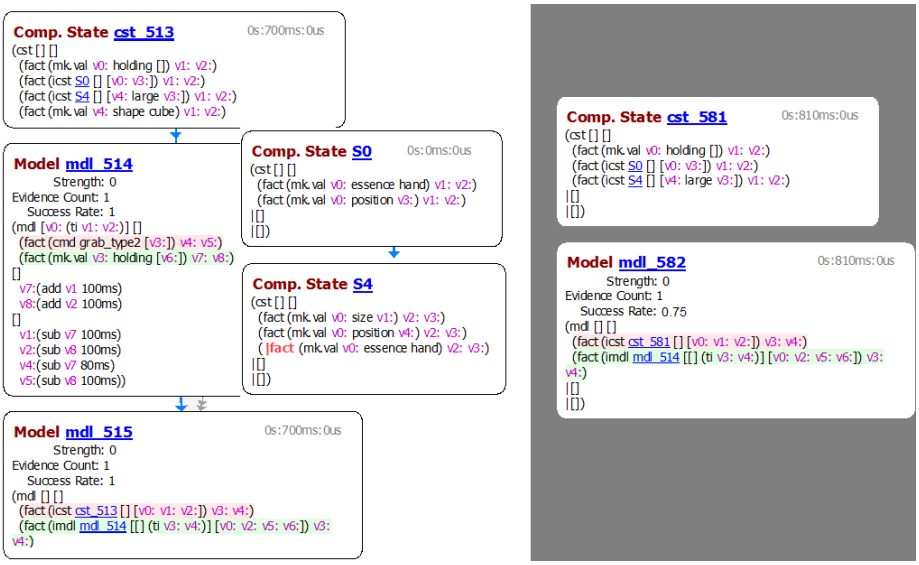

Fig. 3. Left: The composite state *cst*_513 shows the conditions under which the causal model *mdl*_514 holds. The requirement model *mdl*_515 instantiates *cst*_513 on its LHS (highlighted by red), which leads to the instantiation of the *mdl*_514 on its RHS (highlighted by green). This triad is learned by teaching the arm in the guided experimentation phase. **Right**: The composite state *cst*_581 and the requirement model *mdl*_582 are created via the analogy process in the task-solving phase. The figure provides snapshots of OpenAERA's Visualizer software, showing pieces of the results.

Abduction flexibility. Model *mdl*_514 and its initial preconditions, *cst*_513 and *mdl*_515, can be used by the OpenAERA's standard planner only if the target object for grabbing by *grab_type*_2 has a *cube shape*, as constrained by *cst*_513. The analogy process, however, prunes the initial preconditions, induces new preconditions, and injects them into OpenAERA's model-base such that they can be immediately used within the same planning process. In the task-solving phase, the target object $s1$ has a *sphere shape*, which does not match *cst*_513. Yet, $s1$ has another property that matches *cst*_513 - it has *large size*. The analogy process prunes the non-matching fact of *cst*_513, generates a new composite state *cst*_581 and a new requirement model *mdl*_582, shown in Fig. 3. Right, and immediately uses them for grasping the new objects, e.g., s_1.

Familiarity and selective attention. When inducing *cst*_581, the ratio of matching facts of the original composite state *cst*_513 to all components is 3/4 or 0.75, which specifies the success rate (a.k.a. confidence) of the newly created requirement model *mdl*_582 that instantiates *cst*_581 and *mdl*_514. Note that the new requirement model *mdl*_582's success rate increases when grabbing experience succeeds, as the model collects more positive evidence, helping the agent selectively pay attention to similarities that are helpful in goal achievement.

8 Conclusions

We have presented an implemented goal-driven analogy mechanism for generalizing causal knowledge in autonomous cognitive agents. The mechanism has been subjected to empirical validation in robotic experiments in the AERA system operating in a 3D robot simulation environment. The results demonstrate the effectiveness of the proposed approach. Future work calls for exploring the mechanism's scalability and efficacy in real-world applications.

Acknowledgements This work was supported in part by Reykjavik University, Icelandic Institute for Intelligent Machines, and Cisco Systems Inc.

References

1. A. Aamodt and E. Plaza. Case-based reasoning: Foundational issues, methodological variations, and system approaches. *AI communications*, 7(1):39–59, 1994.
2. G. L. Drescher. *Made-up minds: a constructivist approach to artificial intelligence*. MIT press, 1991.
3. L. M. Eberding, A. Sheikhlar, and K. R. Thórisson. Comparison of machine learners on an ABA experiment format of the cart-pole task. *Proceedings of Machine Learning Research, International Workshop on Self-Supervised Learning*, 159:49–63, 2022.
4. B. Falkenhainer, K. D. Forbus, and D. Gentner. The structure-mapping engine: Algorithm and examples. *Artificial intelligence*, 41(1):1–63, 1989.
5. O. Michel. Cyberbotics ltd. webots™: professional mobile robot simulation. *International Journal of Advanced Robotic Systems*, 1(1):5, 2004.
6. M. Mitchell. Abstraction and analogy-making in artificial intelligence. *Annals of the New York Academy of Sciences*, 1505(1):79–101, 2021.
7. E. Nivel, K. R. Thórisson, B. R. Steunebrink, H. Dindo, G. Pezzulo, M. Rodriguez, C. Hernández, D. Ognibene, J. Schmidhuber, R. Sanz, et al. Bounded recursive self-improvement. *arXiv preprint arXiv:1312.6764*, 2013.
8. J. Pearl and D. Mackenzie. *The book of why: the new science of cause and effect*. Basic books, 2018.
9. A. Sheikhlar, K. R. Thórisson, and L. M. Eberding. Autonomous cumulative transfer learning. In *International Conference on Artificial General Intelligence*, pages 306–316. Springer, 2020.
10. A. Sheikhlar, K. R. Thórisson, and J. Thompson. Explicit general analogy for autonomous transversal learning. In *International Workshop on Self-Supervised Learning*, pages 48–62. PMLR, 2022.
11. K. R. Thórisson. A new constructivist AI: from manual methods to self-constructive systems. In *Theoretical Foundations of Artificial General Intelligence*, pages 145–171. Springer, 2012.
12. K. R. Thórisson. Seed-programmed autonomous general learning. In *Proceedings of Machine Learning Research*, volume 131, pages 32–70, 2021.
13. P. Wang. *Non-axiomatic reasoning system: Exploring the essence of intelligence*. Indiana University, 1995.
14. F. Zhuang, Z. Qi, K. Duan, D. Xi, Y. Zhu, H. Zhu, H. Xiong, and Q. He. A comprehensive survey on transfer learning. *Proceedings of the IEEE*, 109(1):43–76, 2020.