SNN-Based Online Learning of Concepts and Action Laws in an Open World

Christel Grimaud, Dominique Longin, Andreas Herzig

SNN-Based Online Learning of Concepts and Action Laws in an Open World

Christel Grimaud, Dominique Longin, Andreas Herzig
Université de Toulouse, CNRS, Toulouse INP, UT3, IRIT, France

Abstract

We present the architecture of a fully autonomous, bio-inspired cognitive agent built around a spiking neural network (SNN) implementing the agent's semantic memory. This agent explores its universe and learns concepts of objects/situations and of its own actions in a one-shot manner. While object/situation concepts are unary, action concepts are triples made up of an initial situation, a motor activity, and an outcome. They embody the agent's knowledge of its universe's action laws. Both kinds of concepts have different degrees of generality. To make decisions the agent queries its semantic memory for the expected outcomes of envisaged actions and chooses the action to take on the basis of these predictions. Our experiments show that the agent handles new situations by appealing to previously learned general concepts and rapidly modifies its concepts to adapt to environment changes.

1. Introduction

The ability of a cognitive agent to act adequately in a given environment depends on its ability to predict how performing a given action will affect its current situation; in other words, it depends on the agent's knowledge of its environment's *action laws*. How artificial agents can acquire these laws and how these should be updated if their environment changes has proved a difficult question. In the case where the intended environment is *open*—that is, where the agent's designer cannot foresee all the situations the agent might encounter in the future—, providing a suitable set of action laws to the agent "by hand" is unfeasible. The only viable solution is that the agent continuously learns the relevant laws from experience, just as natural agents (humans and animals) do. Crucially, this learning process should allow for generalization over disparate experiences, so that the agent is able to behave appropriately in new situations. It should also allow for rapid modification of the learned action laws to accommodate environment changes.

The aim of the present paper is to show how this could be done. Its main hypothesis is that natural agents' ability to perform well in our open and changing world relies on the fact that they store their knowledge in the form of rapidly updatable concepts with various degrees of generality. Presumably, they first form concepts about the encountered objects/situations, and then use these as elements for composing concepts of actions. The latter constitutes their knowledge of their environment's action laws. Another important hypothesis is that natural agents' endless ability to learn new concepts despite having finite brains relies on some efficacious management of forgetting. Most likely, they selectively forget their less important knowledge (such as, e.g., details and unused memories) and reallocate the corresponding neurons to encode new useful knowledge. We suggest that artificial agents could take inspiration from these strategies and use some artificial neural network to learn and store concepts, and query this network to make predictions about the outcome of envisaged actions.

To test this idea, we here build an artificial hybrid agent with an SNN at its core. This agent lives in a very simple virtual world, composed of rooms which may be, or not, accessible (hence, knowable) to it. At first, the agent is confined to one single room and learns by itself how to act in it according to its own interests. Then at some point a door opens to a new room containing some never encountered before objects and situations. Yet, although these are new to the agent, some general laws are preserved from one room to the other. Our experiments show that having learned these laws in the first room allows the agent to act by and large properly in the second one, as soon as it enters it. They also show that the agent is able to learn new laws holding in the second room without catastrophic loss of previous knowledge. Finally, some changes are introduced in the second room, rendering some of the previously learned rules obsolete while some

new rules become true. Again, experiments show that the agent quickly updates its knowledge to account for these changes.

The paper is organized as follows. Section 2 discusses related work, Section 3 presents the agent and its universe, and Section 4 describes the neural network and its functioning. Section 5 describes the agent's general functioning, while Section 6 presents the experiments performed on the agent and discusses the results. Section 7 concludes and outlines future possible developments of the framework.

2. Related works

The research problem addressed in this work is autonomous online learning, generalization and updating of concepts and action laws in an open universe. The intended application is to use the learned concepts in reasoning and planning for autonomous robots. To our knowledge, no existing approach addresses this problem in all its dimensions, but these are investigated in separate research fields.

Continual Learning tackles the problem of lifelong knowledge acquisition [Wang et al., 2024; Lesort et al., 2020]. Its main challenge is to avoid catastrophic loss of previous knowledge when acquiring new knowledge; a secondary research axis is one-shot/few-shots learning, i.e., the ability to learn online from one or few examples [Wang et al., 2020]. However, current approaches mostly consider the learning of *tasks* (mostly image classification/recognition tasks, but also some more complex tasks such as playing games [Kirkpatrick et al., 2017]), not of concepts nor action laws. Furthermore, most of them rely on supervised learning and/or labelled training data, which is unsuitable for open world autonomous agents.

Concept Learning has mainly been studied in view of explainability [Gupta and Narayanan, 2024], mostly of classification models (e.g., [Koh et al., 2020]) but also of decision making in the context of reinforcement learning [Das et al., 2023; Zabounidis et al., 2023]. For this reason, many proposals are dedicated to learning a human-predefined set of concepts using some annotated data. In the field of Image Classification some approaches deal with the extraction of concepts from data [Wang et al., 2022; Ghorbani et al., 2019; Hase et al., 2019], but then these are extracted from labelled classes of images, which again is unsuitable for open world autonomous agents. The vast majority of these methods also disregards the hierarchical organization of concepts from particular to more general, and they generally do not address one-shot learning, online revision or updating of concepts.

Action Laws Learning has been studied from various perspectives. In Dynamic Epistemic Logic, [Bolander and Gierasimczuk, 2018] proposed a method to learn an action model through successive observations of transitions between states. But this method does not achieve generalization nor accommodate environment changes and only considers *universally applicable* actions (i.e., actions that can be executed in every logically possible state), a condition real-world actions rarely satisfy. In the field of Planning, [Bonet et al., 2019] showed how to learn abstract actions from a few carefully chosen instances of some general planning problem. But said instances come with their own set of ground actions which must be known beforehand, so this approach cannot be used in open worlds, where an agent needs to incrementally learn from experience.

Reinforcement Learning (RL) is concerned with agents learning by experience how to achieve a goal in an optimal manner. A number of variants have been developed, which all have in common that the agent learns through trials and errors by means of a reward system, and that what is finally learned is a *policy*, i.e., a function which takes a given state of the agent as an argument and returns an action (or a set of actions) to be taken. RL approaches have proved very successful in a wide range of domains, and for this reason they are quite popular. However, they struggle to adapt to environment changes and to revise the learned policies [Kirk et al., 2023; Farebrother et al., 2018]. This could be related with the fact that in natural agents, learning policies is rather a matter of procedural memory ("memory of procedures"), a memory system that typically requires a large number of learning trials [Tulving, 1985; Knowlton et al., 2017]. By contrast, the semantic memory, which stores facts about the world in the form of concepts, is a fast learning system that can learn or update a concept over one single experience. It thus seems that an artificial agent able to learn and update in real time a conceptual model of its environment and to use it to make its decisions would be able to quickly adapt to environment changes.

RL approaches also typically encounter difficulties in contexts where rewards are scarce or deceptive [Hare, 2019; Ocana et al., 2023], which is the case of most real world contexts. Although some solutions have been proposed (see notably [Ecoffet et al., 2021]), dealing with sparse rewards is still an active field of research in RL. But since concept learning does not depend on rewards, an agent able to learn concepts and to use them to make its decisions would be immune to the problem.

In the field of robotics, semantic information has long been recognized as instrumental in autonomous robots' decision making [Prasad and Ertel, 2020], but there have been very few attempts to make autonomous robots capable of learning semantic knowledge from experience. The most prominent are probably [Wang et al., 2016] and [Nasir et al., 2019], which both use fusion ART neural networks to encode concepts. However these neural networks rely on the extension of their architecture to learn new knowledge (i.e., new neurons are created to encode

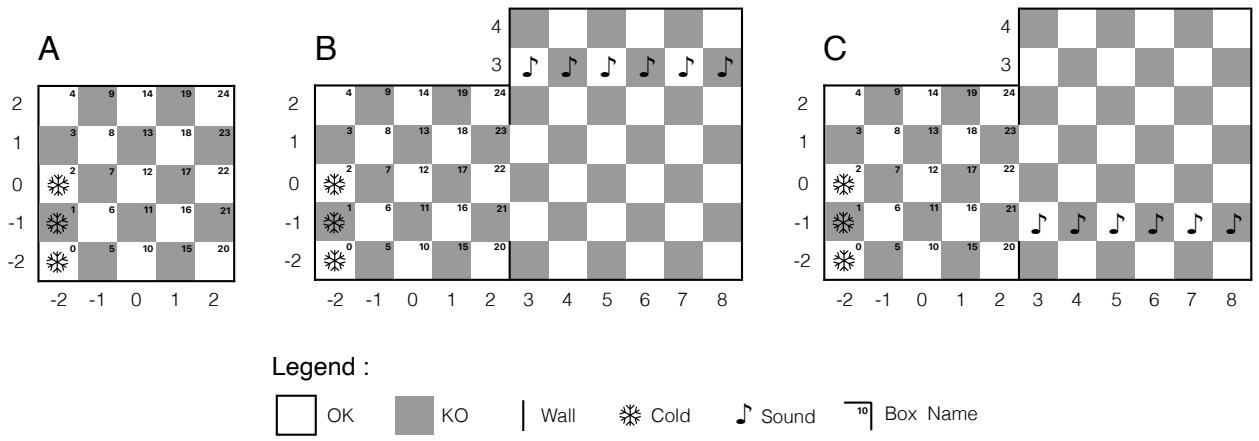

Figure 1. The agent's accessible world. A: in the first phase, Room 1 only; B: after opening the door, Rooms 1 and 2. C : after the sound feature is moved downwards (from row 3 to row -1) in Room 2.

new concepts), which is unsuitable for agents with finite memory resources living in an open world.

In open worlds, the finiteness of autonomous agents' memory makes forgetting unavoidable. For this reason, memorizing the whole of the experienced objects and events cannot be taken as an ultimate goal or criterion for such agents. Instead, forgetting should be accepted as a necessary counterpart of learning, and managed so as to preserve the agent's performance as much as possible. That is, what should be avoided is *catastrophic* forgetting. The most natural way to do this is by ensuring that the knowledge lost through learning is always the agent's less useful knowledge at the time of the learning. How to characterize "less useful knowledge" and how to identify the neurons supporting it in order to preferentially select them for encoding new knowledge is a key issue to be addressed in the sequel.

The present paper aims at providing a proof of concept of a cognitive agent satisfying the above requirements. As a first step, it focuses on the learning of concepts and only uses them in some basic decision making to demonstrate the agent's learning abilities. However the agent is designed so as to allow the retrieval of the learned concepts and action laws, which could then be encoded in symbolic format and used in more complex decision making involving explicit reasoning and action planning. The implementation of such complex reasoning abilities in the agent is left to future work.

In the absence of studies investigating the considered research problem in all its dimensions, comparing the proposed setup with existing approaches seems inappropriate. Indeed, it would only take into account some of the dimensions of the problem and fail to evaluate the proposal relative to its goal. Therefore, a suitable set of experiments was devised (see Section 6).

3. The agent and its universe

This section describes the universe the agent lives in and gives a general overview of the agent. It also clarifies the notions of concepts and action laws used in the sequel.

3.1. The universe

The universe is designed so as to support a set of experiments intended to assess the agent's abilities. It presents a number of challenges for the agent to face, while being kept as simple as possible to allow the precise monitoring of the agent's performances.

The universe is built over a grid of boxes, which we (not the agent) identify using an orthonormal coordinate system (see Figure 1). Each box represents a particular location in the agent's universe and possesses a particular set of features the agent is able to perceive, drawn from the set $LF = \{$*OK*, *KO*, *NorthWall*, *EastWall*, *SouthWall*, *WestWall*, *Cold*, *Sound*, *#0*, *#1*, ..., *#24*$\}$. For example, the box with coordinates $(-2, -2)$ has the feature set $LF_{(-2,-2)} = \{$*OK*, *SouthWall*, *WestWall*, *Cold*, *#0*$\}$, while the box with coordinates $(5, 0)$ has the feature set $LF_{(5,0)} = \{$*OK*$\}$. The expression "*#n*" is to be taken as a particular name for a box. It is therefore a feature, and not all boxes need to have one. Two boxes with the same feature set are indistinguishable for the agent.

The rooms are made out of boxes, and delimited with impassable walls. Opening a door amounts to removing the wall features from the corresponding boxes' feature sets (as in Figure 1.B, where the *EastWall* feature has been removed from boxes $(2, -1)$ to $(2, 1)$). More generally, changes in boxes' feature sets model environment changes. For example, at the time the door opens the boxes $(3, 3)$ to $(8, 3)$ all have *Sound* in their respective feature sets, but at some point the *Sound* feature moves to the boxes $(3, -1)$ to $(8, -1)$ (Figure 1.C).

When considering the agent's universe, at any point in time we only consider the boxes to which it has access, the set of which is always finite. Note that this does not contradict the unboundedness of the universe: although each possible room is finite (as are rooms in real world), the number of possible rooms the agent may discover in its life and the number of new objects it may encounter in them, as well as the number of changes that may occur in these rooms, are unbounded. Being able to handle to these novelties (react appropriately and update its knowledge on the fly) is the agent's main challenge.

As regards the features, *KO* corresponds to some unpleasant stimulus the agent spontaneously wants to avoid, and *OK* to the absence of such a stimulus. The other features convey some indifferent information. It should be stressed that *OK* and *KO* are not rewards in the sense of RL, as they play no role in the neural network's learning process (see section 4.2 for details). Their only purpose is to motivate the agent's choices and to allow us to assess its ability to make appropriate decisions.

The distribution of OK and KO boxes, which is common to both rooms, provides some general (non-monotonic) laws of the universe (such as *"going North-East from an OK box leads to another OK box"*), the learning of which is the agent's second challenge. Boxes' names in the first room are particular features. They are used to check that the agent also forms particular concepts of individual boxes and uses them to learn particular rules (such as *"going North-East from the OK box with name '#12' leads to the OK box with name '#18'*). *Cold* is a feature with an intermediate degree of generality. Being able to handle concepts with various degrees of generality is a third challenge. *Sound* is used to test the agent's ability to use the learned general rules in the presence of novelty, and to check that it can learn new concepts and action laws without suffering catastrophic forgetting. Notably, after the door opens (Figure 1.B) the agent is expected to learn new general rules such as *"going North from a box with sound leads to a box with a north wall"*. Moving the *Sound* feature from its "up" position (Figure 1.B) to its "down" position (Figure 1.C) is used to test the agent's ability to update its knowledge when the environment changes.

3.2. The agent

The agent is composed of a set of sensors, a perceptual system, a semantic memory, a decision system, a motor system and a set of actuators (see Figure 2). Sensors collect data from the external world and feed it to the perceptual system, which performs feature/object recognition. Neural networks doing this while relying on unsupervised learning from unlabelled data already exist (e.g., [Thiele et al., 2018; Kheradpisheh et al., 2017]), so we simply suppose that the agent's perceptual system operates as intended and provides the semantic memory with the appropriate inputs, namely, the correct and complete set of features of the agent's current location. Semantic memory forms concepts by binding together sets of features, and stores them for further retrieval. Its modeling is the main focus of the paper. The decision system is the other important part: it queries the semantic memory to predict the outcome of possible actions, and decides which one to take on the basis of these predictions. This decision is then sent to the motor system, which activates the actuators to perform the corresponding motor activity. Information from the actuators is sent back to semantic memory through proprioception, allowing the agent to memorize the motor-related features of the realized actions.

The agent's possible actions consist in steps from one box to another adjacent box, in any of the eight directions. Formally, an action is a triple made up of a depart location, a motor activity, and an outcome. By "motor activity" we mean the fact that the agent's actuators are activated so as to make it move to the immediate next box in the selected direction. The set of motor activity features the agent is able to perceive by proprioception is the set $MF = \{N, NE, E, SE, S, SW, W, NW, Diag., Orth.\}$, where the first eight are specific to each particular direction, while *Diag.* and *Orth.* are more general features shared by all motor activities yielding diagonal/orthogonal moves. In cases where there is a wall at the edge of the depart box in the selected direction, the agent bumps into it and remains at the same place. We then say that the action's outcome is a failure. Otherwise, the action's outcome is the agent's new location.

3.3. Concepts and action laws

The agent can form two kinds of concepts. First, concepts of "things", in the broad sense. These concepts bind together co-occurrent features, and can be seen as some sort of conjunction in which conjuncts have different "weights", reflecting the fact that some features are more important than others in a concept's definition [Freund, 2008]. They store the agent's knowledge about locations and more generally any object, so we call them *object concepts*. The second kind is relational concepts. These take other concepts as elements, and bind them together into tuples. Concepts of actions are of this kind: they bind together the agent's concepts of a depart location, a performed motor activity, and a subsequent out-

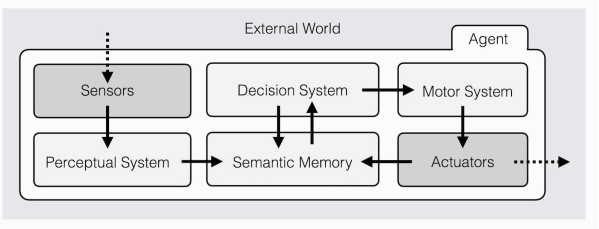

Figure 2. Schema of the agent

come, in the order in which they were experienced.

An object concept X is said to be is *general*, as opposed to *particular*, if there is another concept Y such that the set of features composing X is a strict subset of the set of features composing Y. Y is then said to be *more particular* than X. An action concept is said to be *general* if the object concept of its initial situation is general or its motor activity component only contains *Diag.* or *Orth.*. Generality of concepts is understood relatively to the set of concepts the agent possesses at some point, so no concept is general or particular in itself.

For example, when visiting the box $(0, 0)$ the agent may form the particular object concept `[OK,#12]`, which is a memory of an OK place with name #12, and only applies to this particular box in its accessible universe. If the agent then moves North-East and arrives at box $(1, 1)$, it can form the particular object concept `[OK,#18]`, and also the particular action concept `[[OK,#12],[NE,Diag],[OK,#18]]` which corresponds to the memory of being in an OK place with name #12 and then moving North-East to arrive at another OK place with name #18. Yet, after visiting a number of locations having the feature *OK* in common, the agent may also form the general object concept `[OK]`. Furthermore, it is a general rule in its accessible universe that moving North-East from an OK location always leads to another OK location, except for when there is a wall at the North or East edge of the depart box. Therefore, after having experienced a number of North-East moves from various OK locations, the agent may form general action concepts such as `[[OK],[NE,Diag],[OK]]` and `[[OK,NorthWall][NE,Diag],[Failure]]`. Such general concepts capture the general (non-monotonic) action laws of the agent's universe, and are the ones it shall rely on to behave in never encountered situations.

4. Implementing the agent's semantic memory in the neural network

Spiking Neural Networks (SNNs) are well suited for autonomous learning in open worlds, as they allow for *Spike Time Dependent Plasticity* (STDP), a family of biologically plausible learning rules which can achieve unsupervised online learning from unlabelled data [Thiele et al., 2018]. They are also known for being energy-efficient, which is an interesting property for autonomous robots.

The neural network implementing the agent's semantic memory is inspired from the JAST learning rule [Thorpe et al., 2019; Thorpe, 2023], which is a simplified version of STDP where the sum of the afferent connection weights on any given neuron remains constant through learning. However, contrary to what is the case in JAST here the weights are not binary but are natural numbers, and neurons are not frozen after learning so as to allow updating.

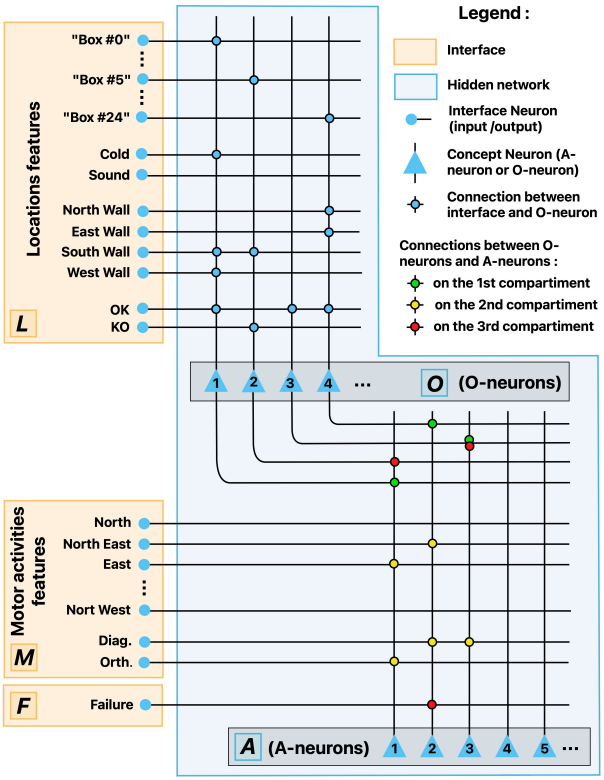

Figure 3. Schema of the SNN. O-neurons #1 and #2 respectively encode the concepts `[#0,Cold,SouthWall,WestWall, OK]` and `[#5,SouthWall,KO]`. A-neuron #1 encodes the concept `[[#0,Cold,SouthWall,WestWall,OK],[E,Orth],[#5,SouthWall,KO]]`. A-neuron #3 encodes the concept `[[OK],[Diag],[OK]]`. For graphical readability, connections weights are omitted.

4.1. The neural network's architecture

The network is composed of an interface, which communicates with the agent's other components, and a body of hidden neurons which is itself divided into two layers (see Figure 3). The first layer learns object concepts and the second learns action concepts. For this reason we call their neurons, respectively, *object concept neurons* (O-neurons for short) and *action concept neurons* (A-neurons). This architecture draws on neuroanatomical studies according to which concepts are represented in the brain by hierarchically organized *concept neurons*, each receiving information from some lower neurons and sending reciprocal connections to these same neurons so that it can reactivate them for information retrieval [Quiroga, 2012; Bausch et al., 2021; Shimamura, 2010]. For simplicity these reciprocal connections are not modeled as such, but instead information is allowed to flow in both directions along the same connections: from interface neurons to O-neurons and from there to A-neurons for learning and querying, and the other way round for retrieving information. A key point is that interface neurons are both input and output neurons, depending on the

phase of the computation.

Interface neurons (I-neurons for short) mainly support the representation of features, be it of the visited locations or of the agent's own motor activities. An additional neuron acts as a failure detector, specifically firing when the agent bumps into a wall and remains at the same place. All of them have their labels fixed from the start.

The first layer of hidden neurons is composed of 100 integrate and fire neurons, with a differentiated dynamics depending on whether their input source is I-neurons or A-neurons. O-neurons learn co-occurrences of features.

The second layer is composed of 400 compartment neurons with three separate input compartments. The first compartment receives connections from O-neurons, the second from motor activities I-neurons, and the third from O-neurons and the Failure neuron. Inputs received at each compartment are unable to trigger a spike by themselves, but the first and second compartments respectively make their next compartment ready to receive and transmit inputs for a certain amount of time. In this manner, inputs can only be efficient if they occur in the correct order, so that A-neurons encode sequences of inputs. The use of compartment neurons to learn sequences of inputs was suggested in [Cui et al., 2016; Hawkins and Subutai, 2016].

For readability, parameter settings are provided in the Appendices.

4.2. The neural network's functioning

We respectively note $\boldsymbol{L}$, $\boldsymbol{M}$, and $\boldsymbol{F}$ the sets of interface neurons that encode locations' features, motor activities' features, and failure. $\boldsymbol{I} = \boldsymbol{L} \cup \boldsymbol{M} \cup \boldsymbol{F}$ is the set of interface neurons. We note $\boldsymbol{O}$ the set of O-neurons and $\boldsymbol{A}$ the set of A-neurons. For convenience we call *"L-neurons"* the neurons from $\boldsymbol{L}$, and similarly for the other above sets.

For any neurons l from $\boldsymbol{L}$ and o from $\boldsymbol{O}$, $w_{l,o}$ denotes the weight of the connection between l and o. Similarly, for any compartment c from $\{1, 2, 3\}$, any neuron n from $\boldsymbol{O} \cup \boldsymbol{M} \cup \boldsymbol{F}$ and any A-neuron a, $w^c_{n,a}$ denotes the weight of n's connection on a's c^{th} compartment. Synapses between any two neurons always have a weight of 1, but an input neuron may have multiple synapses onto a same output neuron, so the weight of a connection between two neurons is always a positive integer (or 0, if there is no connection). Connection weights are initialized at random, in such a way that the sum of weights on any O-neuron is equal to some fixed integer WS_O, and the sum of weights on any compartment of any A-neuron is equal to some fixed integer WS_A. Connexions are stored in separate arrays $cnxLO$, $cnxOA1$, $cnxMA2$ and $cnxOA3$, where $cnxLO$ reads *"connections from* $\boldsymbol{L}$ *to* $\boldsymbol{O}$*"*, and similarly for the others. The agent's learning consists in the update of these arrays.

For each O-neuron o a fixed spike threshold $STO(o)$ is set at random to a value comprised between some minimal and maximal values. This spike threshold governs forward spiking, that is, cases where the input originates from $\boldsymbol{L}$-neurons. Differentiated spike thresholds help the neurons to encode concepts with different degrees of generality, as neurons with lower spike thresholds tend to spike more easily to inputs that do not fit perfectly their connections, which brings them to learn more general concepts. O-neurons also have a backward spike threshold $backwrdSTO(o)$, which governs their spiking when the input originates from A-neurons. To ease information retrieval, $backwrdSTO(o)$ is set to $STO(o)$/3. L-neurons all have the same spike threshold STL, and the Failure neuron has its own spike threshold. Spike thresholds for A-neurons are dynamically adjusted, in a way that is described below.

O-neurons. Given a set $inputLO$ of L-neurons firing in response to an input from either the perceptual system (if the agent is observing its current location) or the decision system (if the agent is querying its semantic memory), we note $firedO$ the set of O-neurons firing in response to $inputLO$. It is computed as follows. First, for any neuron o in $\boldsymbol{O}$ the sum $inputSumO(o)$ of input received by o is computed. Namely, $inputSumO(o) = \sum w_{l,o}$ for $l \in inputLO$. The difference $overT(o)$ with o's spike threshold is then calculated (that is, $overT(o) = inputSumO(o) - STO(o)$). Neurons o for which $overT(o)$ is positive (i.e., which have reached their spike threshold) then fire in turn, starting by those with the highest $overT(o)$, until either no O-neuron with positive $overT$ value remains or the number of firing O-neurons reaches some target number $TnbFiredO$. This target number is a fixed parameter of the network that helps to keep the number of firing O-neurons under control.

For any O-neuron o the number of steps performed by the agent since o's last spiking is noted $lastSpikedO(o)$. $lastSpikedO(o)$ is initialized at random to a value comprised between 0 and 2,000, and increased by 1 at each step. It is reset to 0 each time the neuron spikes after learning.

Learning of object concepts occurs whenever the agent observes its current situation. The O-neurons firing in response to the observation input $inputLO$ are selected for learning. That is, we set $learningO = firedO$, where $learningO$ is the set of O-neurons selected for learning. If there are enough learning neurons, that is, if $|learningO|$ reaches some target number $TnbQueryO$ (with $TnbQueryO$ a fixed parameter of the network such that $TnbQueryO < TnbFiredO$), the network directly proceeds to learn. Otherwise it first looks for additional neurons by boosting the input received by O-neurons. In practice, for any non-firing O-neuron we define $boostedInputSumO(o) = inputSumO(o) * (lastSpikedO(o) + b)/b$, with $b \in \mathbb{N}$ a fixed parameter. $lastSpikedO(o) + b)/b$ is an increasing function of $lastSpikedO(o)$, so neurons having not spiked for a

long time receive more boosting. Neurons with the highest boosted input sum values are then successively added to $learningO$ until $|learningO| \geq TnbQueryO$. In this manner, O-neurons that have not been used for a long time are preferentially recruited for learning new information. Of course, some of the information encoded by these neurons will be lost in the next learning process, bringing the agent to forget some of its previous knowledge. But this will only affect the agent's least used knowledge, as these neurons are the least used at the time of the learning.

The learning of O-neurons essentially consists in replacing their connections from non-firing L-neurons with connections from firing L-neurons (namely, the neurons from $inputLO$). To decide which neuron from $inputLO$ will establish new synapses with a given O-neuron, a probability distribution $probaNewSynapsesLO$ is defined. For any neuron l in $inputLO$, $probaNewSynapsesLO(l)$ represents l's propensity to grow new synapses to O-neurons at time point t. $probaNewSynapsesLO(l)$ is a decreasing function of $\sum w_{l,o}$ for $O \in \boldsymbol{O}$, that is, of the weight sum of all of l's connections to O-neurons. It accounts for the fact that neurons cannot indefinitely grow new outgoing synapses, hence that a neuron having already a lot of them should be less prone to establish new connexions than a neuron having fewer of them. It ensures that L-neurons encoding rarely observed features (such as, e.g., boxes' names) still manage to have connections with at least a few O-neurons.

The learning process then depends on the accuracy of the agent's knowledge relative to the current situation. To assess it, O-neurons from $firedO$ send a backward input to L-neurons, and the resulting set $firedL$ of firing L-neurons is compared with the initial input $inputLO$. This corresponds to the agent comparing what it observes (represented by $inputLO$) with what it would have expected based on its current knowledge (represented by $firedL$).

If $inputLO \subseteq firedL$, that is, if all the observed features can be retrieved from the firing O-neurons, then for any neurons o in $learningO$ and l in $\boldsymbol{L} \backslash inputLO$, the connection weight $w_{l,o}$ is reset to 0 (i.e., all of o's inactive synapses are deleted). They are then replaced with new synapses from neurons from $inputLO$ following the above-mentioned probability distribution, in such a way that $\sum w_{l,o}$ for $l \in \boldsymbol{L}$ remains equal to WS_O. This procedure tends to reinforce O-neurons' connections with L-neurons encoding well-shared features at the expense of connexions with L-neurons encoding more specific features, and is the driving force for the formation of general object concepts.

If, on the contrary, $inputLO \not\subseteq firedL$, then we first check if $|learningO|$ reaches $TnbFiredO$. If not, the network looks for one more neuron to add to $learningO$, using the same method as above except that only one neuron is added. The learning process is then similar as before, except that for one O-neuron o with maximal $lastSpikedO(o)$, all synapses whether active or inactive are replaced. Doing so favors the establishment of connections with all the neurons from $inputLO$, so that this particular neuron tends to learn the particular situation with all its features.

Querying the network. At any point in time, we respectively note $inputOA1$, $inputMA2$ and $inputOA3$ the sets of neurons from $\boldsymbol{O}$, $\boldsymbol{M}$ and $\boldsymbol{O} \cup \boldsymbol{F}$ that send inputs to A-neurons' first, second and third compartments. Inputs on the first compartment encode an initial situation, inputs on the second compartment encode motor activity features, and inputs on the third compartment encode an action's outcome. These inputs may occur in two different situations: when the decision system queries the neural network for the expected outcome of an envisaged action, or when the neural network learns about a given action.

To query the neural network, the decision system first sends an input to the L-neurons that encode the features of the initial situation, bringing a set $inputLO$ of L-neurons to fire. This in turn triggers the firing of a set $firedO$ of O-neurons, following the above-described process. $firedO$ is an input to A-neurons' first compartment, that is $inputOA1 = firedO$. Given an envisaged motor activity $Move$, the decision system then sends an input to the M-neurons that encode $Move$'s features. These fire, sending an input $inputMA2$ to A-neurons' second compartment. Querying the network amounts to bringing some A-neurons to spike in response to this incomplete input and send backward input to O-neurons, so that these spike and send backward input to interface L-neurons. The resulting firing of interface L-neurons is the network's answer to the query.

Now, among the A-neurons that receive some input from O- and M- neurons, some support the representation of general action concepts (such as, e.g., `[[OK], [NE], [OK]]`), while others support the representation of more particular concepts (such as, say, `[[OK, NorthWall], [NE], [Failure]]`). Obviously, it is desirable that the neurons supporting the more particular concepts take precedence over those supporting more general concepts, as more particular concepts convey more adequate information. To achieve this, the input sent to A-neurons is modulated by applying a multiplying factor to the neurons' connections weights. More precisely, for any neuron o in $inputOA1$, this multiplying factor is a decreasing function of $\sum w^1_{o,a}$ for $a \in \boldsymbol{A}$, and similarly for $inputMA2$. The idea behind this is that neurons with fewer output connections tend to support more particular concepts than neurons with more output connections, hence that the A-neurons to which they connect also tend encode more particular concepts.

Modulated input sums are computed separately for each compartment of each A-neuron a, in a way similar to that used for O-neurons. If the input sum received by

a on its first compartment is above some noise threshold $noiseA$, then a's second compartment is able to transmit its input to a. The total input received by a is then computed as the sum of the two inputs. Otherwise, both inputs are discarded.

Due to input modulation, the input sum received by a given A-neuron a indicates both the accuracy of the action concept encoded by a relative to the input pair ($inputOA1$, $inputMA2$), and the concept's degree of generality: the more adequate and particular the encoded concept, the higher this input sum.

The set *firedA* of firing A-neurons is obtained by first setting A-neurons' spike threshold to the highest input sum value (so only A-neurons with the highest input sum fire), and then repeatedly lowering it to the next highest value until the number of firing A-neurons reaches a target number $TnbQueryA$. Each time the spike threshold is lowered all the A-neurons with input sums above the threshold spike again, so A-neurons with higher inputs sums spike repeatedly.

Each time an A-neuron spikes, it sends a backward input to O- and Failure neurons through its third compartment's connections. Backward input sums are computed incrementally as A-neurons fire, bringing O- and Failure neurons to fire as soon as their spike threshold is reached. If some O-neuron fires first, it inhibits the Failure neuron and prevents it from further firing. Similarly, if Failure neuron spikes first, it inhibits all the O-neurons and prevents them from further firing. Firing O-neurons in turn send backward inputs to L-neurons and the input sum received by L-neurons is computed incrementally. L-neurons spike as soon as they reach their threshold, and their firing inhibits the L-neurons that encode incompatible features (for example, the firing of the L-neuron encoding the feature ***OK*** prevents the L-neuron encoding the feature ***KO*** from further firing).

The value of A-neurons' spike threshold at the moment a given neuron $n \in \boldsymbol{L} \cup \boldsymbol{F}$ spikes measures the relevance of the A-neurons involved in n's spiking relative to the input pair ($inputOA1$, $inputMA2$). It can therefore be used by the agent to assess the reliability of its prediction. In other words, it indicates the degree of confidence the agent has in its prediction of the feature encoded by n: the higher this value, the higher the confidence. So, formally, the querying process returns a prediction set $P_{Move} = \{(f_1, conf_1), ...(f_n, conf_n)\}$, where f_i is a feature and $conf_i$ the degree of confidence the agent has in its prediction.

A-Neurons' learning. Learning of action concepts takes place after each step made by the agent. In this case, $inputOA1$ is the set *firedO* resulting from the agent observing its depart location, $inputMA2$ is the set of M-neurons firing by proprioception as the agent performs the motor activity, and $inputOA3$ is the set *firedO* resulting from the agent observing its arrival location.

For any neuron a in $\boldsymbol{A}$, $inputSumA1A2(a)$ is sum of the inputs received by a on its first and second compartments. That is, $inputSumA1A2(a) = \sum w^1_{o,a}$ for $o \in inputOA1 + (\sum w^2_{m,a}$ for $m \in inputMA2)$.
If $inputSumA1A2(a)$ reaches a certain threshold $learnAT$ and a fired in the querying process by which the agent predicted the outcome of the current action (see Section 5 below), then a is selected for learning. The set of such A-neurons is denoted $learningA$. If $|learningA|$ reaches a target number $TnbQueryA$ the network directly proceeds to learn. Otherwise, additional neurons are selected by boosting the input sums received by A-neurons in a way similar to the one previously described for O-neurons.

For each A-neuron a, the number of steps performed by the agent since a's last spiking is noted $lastSpikedA(a)$. $lastSpikedA(a)$ is initialized at random to a value comprised between 0 and 20,000, and increased by 1 at each step. It is reset to 0 each time the neuron spikes after learning (i.e., spikes occurring during the querying process are discounted).

For each input set i in $\{inputOA1$, $inputMA2$, $inputOA3\}$, a probability distribution $probaNewSynapsesA_i$ is computed on i that models the propensity of each neuron n from i to grow new synapses to A-neurons' corresponding compartments. This probability is a decreasing function of $\sum w^c_{n,a}$ for $a \in \boldsymbol{A}$, where $c \in \{1, 2, 3\}$ depends on the input set.

For each A-neuron a in $learningA$, a learning rate $LR(a)$ is computed. This learning rate is an integer comprised between 1 and WS_A (Remember that WS_A is the sum of connections weights on any A-neuron compartment). $LR(a)$ represents a's propensity to modify its connections through learning, and is calculated in a way such that A-neurons with maximal $lastSpikedA$ value among those in $learningA$ get a learning rate equal to WS_A, those with minimal $lastSpikedA$ value get a learning rate equal to 1, and the others get learning rates in between.

The learning process then depends on the accuracy of the agent's expectations relative to the action's outcome. These expectations were obtained by querying the network before choosing to perform the action. If these expectations were both correct (i.e., all expected features are actually present) and complete (i.e., all actually present features were expected), then for any A-neuron a in $learningA$, the learning process deletes all of a's inactive synapses in the first two compartments, but only some of them in the third one. More precisely, if n is the number of inactive synapses on a's third compartment, the number of inactive synapses to delete is $min(LR(a), n)$. Then, in all three compartments new synapses replacing the deleted ones are chosen at random according to $probaNewSynapsesA_i$ (where i is the input set corresponding to the considered compartment), in such a way that the sum of connection weights (i.e., the number of synapses) on the compartment remains equal to WS_A.

If the agent's expectations were incorrect, then for any a in $learningA$ the learning process not only replaces all inactive synapses on a's first two compartments, but also some active ones. Here the number of synapses to be replaced is $max(LR(a), n)$, where n is the number of inactive synapses, and the deleted active synapses are chosen at random. As for a's third compartment, the learning process is the same as in the first case. As previously with O-neurons, replacing some active synapses in addition to the inactive ones allows the considered compartments to get connections from more neurons from the input set, and so to "re-specialize" relative to the input. Making the first two compartments more specialized to, respectively, the initial situation and the motor activity helps the neuron to fire only in response to this precise action, hence to retain more accurate information about its outcome.

If the agent's expectations were incomplete, then for any A-neuron a in $learningA$ the number of replaced synapses is $max(LR(a), n)$ in all three compartments. In the case where the predictions were neither correct nor complete, the learning process successively runs the two above procedures.

5. Functioning of the agent

Suppose the agent is at some initial location and observes it: information from its perceptual system triggers the firing of the L-neurons encoding the location's features, which brings a number of O-neurons to fire and send an input $inputOA1$ to A-neurons' first compartment. Concomitantly, the information is transmitted to the agent's decision system, which decides to make a step.

The agent's choice of a motor activity depends on whether it wants to exploit its current knowledge about its environment, or to explore it to improve its knowledge. The exploration/exploitation dilemma is a well-known problem in online learning [Watkins, 1989; Sutton and Barto, 2018], and evolving environments make it even more difficult. Therefore, we do not try to reach an optimal solution here, but instead we simply make the agent's decision system choose at random with equal probability between an *Exploration* and an *Exploitation* mode.

This choice being made, for each motor activity $Move$ out of the eight possible ones the decision system queries the semantic memory for the outcome of the action having the current location as initial situation and $Move$ as motor activity. It then rates each of them for its suitability, by building the sets S (for *"Suitable"*), US (*"Unsuitable"*) and UD (*"Undecided"*):

- $S = \{(Move, conf) \mid (\textbf{\textit{OK}}, conf) \in P_{Move}\}$
- $US = \{(Move, conf) \mid (\textbf{\textit{KO}}, conf) \in P_{Move}$ or $(\textbf{\textit{Failure}}, conf) \in P_{Move}\}$
- $UD = \{Move \mid \nexists conf)$ s.t. $(Move, conf) \in S \cup US\}$

The decision system then chooses a motor activity depending on the selected mode. In *Exploration* mode, the agent is willing to take risks and chooses an action with the most uncertain outcome possible so to increase its knowledge: if $UD \neq \emptyset$ it picks one from UD, otherwise it goes for one with the least $conf$ in $S \cup US$. In *Exploitation* mode by contrast, the agent just wants to avoid KO boxes and failure as much as possible. So, if $S \neq \emptyset$ it chooses one with the greatest $conf$, otherwise if $UD \neq \emptyset$ it picks one from it, and if both S and UD are empty it chooses one with the least $conf$ in US.

The decision system then transmits its decision to the motor system to perform the selected motor activity. The M-neurons that encode its features are activated by proprioception and send an input $inputMA2$ to A-neurons' second compartment.

The agent's move is simulated by computing its arrival location and its features. If the agent bumps into a wall, the *Failure* neuron fires and sends an input $inputOA3$ to A-neurons' third compartment, and the agent directly learns the action (see A-neurons' learning above). Otherwise, it first learns relevant object concepts about its new location (see O-neurons' learning). The input $inputOA3$ to A-neurons' third compartment is then the set $firedO$ obtained by making O-neurons fire again while observing the new location after learning about it.

The agent is then ready for the next step. As the current arrival location is also next step's depart location, next step input to A-neurons' first compartment $inputOA1$ is also $firedO$. If the action's outcome was failure, $inputOA1$ is kept unchanged.

6. Experiments and results

All tests were realized by placing the agent at location $(0, 0)$ and prompting it to perform a succession of series of steps, each complete sequence of series of steps being called a *trial*. The results presented here are averaged over 50 trials. Three distinct groups of tests ("Experiments") were carried out.

6.1. Experiment #1

In a first group of tests, trials consisted in 65536 steps in total with the door kept closed all along, so that the agent had no access to the second room. A first test concerned the agent's ability to learn an action over one single experience ("one-shot learning"). To assess it, after each step the agent was asked to redo the prediction that led to the just realized action, and this new prediction was compared with the action's actual outcome (see Table 1).

A prediction is said to be *Correct and Complete* (CC) if the predicted features are exactly those of the arrival location. The table's first line shows the mean percentage of steps leading to a CC post-learning prediction, for each series of steps. The second line (MF for *"Missed*

Nb of Steps	1	2	4	8	16	32	64	128	256
CC	0.0	100.0	100.0	99.0	98.2	97.8	96.0	91.9	89.9
MF	100.0	0.0	0.0	0.4	0.9	1.1	2.1	4.2	5.1
PE	0.0	0.0	0.0	0.0	0.0	0.0	0.1	0.1	0.4

Nb of Steps	512	1024	2048	4096	8192	16384	32768	65536	Global
CC	91.3	93.9	94.8	95.8	96.3	95.8	96.2	96.5	96.2
MF	4.1	2.8	2.3	1.9	1.7	1.9	1.7	1.6	1.7
PE	0.4	0.2	0.2	0.2	0.1	0.1	0.1	0.1	0.1

Table 1. Mean post-learning predictions percentages over 50 trials. CC: Correct and Complete, MF: Missing Features, PE: Prediction Errors.

Nb of Steps	Room	OK	KO	Fail.	Wall	Cold	Sound	Box Name
2	1	15.6	6.7	0.0	3.9	2.6	0.0	1.2
	2	14.3	4.7	0.0	1.9	0.0	0.0	0.0
8	1	36.1	29.1	15.4	11.5	8.9	0.0	4.8
	2	31.6	24.2	12.7	7.0	0.0	0.0	0.0
64	1	83.2	81.5	42.9	35.8	28.2	0.0	22.1
	2	76.6	79.8	34.2	23.0	0.0	0.0	0.0
512	1	95.5	93.3	67.2	64.5	58.3	0.0	51.2
	2	87.8	88.2	36.1	23.7	0.0	0.0	0.0
4096	1	96.7	94.7	78.8	79.2	74.5	0.0	70.6
	2	91.2	88.3	44.3	27.5	0.0	0.0	0.0
16384	1	96.8	94.9	82.0	81.5	78.8	0.0	75.6
	2	89.9	84.7	48.4	28.7	0.0	0.0	0.0
32768	1	97.1	94.7	80.7	81.9	79.5	0.0	74.7
	2	90.9	87.8	45.9	29.4	0.0	0.0	0.0
65536	1	97.3	94.0	83.0	81.9	78.2	0.0	75.3
	2	92.1	85.2	50.4	29.6	0.0	0.0	0.0

Table 2. Predictions' *Hit Rates* after n steps in the first room with the door kept closed. Results averaged over 50 trials. For conciseness results are only shown for some series.

Nb of Steps	Room	OK	KO	Fail.	Wall	Cold	Sound	Box Name
2	1	24.8	22.6	0.0	20.5	1.3	0.0	9.2
	2	13.9	3.9	0.0	1.3	0.0	0.0	0.0
8	1	60.8	60.0	39.5	26.7	5.5	0.0	10.9
	2	59.7	49.8	30.1	17.3	0.0	0.0	0.0
64	1	76.1	77.4	62.4	33.4	26.3	0.0	21.7
	2	80.0	80.0	46.4	19.1	0.0	0.0	0.0
512	1	89.5	85.8	87.0	70.0	75.7	0.0	65.4
	2	83.2	82.4	52.5	26.8	0.0	0.0	0.0
4096	1	92.1	90.6	93.0	84.0	91.2	0.0	81.7
	2	83.7	84.9	63.7	32.8	0.0	0.0	0.0
16384	1	92.2	92.4	94.6	87.6	92.3	0.0	84.1
	2	83.6	85.6	63.1	32.7	0.0	0.0	0.0
32768	1	91.9	91.6	95.0	86.7	90.8	0.0	84.2
	2	83.4	84.8	67.6	33.1	0.0	0.0	0.0
65536	1	92.1	92.7	94.3	87.5	92.9	0.0	84.2
	2	83.8	86.2	64.8	34.3	0.0	0.0	0.0

Table 3. Predictions' *Correctness* after n steps in the first room with the door kept closed. Results averaged over 50 trials. For conciseness results are only shown for some series.

Features") shows the average percentage of features occurrences that the agent failed to predict after learning. The third line (PE for *"Predictions Errors"*) shows the average percentage of wrongly predicted features occurrences.

These results show a good performance at immediate recall after learning. The fact that the network does not learn anything at the first step is expected, as there is no previous step so no arrival location either. Therefore, $inputOA1$ is the empty set, which means that A-neurons' first compartments receive no input at all. This prevents them from integrating inputs received at their second compartment, so the total input received by A-neurons at first step is 0 and no learning can occur. The lower CC prediction scores around 128-1024 steps are explained by the fact that at start-up the network has a lot of unused neurons at its disposal and simply learns the particular concepts with all their features, so specific features are well learned. As the network accumulates knowledge, unused neurons get scarce and the network tends to recruit less-used neurons for learning instead. A trade-off is then made between learning the current observation's specific features and not forgetting the common features already encoded by the neurons, which makes specific features more difficult to learn. Yet after repeated encounters the specific features are finally learned, so the score improves again.

To test whether the acquired knowledge was retained in the long run, after each series of steps the simulation was frozen and learning was deactivated, and the agent was placed successively in each location of each room. There, it was queried for its predictions for each of the eight possible motor activities, and its predictions were compared with the actions' actual outcomes. Tables 2 and 3 show each feature's mean *Hit Rate* (that is, its chances of being predicted when effectively present), and *Correctness* (its chances of being effectively present when predicted)[1] for each room.

Values for the first room (white lines) show that learned actions are indeed recalled long after having been per-

[1] *Hit Rate* is also known as *True Positive Rate*, *Recall* or *Sensitivity*, while *Correctness* is also known as *Precision* or *Positive Predictive Value* – see for example [Kohavi and Provost, 1998] for definitions. Here we multiplied the obtained figures by 100 to get percentages.

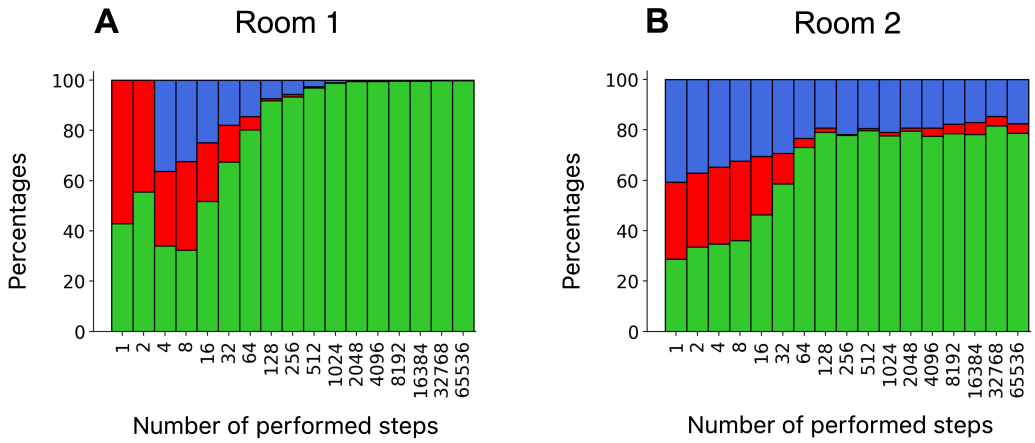

Figure 4. Actions' mean outcomes with door closed; Green: OK, Red: KO, Blue: Failure.

formed. Values for the second room (grey lines) show that despite never having been in this room (since the door was kept closed) the agent was able to correctly predict *OK* and *KO* features and to a lesser extent *Failure* —and this, even though locations from the second room have different sets of features, including for some of them a new feature, *Sound*. The poor performance at wall prediction is due to the lack of general rules of the universe observable in the first room regarding the presence of walls in adjacent boxes. This brings the agent use the particular concepts it has of individual boxes to predict walls in the first room, but these cannot be used to predict walls in the second room. The mixed result at failure prediction comes from a competition between general action concepts, the control of which needs to be improved.

To test the agent's ability to use its knowledge to make appropriate decisions, the outcome of each action taken in *Exploitation* mode was recorded throughout the trials. Figure 4.A shows the percentages of *OK*, *KO*, and *Failure* outcomes thus obtained for each series of steps.

Visited locations were logged to check that the agent was not looping indefinitely on the same boxes: all boxes kept being visited, be it very rarely, at any point of the trials, due to the *Exploration* mode.

Finally, the agent's ability to use the knowledge acquired in the first room to act judiciously in the second room was tested. At the end of each series of steps the agent was asked to chose a move from each of the second room's type of locations (where two locations are said to be of the same type if they have exactly the same features). Figure 4.B shows the percentages of *OK*, *KO*, and *Failure* obtained in this manner. These results reflect the agent's performance at making predictions about the second room's locations: it successfully predicts OK and KO boxes, but has more difficulties predicting failure.

6.2. Experiment #2

To test the agent's ability to handle changes, a second group of tests was carried out. The setup was the same as in Experiment #1, except that the door was opened at the 2048^{th} step (as in Figure 1.B). The agent spontaneously went in the second room, and spent a variable but significant amount of time in it (34.6% of steps on average, standard deviation = 11.3). Tables 4 and 5 show the features' *Hit Rates* and *Correctness* from the moment the door was opened.

These results show that the agent was able to learn new concepts involving the *Sound* feature. The rather low *Hit Rates* for the feature are due to the lack of observable cues in boxes at the direct south of boxes with sound, which prevents the agent from being able to predict it in 40% of the cases.

It should be remarked that this learning did not lead to significant loss of previous knowledge regarding the first room: a moderate drop in *Hit Rates* can be observed for boxes' names, *Cold* and walls, but not for *OK*, *KO*, and *Failure*. Furthermore, *Correctness* is barely impacted. This is due to the use of selective forgetting in the learning process (remember O- and A- neurons' boosting of inputs in Section 4.2). As neurons encoding more particular concepts (such as detailed descriptions of particular locations) are less often reactivated, they tend to be reallocated over time, which leads to the loss of the particular features they encode. But neurons encoding more general concepts are more often reactivated, which preserves them from reallocation. As a result, particular and less-used concepts tend to fade out over time, while well-used and general concepts are preserved.

As a consequence, the agent's ability to make appropriate decisions in the first room was not impacted by the door being opened. In fact, the bar chart of *OK*, *KO*, and *Failure* outcomes obtained in this second run of tests showed no visible difference with the one obtained with the door closed shown in Figure 4.A.

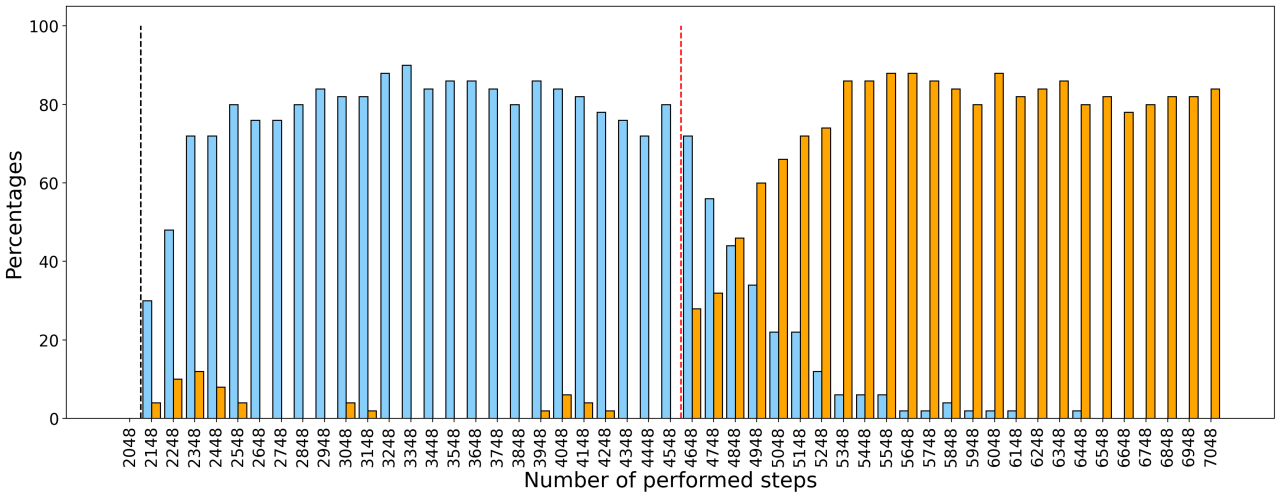

Figure 5. Prediction of walls in arrival boxes given *Sound* as information about the depart location. Black dotted line: door opens; Red dotted line: sound changes location; Blue: *NorthWall* predictions considering a north move; Orange: *SouthWall* predictions considering a south move (percentages over 50 trials).

Nb of Steps	Room	OK	KO	Fail.	Wall	Cold	Sound	Box Name
2048	1	96.5	94.4	75.3	75.7	72.0	0.0	64.6
	2	90.8	88.9	37.4	26.8	0.0	0.0	0.0
4096	1	96.8	97.6	74.4	63.9	59.4	0.0	53.9
	2	91.4	94.0	51.5	34.2	0.0	34.0	2.7
8192	1	97.1	96.9	77.6	68.3	64.6	0.0	59.5
	2	91.2	92.7	55.8	32.6	0.0	28.6	3.6
16384	1	97.0	96.5	80.4	66.7	63.2	0.0	57.9
	2	91.9	91.7	63.7	37.3	0.0	35.1	4.0
32768	1	96.9	97.3	82.9	67.8	64.8	0.0	60.3
	2	92.4	92.4	68.1	36.9	0.0	41.8	4.7
65536	1	98.0	97.3	84.4	70.3	64.5	0.0	61.4
	2	93.0	91.5	71.0	39.7	0.0	36.5	6.9

Table 4. Predictions' *Hit Rates* after n steps, with the door opened at the 2048^{th} step.

Nb of Steps	Room	OK	KO	Fail.	Wall	Cold	Sound	Box Name
2048	1	91.9	88.8	91.4	79.8	88.0	0.0	78.5
	2	83.2	83.3	56.9	30.7	0.0	0.0	0.0
4096	1	94.3	90.5	92.2	74.2	96.0	0.0	86.7
	2	89.8	86.0	73.8	43.0	0.0	57.6	0.7
8192	1	94.5	92.0	92.3	77.7	95.4	0.0	89.4
	2	90.0	86.6	71.7	41.5	0.0	57.9	0.9
16384	1	94.7	93.1	91.8	77.2	96.1	0.0	89.9
	2	90.1	89.0	76.3	46.9	0.0	60.3	1.4
32768	1	95.4	93.8	93.7	79.9	94.0	0.0	92.9
	2	91.3	89.8	78.8	48.2	0.0	65.6	1.1
65536	1	95.6	94.7	95.3	78.8	98.2	0.0	92.5
	2	91.1	90.9	79.0	48.2	0.0	58.6	1.8

Table 5. Predictions' *Correctness* after n steps, with the door opened at the 2048^{th} step.

6.3. Experiment #3

To better assess the agent's ability to update its knowledge to accommodate environment changes, a third experiment was conducted. As previously, the agent was kept in the first room for 2048 steps before the door was opened, but then it was prompted to run 50 series of 100 steps. Half way through these, the *Sound* feature was moved from the boxes $(3, 3)$ - $(8, 3)$ to the boxes $(3, -1)$ - $(8, -1)$ (as in Figure 1.C). At the end of each of these series of steps, the agent was asked for its predictions about the outcome of two actions having *Sound* as the only available information about the depart location: the first one when considering a north move, and the second, when considering a south move. It was recorded whether *NorthWall* (resp., *SouthWall*) was predicted. Other predicted features, if any, were discarded.

Figure 5 shows the percentage over 50 trials of *NorthWall* (resp., *SouthWall*) predictions after each series of steps.

These results show that after the door opened the agent first learned to predict a north wall in the arrival box when considering a north move, but that once the *Sound* feature was moved to its "down" position it stopped doing so and learned to predict a south wall in the arrival box when considering a south move instead. This learning was both fast and robust, considering that the agent had very few learning occasions. Indeed less than 0.6% (on average) of taken steps were north moves from a box with sound, and similarly for south moves from a box with sound. This means that very few experiences were enough for the agent to learn / update its knowledge, which is consistent with the results obtained in the first experiment.

6.4. Computing time

Computing time was simply estimated by running trials consisting in a single series of 2024 steps. An average of

6.5 seconds (standard deviation 0.13) was obtained on a conventional computer. However this is to be taken as an upper bound, since to ease reproducibility some testing carried out in the course of the series was not deactivated. No attempt to optimize the computing time was made, as it seems less critical in the case of online learning of autonomous agents which can learn while physically performing their actions.

7. Conclusion and Future Developments

In this paper a proof of concept of the architecture of a fully autonomous agent learning action laws online and accommodating environment changes was provided. This agent relies on general concepts to handle new situations and dynamically adjusts its concepts to its current environment. This makes it well suited for open worlds: if a new door were to open to a third room with new objects and laws, it would learn them just as it did in the second room. Of course, this would come at the cost of the forgetting of its least used concepts, but these are precisely the ones it needs the less. In fact, the agent's ability to selectively forget its least used knowledge ensures that it will always be able to adapt to new environments by replacing old unused concepts by new useful ones.

As regards scalability to larger environments, the number of neurons needed in each layer of the network (interface, O-neurons, A-neurons) grows linearly with the number of items (features, object concepts, action concepts) to encode; the number of synapses needed on each O-neuron grows linearly with the maximum number of features to be learned for a given object/situation. However, it is important to observe that these are upper bounds: the agent's ability to form general concepts out of individual experiences and to use them to make decisions makes it unnecessary to encode each individual object with all its features and each performed action.

Future developments of the architecture include endowing the agent with planning abilities. To do so, a notion of *applicable* action law should be defined. In a first approach, an action law represented by an action concept `[x, y, z]` could be deemed applicable in a situation s if s satisfies all the features in `x` and `z` $\neq$ `[Failure]`. The agent would also need to build its own set of possible situations (states) online. The set of its object concepts could probably be used to this end. The decision system should also be augmented to represent goals and embed a cost function and a planning algorithm.

However further work remains to be done to allow the agent to live in more realistic environments. Notably, it would be necessary to make the agent able to use incomplete information as input for learning and querying, as real world agents' observations are rarely complete. It would then be interesting to make the agent able to query its semantic memory for object properties given some partial input. It would also be useful to implement negation in the network, to allow the agent to represent the fact that a given object does not have a given feature. Neural inhibition could probably be used for this, but the appropriate learning rules remain to be found. It seems that taken together these two improvements would bring the agent to draw non-monotonic inferences in the spirit of [Grimaud, 2016].

It would also be suitable to make the agent able to distinguish between objects and their locations, as actions can modify one, the other, or both. Biological brains process the "what" and the "where" components of observations in two separate pathways before reunifying them, and this could be an inspiration source.

A further line of research would be to investigate how an agent should decide between *Exploration* and *Exploitation* modes in an open world. Intuitively, it seems that the choice between these two operating modes should depend on the agent's estimation of the risk associated with an explorative behavior, and that the agent should refrain from entering in exploration mode in situations where the assessed risk is high, while allowing itself more exploration in situations where it is low. Yet a correct formalization of this idea remains to be found.

Acknowledgements

This work is supported by the ANR (Agence Nationale de la Recherche) project ALoRS ("Action, Logical Reasoning and Spiking networks") [ANR-21-CE23-0018-01]

References

M. Bausch, J. Niediek, T. P. Reber, S. Mackay, J. Boström, C. E. Elger, and F. Mormann. Concept neurons in the human medial temporal lobe flexibly represent abstract relations between concepts. *Nature communications*, 12(1):6164, 2021.

T. Bolander and N. Gierasimczuk. Learning to act: qualitative learning of deterministic action models. *Journal of Logic and Computation*, 28(2):337–365, 2018.

B. Bonet, G. Frances, and H. Geffner. Learning features and abstract actions for computing generalized plans. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 33, pages 2703–2710, 2019.

Y. Cui, S. Ahmad, and J. Hawkins. Continuous online sequence learning with an unsupervised neural network model. *Neural computation*, 28(11):2474–2504, 2016.

D. Das, S. Chernova, and B. Kim. State2explanation: Concept-based explanations to benefit agent learning and user understanding. *Advances in Neural Information Processing Systems*, 36:67156–67182, 2023.

A. Ecoffet, J. Huizinga, J. Lehman, K. O. Stanley, and J. Clune. First return, then explore. *Nature*, 590(7847): 580–586, 2021.

J. Farebrother, M. C. Machado, and M. Bowling. Generalization and regularization in dqn. *arXiv preprint arXiv:1810.00123*, 2018.

M. Freund. On the notion of concept I. *Artificial Intelligence*, 152(1):105–137, 2008.

A. Ghorbani, J. Wexler, J. Y. Zou, and B. Kim. Towards automatic concept-based explanations. *Advances in neural information processing systems*, 32, 2019.

C. Grimaud. Modelling reasoning processes in natural agents: a partial-worlds-based logical framework for elemental non-monotonic inferences and learning. *Journal of Applied Non-Classical Logics*, 26(4):251–285, 2016.

A. Gupta and P. Narayanan. A survey on concept-based approaches for model improvement. *arXiv preprint arXiv:2403.14566*, 2024.

J. Hare. Dealing with sparse rewards in reinforcement learning. *arXiv preprint arXiv:1910.09281*, 2019.

P. Hase, C. Chen, O. Li, and C. Rudin. Interpretable image recognition with hierarchical prototypes. In *Proceedings of the AAAI Conference on Human Computation and Crowdsourcing*, volume 7, pages 32–40, 2019.

J. Hawkins and A. Subutai. Why neurons have thousands of synapses, a theory of sequence memory in neocortex. *Frontiers in Neural Cicuits*, 10, 2016. doi: 10.3389/fncir.2016.00023. URL https://www.frontiersin.org/articles/10.3389/fncir.2016.00023.

S. R. Kheradpisheh, M. Ganjtabesh, S. J. Thorpe, and T. Masquelier. Stdp-based spiking deep convolutional neural networks for object recognition. *Neural Networks*, 99:56–67, 2017.

R. Kirk, A. Zhang, E. Grefenstette, and T. Rocktäschel. A survey of zero-shot generalisation in deep reinforcement learning. *Journal of Artificial Intelligence Research*, 76:201–264, 2023.

J. Kirkpatrick, R. Pascanu, N. Rabinowitz, J. Veness, G. Desjardins, A. A. Rusu, K. Milan, J. Quan, T. Ramalho, A. Grabska-Barwinska, et al. Overcoming catastrophic forgetting in neural networks. *Proceedings of the national academy of sciences*, 114(13): 3521–3526, 2017.

B. J. Knowlton, A. L. Siegel, and T. D. Moody. 3.17 - procedural learning in humans. In J. H. Byrne, editor, *Learning and Memory: A Comprehensive Reference (Second Edition)*, pages 295–312. Academic Press, Oxford, second edition edition, 2017. ISBN 978-0-12-805291-4. doi: https://doi.org/10.1016/B978-0-12-809324-5.21085-7. URL https://www.sciencedirect.com/science/article/pii/B9780128093245210857.

P. W. Koh, T. Nguyen, Y. S. Tang, S. Mussmann, E. Pierson, B. Kim, and P. Liang. Concept bottleneck models. In *International conference on machine learning*, pages 5338–5348. PMLR, 2020.

R. Kohavi and F. Provost. Glossary of terms, 1998.

T. Lesort, V. Lomonaco, A. Stoian, D. Maltoni, D. Filliat, and N. Díaz-Rodríguez. Continual learning for robotics: Definition, framework, learning strategies, opportunities and challenges. *Information fusion*, 58: 52–68, 2020.

J. Nasir, D.-H. Kim, and J.-H. Kim. Art neural network-based integration of episodic memory and semantic memory for task planning for robots. *Autonomous Robots*, 43(8):2163–2182, 2019.

J. M. C. Ocana, R. Capobianco, and D. Nardi. An overview of environmental features that impact deep reinforcement learning in sparse-reward domains. *Journal of Artificial Intelligence Research*, 76:1181–1218, 2023.

P. K. Prasad and W. Ertel. Knowledge acquisition and reasoning systems for service robots: A short review of the state of the art. In *2020 5th International Conference on Robotics and Automation Engineering (ICRAE)*, pages 36–45. IEEE, 2020.

R. Q. Quiroga. Concept cells: the building blocks of declarative memory functions. *Nature Reviews Neuroscience*, 13:587–597, 2012.

A. P. Shimamura. Hierarchical relational binding in the medial temporal lobe: the strong get stronger. *Hippocampus*, 20(11):1206–1216, 2010.

R. S. Sutton and A. G. Barto. *Reinforcement learning: An introduction*. MIT press, 2018.

J. Thiele, O. Bichler, and A. Dupret. Event-based, timescale invariant unsupervised online deep learning with stdp. front. comput. neurosci. 12, 46 (2018), 2018.

S. Thorpe. Timing, spikes, and the brain. In *Time and Science: Volume 2: Life Sciences*, pages 207–236. World Scientific Publishing Europe Ldt, 2023.

S. Thorpe, T. Masquelier, J. Martin, A. R. Yousefzadeh, and B. Linares-Barranco. Method, digital electronic circuit and system for unsupervised detection of repeating patterns in a series of events. Patent US20190286944A1, 2019.

E. Tulving. How many memory systems are there? *American psychologist*, 40(4):385, 1985.

A. Wang, W.-N. Lee, and X. Qi. Hint: Hierarchical neuron concept explainer. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 10254–10264, 2022.

L. Wang, X. Zhang, H. Su, and J. Zhu. A comprehensive survey of continual learning: theory, method and application. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2024.

W. Wang, A.-H. Tan, and L.-N. Teow. Semantic memory modeling and memory interaction in learning agents. *IEEE Transactions on Systems, Man, and Cybernetics: Systems*, 47(11):2882–2895, 2016.

Y. Wang, Q. Yao, J. T. Kwok, and L. M. Ni. Generalizing from a few examples: A survey on few-shot learning. *ACM computing surveys (csur)*, 53(3):1–34, 2020.

C. J. C. H. Watkins. Learning from delayed rewards, 1989.

R. Zabounidis, J. Campbell, S. Stepputtis, D. Hughes, and K. P. Sycara. Concept learning for interpretable multi-agent reinforcement learning. In *Conference on Robot Learning*, pages 1828–1837. PMLR, 2023.

Appendices

The network's parameters were set as follows:

- Interface neurons:
  - STL = 50
  - Spike threshold for Failure = 120
- O-neurons:
  - WS_O = 40
  - for o in $\boldsymbol{O}$, $22 \leq STO(o) \leq 32$
  - $TnbFiredO$ = 12
  - upper bound for $lastSpikedO$ initialization = 2,000
  - $TnbQueryO$ = 6
  - b = 50
- A-neurons:
  - WS_A = 30
  - $noiseA$ = 2
  - $TnbQueryA$ = 4
  - $learnAT$ = 55
  - upper bound for $lastSpikedA$ initialization = 20,000