Recombinant building: the ability to generate and recombine navigation structures is hard to acquire using just reinforcement learning

Ganesh Shinde[a,b] (ganeshinde@outlook.com), Harshit Agrawal[a] (harshitagrawal.iitr@gmail.com), Sanjay Chandrasekharan[a] (sanjay@hbcse.tifr.res.in)

[a]The Learning Sciences Research Group, Homi Bhabha Centre for Science Education, Tata Institute of Fundamental Research, Mumbai, India

[b]Centre for Modeling and Simulation, Savitribai Phule University, Pune, India

Abstract

Humans build novel tools, external knowledge structures (markers, maps etc.), and internal structures (analogies, mental models etc.) to facilitate cognition. Humans also recombine these building strategies to suit any task. Other organisms generate such structures as well, but they use them to optimize single tasks. This suggests that the human species' cognitive advantage stems from the capability to recombine built structures, and the resulting extended mind. Chandrasekharan & Stewart (2007) hypothesized that this capacity could emerge from reinforcement learning. We tested this proposal, by studying three foraging models, which examined whether novel recombinations of building (external and internal navigation structures) emerged in reactive agents, from just reinforcement learning. Results showed that recombination does not emerge with just reinforcement. This was because the building of external structures provided a very high reward profile, including free riding, thus acting as an attractor, blocking the recombination strategy. We discuss the implications of these results.

Keywords: Recombination, Distributed Cognition, Extended Mind, Agent-based Modeling

Introduction

The human mind is a rare adaptation (Donald, 1998). Its rarity stems from the following four building (Chandrasekharan, 2009; Chandrasekharan & Nersessian, 2015) and incorporation (Maravita & Iriki, 2004; Chandrasekharan, 2014) strategies, which together set apart human cognitive practices from that of other organisms.

- The capacity to build external physical structures (including tools), and incorporate (i.e. integrate with the body) such structures made by others.
- The capacity to build external knowledge structures (including markers, maps and language), and incorporate (i.e. integrate with internal structures) such structures made by others.
- The capacity to build new internal structures (such as landmark grids, categories, analogies, and models) based on interaction with tools and external knowledge structures, and incorporate (i.e. integrate with internal structures) such internal structures when they are externalized by others.
- The capacity to *recombine* these building and incorporation strategies, to perform better in *any* task.

The first three strategies are present to some extent in other organisms (Laland, Odling-Smee, & Feldman, 2000), including insects (Mhatre, 20018; Mhatre & Daniels, 2018; Bradbury & Vehrencamp, 1998; Sanders et al, 2015). However, the building processes are usually limited to particular tasks (such as building of nests, rudimentary tools, signaling structures and internal landmarks).

The presence of these strategies in other organisms, and the absence of human-like cognitive practices in these and other organisms, together suggests that the rare mind that sets humans apart emerges from the recombination strategy, which allows the first three strategies to be combined fluidly, across any task (Vygotsky, 1980).

The first strategy has mostly been studied in anthropology (Tomasello, 2009), though there are some recent neuropsychological (Putt et al, 2017) and behavioral studies (Emery & Clayton, 2009) as well. The second strategy has been studied empirically to some extent within the Distributed Cognition framework (Hutchins, 1995; Kirsh, 2010; Chandrasekharan, 2009), and analysed conceptually within the Extended Cognition framework (Clark & Chalmers, 1998; Menary, 2010; Adams & Aizawa, 2008). There is significant empirical and analytical work on the third strategy (Pulvermuller, 2001; Nersessian, 2010).

However, very little known is about the recombination strategy, particularly how it emerges, and the cognitive/neural machinery that is involved. This is mostly because the process of building external physical and knowledge structures is poorly understood. As Schwartz and Martin (2006) observes, "most cognitive research has been silent about the signature capacity of humans for altering the structure of their social and physical environment." Even in the study of internal structures, most studies have focused on their use, rather than the process by which they are built.

Chandrasekharan & Stewart (C&S, 2007) presented a preliminary model of the process leading up to the building of external knowledge structures. They showed that reinforcement learning could lead to the emergence of a stable building process, where agents learn to systematically generate external knowledge structures (termed 'epistemic structures') within their lifetime. In this model, reactive agents (agents that only sense and act) learn to systematically build epistemic structures to optimize food foraging behavior (ES model, figure 1). Each agent could initially drop two chemicals (in analogy to pheromones) randomly (in analogy to perspiration), during a food-gathering task. They could also follow the dropped pheromones in some random instances. Pheromones could

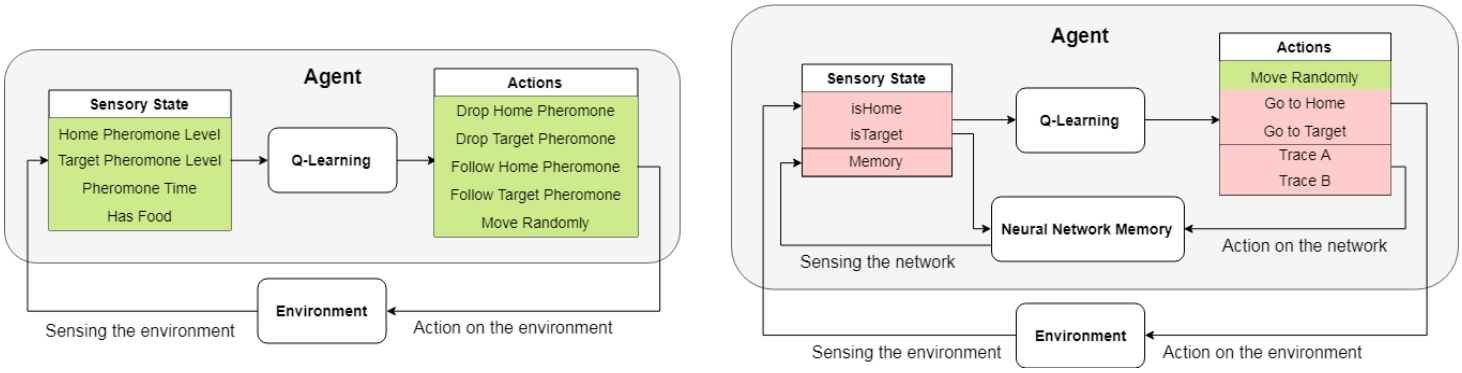

Figure 1: Model architecture used by C&S, for the Epistemic Structure (ES) model (Left) and the Trace model (Right).

be dropped only when an agent was still, and every action an agent made (dropping pheromones, following pheromones, wandering) had a energy/tiredness cost. Given this basic structure, and a reinforcement learning algorithm (Q-learning) that selected agent actions to lower the energy cost, the agents learned over time to systematically build pheromone structures, and maximize food gathering based on this building behavior.

C&S then extended this basic building structure (random behavior moving to systematic building, based on feedback about energy use), to develop a Trace model (figure 1), which examined whether reactive agents, with the capability to drop two random traces (A,B) in an internal neural network, could learn to generate these traces systematically in relation to two landmarks (Home and Target). This type of systematic generation of traces could be considered equivalent to the building of a rudimentary version of a place cell (Sanders et al, 2015). Results showed that the Q-learning algorithm allowed the agent to generate such a place-cell-like structure, by systematically 'grounding' the initially random internal traces to the landmarks (A only at Home and B only at Target, or vice versa), and maximize food gathering based on this internal structure.

Recombinant Building

These models showed that systematic building of ESs and Traces could emerge from random behavior, based on the same underlying reinforcement learning mechanism. C&S then hypothesized that since the underlying learning model is the same, the two strategies could easily be recombined, and reinforcement learning could thus account for the emergence of extended minds.

However, extended minds do not exist widely, as would be expected if reinforcement learning is enough to generate such minds. We therefore tested this hypothesis, by developing a Recombinant Building model (Figure 2). This model is based on the basic structure of the ES model reported by C&S. To test the possibility of recombinance, we added the Trace model's capabilities to the ES model. That is, apart from the ES actions, agents could also drop two types of traces randomly in an internal network.

Our hypothesis was: reinforcement based on tiredness would lead to the emergence of a recombinant building strategy, where internal traces and external structures would be recombined systematically, thus developing a more optimal building strategy than the ES one, to solve the food-gathering task. One example of a recombinant building behavior would be reaching Home and recognizing Home (based on the systematic activation of the Home trace in this location), and then on dropping only target pheromones, until Target is reached. For the Target location, the equivalent recombination would be recognizing Target, and then on dropping only home pheromones, until Home is reached. Another example would be storing a trace of the pheromone level in each location, and then lowering/raising the generation of pheromones, based on this trace.

This operationalisation of the recombinant building hypothesis was investigated using four studies.

Study 1 was a baseline, replicating the original ES study (Figure 4). The only minor change was in the exploration rate, which was set to decrease steadily. Also, each simulation (100 trials) ran for a million time steps, instead of the maximum of 3000 reported in the C&S study.

In study 2, we added to each agent the ability to drop two traces in an internal neural network. The structure of this internal network trace was exactly the same as reported in the C&S model.

In study 3, instead of storing traces of Home and Target, which are stable states of the world, the agents could store traces of the levels of Home Pheromone and Target Pheromone, which are transient states of the world.

In study 4, agents could store traces of both permanent and transient states of the world (Home, Target and the levels of Home Pheromone and Target Pheromone).

The architecture of the system and the main results are outlined in the sections below. For more details see the code (recombinant.surge.sh). See the C&S paper for other details of model design and their rationale.

Model Architecture

World: Following C&S, the task was analogous to foraging behavior (i.e. navigating from a home location to a target

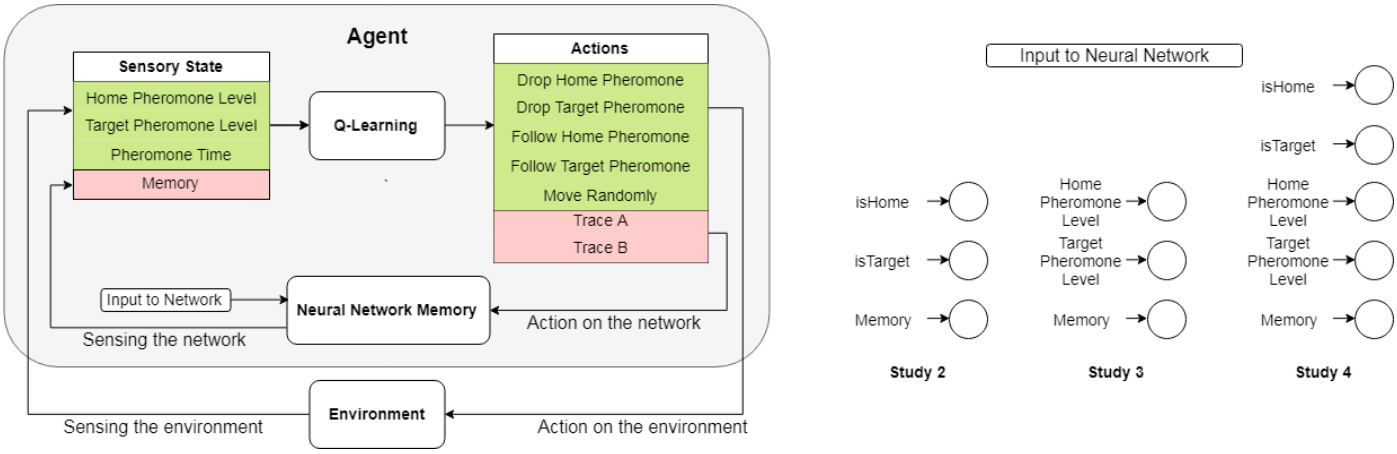

Figure 2: (Left) Recombinant Building architecture. (Right) Inputs to the network for the studies.

location and back again). The environment consisted of a 30 × 30 toroidal (doughnut-shaped) grid world, with one 3 × 3 square patch representing the agent's home, and another representing the target (as depicted in figure 3). This target was the food source.

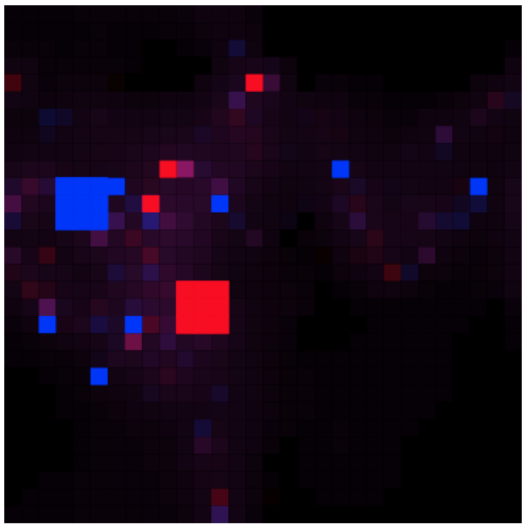

Figure 3: The world grid. The big blue square is the Target, the big red square is the Home. The small bright red squares are agents without food. The small bright blue squares are agents with food. The faded red color indicates dropped home pheromones, and the faded blue color indicates the dropped target pheromones.

Reinforcement Learning: Q-learning was used as the reinforcement learning algorithm. It was present in every agent, and sought to minimize individual energy expenditure while foraging. The space of states and actions for study 1 is shown in figure 1 (ES architecture). Figure 2 shows the same for studies 2, 3 and 4. The different inputs to the neural network for these studies are also shown.

Neural Network: The neural network was a feedforward multilayer perceptron, trained using back-propagation of error (Rumelhart, Hinton, & Williams, 1986). It had three input nodes, one output node, and three hidden nodes for studies 2&3. In study 4, where we combined the transient and permanent features, the network had 5 input nodes, 1 output node, and 7 hidden nodes. The activation function for all nodes was the hyperbolic tangent, and the learning rate (α) was 0.2. To handle the feedback between the output value and the input state, the network was run 100 times.

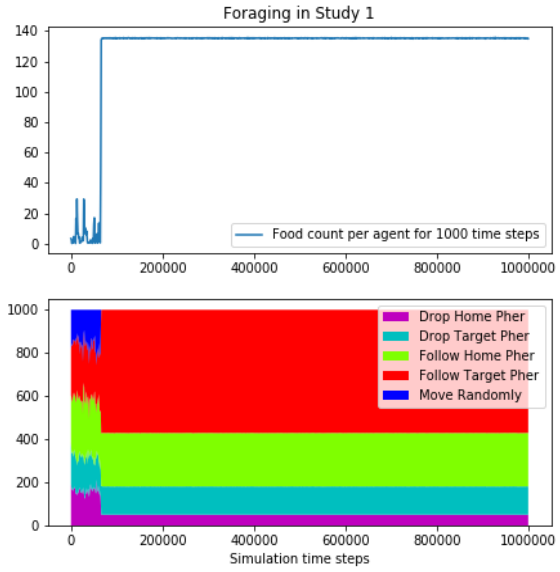

Figure 4: Food count and average action distribution (per 1000 time steps) for a trial in study 1.

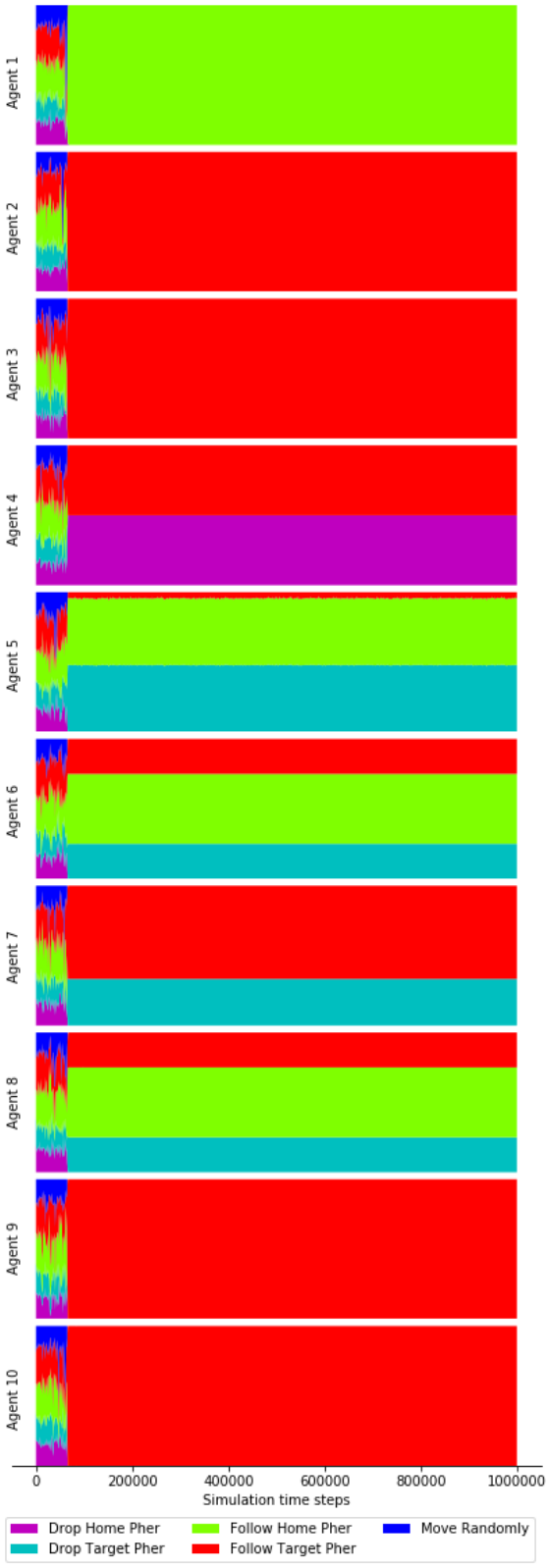

Figure 5: Action distribution of agents (per 1000 time steps) in a trial of study 1. Notice some agents only follow pheromones. They are 'free riders'.

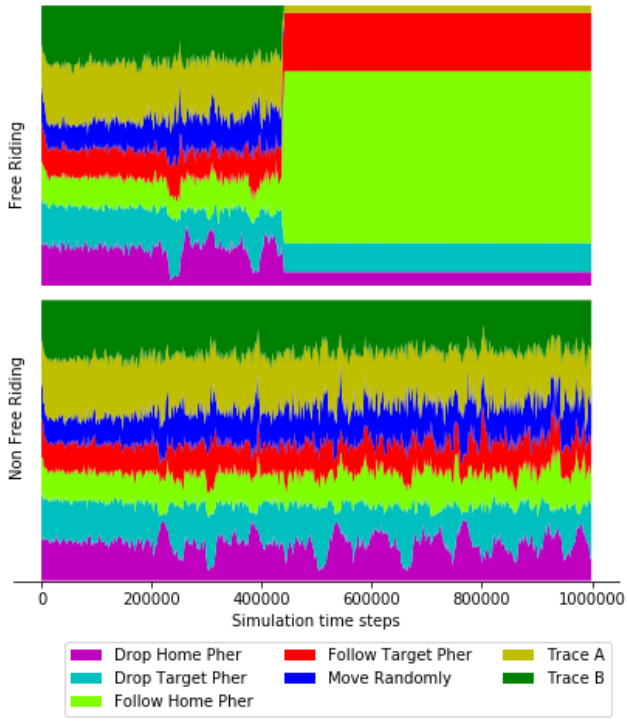

Figure 6: Average action distribution for study 4. The top one shows the average action distribution in a trial where free-riding emerged. The graph below shows the average action distribution during a trail when free riding did not emerge.

Results

Study 1

One hundred trials (each with one million time steps) were run, with the exploration rate of Q-learning reducing by 1% every 100 time steps. Remarkably, in almost all trials, the system improved food gathering performance after around 10000 time steps, by discovering a second optimization after the ES one -- *free riding* (see figure 4 for one such instance). All hundred trials converged to the free-riding strategy.

The free-riding strategy, which is very stable once discovered, is marked by a consistent pattern, where some agents only follow pheromones -- they don't drop any pheromones (figure 5). They thus free-ride on the work of agents that do the pheromone dropping actions. More interestingly, across agents, the pheromone following actions (follow home, follow target) dominate, forming the largest action component. Dropping actions (drop home pheromone, drop target pheromone) are present in relatively smaller proportions. There are thus two optimizations: one where individual agents move to just following pheromones, and secondly, the colony as a whole lower dropping actions in general. This 'super-optimization' result was missed by C&S, as they ran less time steps.

Once this action pattern sets in, the food collection increases dramatically across agents. This is because the dropping actions, which require energy and staying still, are less overall. The dropping actions do not lead the agent

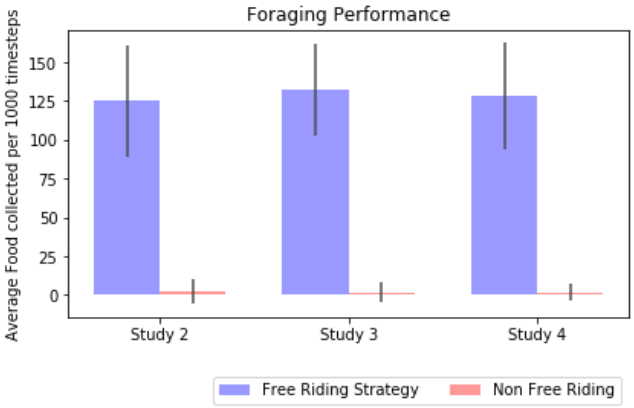

Figure 7: Foraging performance when free riding was present and absent.

closer to the Home or Target. As would be expected, individual free riders perform better than agents who drop pheromones, as they make more movement actions than their dropping counterparts, who stay still while dropping.

Study 2,3&4

As discussed, these studies examined the recombination hypothesis, by providing agents with both external and internal actions. Each study examined different internal actions, where the inputs given to the internal neural networks were different (see figure 2). One hundred trials (each with one million time steps) were run for each study.

Results showed that in each study, either the system did not perform well, or it settled into the free riding strategy, where the performance was at par with study 1. Figure 6 shows an example of both these cases.

In most of trials where the system settled into free riding, only the dropping and following actions were present. Table 1 provides an overview of the emergence of the free-riding strategy across these studies. In a few cases, dropping internal traces and move randomly actions were also present, usually in only one agent among the 10 agents. Figure 6 shows such a case where dropping of trace A was also present. Figure 7 shows a comparison of foraging performance (food collected per 1000 time steps) across time where free riding was present vs. where it was not. Table 1 also logs this performance.

Table 1 (per 1000 time steps)

Name	Iterations where free riders were present	Free riding Performance Avg(std)	Non-free riding Performance Avg(std)
Study 2	57	125(35)	2.5(8)
Study 3	55	132(30)	1.7(6)
Study 4	29	128(34)	1.6(5)

Discussion

The consistent results across the studies indicate three major trends: free-riding, the attractor role of external building, and the absence of recombination. We discuss each below.

Free-riding (FR): Across the three recombination cases, most agents moved to the ES strategy, in most trials. In every trial where this shift happened, FR emerged after around 10000 trials. This suggests Q-learning has the capability to do super-optimizations over time. However, a Tragedy of the Commons (Hardin, 1968) situation, where every agent turns into a free rider to maximize individual returns and thereby destroy a resource (the pheromone trails in this case), did not emerge. These results indicate that Q-learning, even though embedded within each individual agent, managed to discover, and maintain, a globally optimal strategy.

FR is well-documented in biology, and the emergence of stability despite FR is usually attributed to mechanisms wider than the organism, such as kinship, policing and diminishing returns (Rankin, Bargum, & Kokko, 2007). Our results suggest that a within-organism feature -- the estimates of reward made by reinforcement learning -- could also lead to stable behavior when FR is present. Overall, the FR result provides significant support to the ES model presented by C&S, as it is close to behavior seen in nature, and emerged inadvertently, based on a mechanism different from commonly cited ones.

The attractor role of external building: Most trials saw the recombinant models moving predominantly to the external structure strategy. This suggests building stable internal structures is either difficult, or not optimal, when external building is also possible.

In contrast to this result, internal structures such as place cells and grid cells are more widespread in nature than built trails, possibly because built external structures are transient. Building internal structures like grid cells allow organisms to navigate using more stable properties of the world (like landmarks), and also navigate in air and water, where trails cannot be built and maintained easily.

Our results suggests that such internal structures are generated only in cases where building external structures is not an option, as the possibility of building external structures would lead to the learning system getting attracted to this strategy, given its better reward profile. A corollary is that organisms which navigate mostly using built external structures (usually lower organisms) could be considered as developing a niche that requires, and thus leads to, minimal neural development, as the building of external structures limit the emergence of complex landmark-based navigation, and related neural complexity.

Even though the C&S study showed that rudimentary landmarks could emerge from reinforcement learning, it is unclear whether a wider navigation system, where many landmarks are stored using complex grid-cell-like structures, could emerge just from reinforcement learning. It is possible

that mechanisms other than reinforcement learning are recruited in the development of complex internal storage systems such as grid cells.

A candidate mechanism could be learning based on 'offline' simulation of stored events (Schubotz, 2007). Such simulation is possible through the re-activation of the traces stored in the grid cells (Buzsáki & Moser, 2013). Interestingly, the C&S model shows that this simulation capability could emerge from just reinforcement learning. This is because the storage of internal memories emerge solely from actions in the C&S model. Such memories are thus 'constituted' by actions, and they thus embed action possibilities, which can be simulated offline, to generate feedback for reinforcement.

Absence of Recombination: Given the starting premise (the absence of human-like cognition in other organisms), and the inability of the models in our study to recombine the two building strategies (probably due to the attractor role played by the ES strategy), a possible interpretation could be that mechanisms other than reinforcement (such as simulation, metacognition etc.) are needed for the emergence of the recombination strategy. However, such an interpretation is not justified at this point, for three reasons.

First, our implementation expected the agents to learn both the building strategies, as well as recombine the strategies, in real-time. This is unrealistic, as organisms have the possibility of stabilizing each building process (external storage, internal storage) first, and then recombining them.

Second, more sophisticated reinforcement learning algorithms, such as Deep Q-learning, could allow the model to discover the recombination strategy. We are currently testing this possibility.

Finally, the models reported here assume there is only one reinforcement learning mechanism. This is also unrealistic, as each building process could have its own reinforcement, and the two reinforcement systems could be working in tandem, with positive feedback loops between them. It is currently unclear how such a system could be implemented.

References

Adams, F., & Aizawa, K. (2008). *The bounds of cognition*. John Wiley & Sons.

Bradbury, J. W., & Vehrencamp, S. L. (1998). *Principles of animal communication*. Sunderland, MA: Sinauer Associates.

Buzsáki, G., & Moser, E. I. (2013). Memory, navigation and theta rhythm in the hippocampal-entorhinal system. *Nature Neuroscience*, 16(2), 130.

Chandrasekharan, S., & Stewart, T. C. (2007). The origin of epistemic structures and proto-representations. *Adaptive Behavior*, 15(3), 329–353.

Chandrasekharan, S. (2009). Building to discover: a common coding model. *Cognitive Science*, 33 (6), 1059-1086.

Chandrasekharan, S. (2014). Becoming Knowledge: Cognitive and Neural Mechanisms That Support Scientific Intuition. In Osbeck, L.M. & Held., B.S. (Eds.), *Rational Intuition: Philosophical Roots, Scientific Investigations*, pp. 307-337, Cambridge University Press.

Chandrasekharan, S., Nersessian, N.J. (2015). Building Cognition: the Construction of Computational Representations for Scientific Discovery. *Cognitive Science*, 33, 267–272.

Clark, A., Chalmers, D.J. (1998). The Extended Mind, *Analysis*, 58(1), 7–19.

Donald, M. (2001). *A mind so rare: The evolution of human consciousness*. WW Norton & Company.

Hardin, G. (1968). The Tragedy of the Commons. *Science*, 162, 1243–1248.

Hutchins, E. (1995). How a cockpit remembers its speeds. *Cognitive Science*, 19, 265–288.

Kirsh, D. (2010). Thinking with external representations. *AI & Society*, 25(4), 441–454.

Maravita, A., & Iriki, A. (2004). Tools for the body (schema). *Trends in Cognitive Sciences*, 8(2), 79–86.

Menary, R. (2010). *The extended mind*. MIT Press, Cambridge,MA.

Mhatre, N. (2018). Tree cricket baffles are manufactured tools. *Ethology*, 124(9), 691-693.

Mhatre, N. & Daniel, R. (2018). The drivers of heuristic optimization in insect object manufacture and use. *Frontiers in Psychology*, 9, DOI: 10.3389/fpsyg.2018.01015

Rankin, D. J., Bargum, K., & Kokko, H. (2007). The tragedy of the commons in evolutionary biology. *Trends in Ecology & Evolution*, 22(12), 643-651.

Rumelhart, D. E., Hinton, G. E., & Williams, R. J. (1986). Learning representations by back-propagating errors. *Nature*, 323(6088), 533.

Schwartz, D. L., & Martin, T. (2006). Distributed learning and mutual adaptation. *Pragmatics & Cognition*, 14, 313–332.

Sanders, H., Rennó-Costa, C., Idiart, M., & Lisman, J. (2015). Grid cells and place cells: an integrated view of their navigational and memory function. *Trends in Neurosciences*, 38(12), 763-775.

Vygotsky, L. S. (1980). *Mind in society: The development of higher psychological processes*. Harvard University Press, Cambridge, MA.