A Global Workspace model implementation and its relations with Philosophy of Mind

Eduardo C. Garrido-Merchán, Martín Molina, Francisco M. Mendoza-Soto

[1] Universidad Pontificia de Comillas, Madrid, Spain eduardo.garrido@uam.es
[2] Universidad Politécnica de Madrid, Madrid, Spain martin.molina@upm.es
[3] franciscomanuel.mendoza.soto@alumnos.upm.es

Abstract. This work seeks to study the beneficial properties that an autonomous agent can obtain by imitating a cognitive architecture similar to that of conscious beings. Throughout this document, a cognitive model of an autonomous agent based in a global workspace architecture is presented. We hypothesize that consciousness is an evolutionary advantage, so if our autonomous agent can be potentially conscious, its performance will be enhanced. We explore whether an autonomous agent implementing a cognitive architecture like the one proposed in the global workspace theory can be conscious from a philosophy of mind perspective, with a special emphasis on functionalism and multiple realizability. The purposes of our proposed model are to create autonomous agents that can navigate within an environment composed of multiple independent magnitudes, adapting to its surroundings to find the best possible position according to its inner preferences and to test the effectiveness of many of its cognitive mechanisms, such as an attention mechanism for magnitude selection, possession of inner feelings and preferences, usage of a memory system to storage beliefs and past experiences, and incorporating the consciousness bottleneck into the decision-making process, that controls and integrates information processed by all the subsystems of the model, as in global workspace theory. We show in a large experiment set how potentially conscious autonomous agents can benefit from having a cognitive architecture such as the one described.

1 Introduction

Artificial consciousness [9], also known as machine consciousness, is a subfield of artificial intelligence that studies how computer science implementations of models coming from the neuroscience community [13] that are related to consciousness can add value to the behavior shown by computational models or in the most optimistic scenario arise consciousness according to different philosophy of mind theories [8]. With the recent advancement of technology and machine learning methods [36], the field of machine consciousness can propose new implementations of cognitive architectures using the latest proposed methodologies and techniques. For example, we could simulate dreams or imagination being theoretically perceived by a machine by training generative adversarial networks [12] on huge datasets of images like Imagenet [16] being feed by more images cap-

tured by the robot while it is awake and classified by a deep neural network [35]. We could also simulate measures of different conflicting emotions, like joy and the necessity of sleep, in constrained multi-objective black-box optimization problems [19] where each black-box receives rewards from a deep reinforcement learning setting [33] where each agent interacts with its environment. These architectures were infeasible to implement in a robot without the current technologies and hardware advancements. Hence, we believe that the machine consciousness field must focus on implementing cognitive architectures, such as the global workspace theory [38], on, first simple but showing its effectiveness, computational models, and then complex models such as the ones described.

Not only has artificial consciousness advances but also discussions coming from philosophy of mind regarding consciousness [31]. As artificial intelligence [42] and neuroscience advance, new questions are being discussed by philosophers over these improvements. For example, the amazing results given by the natural language GPT-3 deep generative model [20] have been analyzed from the point of view of consciousness by David Chalmers, Amanda Askell or Carlos Montemayor [52]. From the neuroscience perspective, Stanislas Dehaene uses the concepts of access consciousness, concretely what information are we aware of and what information is processed unconsciously, and phenomenal consciousness, why and how do we subjectively perceive the world as we do, originally distinguished by Ned Block [6] for neuroscience research, studying in particular access consciousness and leaving phenomenal consciousness apart as a problem that is not yet solved [14]. Nevertheless, at least as far as we know, this does not happen the other way round and computer science does rarely use philosophy of mind concepts and neuroscience ideas to keep developing its methods. For example, deep learning methods do not worry about the hypothesis that if living beings are the result of evolution and are conscious, then consciousness must give an evolutionary advantage over mere unconscious processing of information such as the one happening in deep learning. The motivation of the present work is to explore the utility of cognitive models that in the future could be used, for example, to improve the quality of autonomous robotic systems (e.g., better performance efficiency). We do not intend to deliver a state-of-the-art deep learning method but only to start exploring how computer science models may imitate the flow of information that living beings do to see whether that processing of information produces a reasonable behavior. The obtained results from the experiments of the implementation of our approach may also be useful to provide quantitative empirical evidence that supports the hypothesis that simulating consciousness of certain internal perceptions, like being conscious of happiness, in biological or artificial beings can enhance the behavior of an agent concerning its environment. We also discuss how our proposed cognitive architecture relates with the main philosophy of mind ideas regarding consciousness in an attempt to be a bridge between the two disciplines and motivate the study and implementation of philosophy of the mind ideas in computer science. In particular, as our work depends on some of the assumptions taken by functionalism [32], we describe this view, and other related ones such as logical behaviorism by Ryle, in detail

and review the work of authors such as Smart, Putnam, or Dennett. We will analyze how, according to this view, machine consciousness and, in particular, our proposed model could potentially simulate the perception of measures such as happiness by an autonomous agent that enhances its performance. We would also like to briefly mention and comment how other theories such as dualism from Descartes and Chalmers point of view or panpsychism could view our work to offer to the artificial intelligence reader a quick overview of how different philosophy of mind ideas can impact an artificial consciousness model or derive different artificial consciousness models.

Previous work has been done regarding artificial consciousness [40,48]. The literature contains the implementation of robots simulating correlated behaviors of beings exhibiting consciousness [45]. Other works include artificial consciousness cognitive software architectures such as a neuronal model of a global workspace for cognitive tasks [15]. The consciousness prior model [4] is an alternative to this approach based on sparse factor graphs that have some similar elements to our proposed architecture. It will be further discussed in the philosophy of mind section of this paper. Machine learning [1] and deep reinforcement learning [33] techniques can manage a big quantity of information and could be incorporated in both our proposed architecture and in other ones as further extensions.

It is important to classify our model to have a perspective about the usefulness and significance that it adds to the artificial consciousness community. Artificial consciousness models and implementations are classified into 4 levels depending on the characteristics of the implementation [24]. The models belonging to level 1 just exhibit external behavior associated with consciousness. For example, chatbots that emulate conversations with humans successfully with generative models are examples of level 1 machine consciousness level implementations. The second level refers to machines with the cognitive characteristics associated with consciousness. These characteristics can include cognitive modules such as inner speech [39]. Machines with an architecture that is claimed to be a, ideally, cause or correlate of human consciousness are included in level 3. Finally, systems that can arise from potential phenomenal experiences would be considered level 4 machine consciousness implementations. Our proposed approach is, arguably from the philosophy of mind perspective and their particular views about consciousness, level 3 cognitive architecture system as, from our point of view, we do not have the tools to objectively determine whether level 4 systems could be implemented in practice. However, this will be discussed in the philosophy of mind section of this work. It is critical to remark the importance of philosophy of mind in machine consciousness models and, in particular, in our model. Concretely, as we objectively can not measure whether a system is conscious without relying on the assumptions given by functionalism, philosophy of mind doctrine that states that a mental state of a particular type does not depend on its internal constitution, but rather on the way it functions, or the role it plays, in the system of which it is a part [32]. Hence, we can emulate mental states with artificial intelligence, which

relies on software, that could be potentially conscious of certain perceptions in an autonomous agent.

This manuscript is organized as follows: We have already discussed related work and an introduction to artificial consciousness. We will now propose the different involved modules and models of the cognitive architecture for a potentially conscious agent and possible enhancements of them. We then present the cognitive architecture that consists of the previously described modules. Once that the model has been explained, we discuss the connections that our proposed approach has with the philosophy of the mind theories and how our model fits some of these theories, like the functionalism theory. We continue with an exhaustive set of experiments where we show how an agent implementing our architecture exhibits an autonomous behavior. Finally, we conclude the paper with some conclusions and future work proposals.

2 The autonomous agent proposal and its environment

In this section, we will describe each of the modules of the robot and how are they related to the Global Workspace Theory. But before describing the modules of the autonomous agent, we need to define the environment where it is going to be placed. The environment is a finite $2D$ space represented as a grid $\mathbf{G}$ of dimension $N \times N$ cells. Each cell contains observable physical magnitudes. In the real world, for example, these magnitudes could correspond to temperature, humidity, light, noise, presence of obstacles, etc. We use abstract magnitudes identified by $m_1, .., m_n$. The value of each physical magnitude is represented by a normalized real number in the interval $[-a, +a]$(for example, $a = 10$).

The robot can perceive magnitudes using sensors. When the robot is located in a particular cell, it perceives information of each adjacent cell from every direction $(up, down, left, right)$. If, for example, there are five magnitudes, the robot perceives a total of 20 values (4 directions multiplied by 5 observed magnitudes). The robot can move in the environment by doing one of four basic motion actions: $up, down, left, right$ in a particular instant. We represent with $\mathcal{X}$ the set of possible actions that the robot may take. Now that we have described the environment of the autonomous agent we will illustrate each of the modules that help it to navigate the environment, but first, we need to explain the theory that has inspired the modules of the autonomous agent. In particular, the global workspace theory, and why is it critical as an inspiration for the design of the agent.

2.1 Global workspace theory

One of the hypotheses that we want to support with empirical evidence with this paper is that consciousness is useful, which is an evolutionary advantage. As we will illustrate in the philosophy of mind section of the paper, we rely on the assumptions of functionalism and multiple realizability to claim that our

proposed autonomous agent is potentially conscious. But a requisite for the agent is to simulate, or more specifically to approximate as the whole set of functional connections of the brain is unknown, the functional connections that happen in the brain. If these connections are emulated in software, then following our assumptions we can state that the agent is potentially conscious. Hence, we can test the hypothesis that consciousness adds an evolutionary advantage if the behavior of the robot gives a good performance. To do so, we propose in this paper an implementation for the autonomous agent of the global workspace theory, that we describe in this subsection.

We find the following definition of the Global workspace theory very useful. In particular, the Global Workspace Theory can be compared to a theater of the mind, in which conscious contents resemble a bright spot on the stage of immediate memory, selected by a spotlight of attention under executive guidance. Only the bright spot is conscious; the rest of the theater is dark and unconscious, where most of the information processing is carried out. [2,3]. However, the conscious experiences recruit widely distributed brain functions that are mostly unconscious, so consciousness must be connected, in some way, with most of the areas of the brain, or modules of the system, where information is processed.

We design our proposed autonomous robot inspired by this definition of the global workspace theory. As we will see in the following subsection, the autonomous agent can be divided into a set of modules. In particular, an attentional system, a potentially conscious decision-making process, a set of memories, a degree of happiness, and an evaluative, motor, and perceptual system, among other minor sets of information. Concretely, we design an attentional system that represents the rest of the theater, dark and unconscious processing of information. On the other hand, we represent the bright spot as the decision-making process, where only the information classified as more important by the whole global workspace is processed and the potentially conscious autonomous system would be aware of the decision-making process involving the most relevant information. Most importantly, the potentially conscious process would be able to access any of the information of the attentional system if several states happen on the whole global workspace model, as we will see, to model that consciousness is widely distributed across the functions of the system. The execute guidance is modeled by the short and long-term memories along with the degree of happiness of the autonomous robot. Hence, the robot would be able to perceive that is making the decisions according to the happiness degree and its memories, completing the theater of the mind. Again, all of these modules are connected to the potentially conscious decision-making process to simulate that potentially conscious states are widely distributed along with all the systems. In the global workspace theory, different modules of the brain compete for being experienced by conscious states, we emulate this competition with the tuples of the attentional system. Concretely, we can see the global workspace as a common, bandwidth-limited communication channel [27], a kind of shared memory, in computer science-related terms. We model the bandwidth-limited communication channel, or consciousness bottleneck,

as the amount of tuples that the agent is going to be aware of at every instant. All the rest of the tuples could be potentially experienced by the decision-making process but the tuples are assigned a score by the attentional system according to other information of the global workspace model as the happiness degree of information coming from previous evaluations. Only the tuples that are assigned a bigger score are going to be potentially perceived by the autonomous agent, passing the consciousness bottleneck and entering the shared memory, simulating the competition and the bandwidth-limited communication channel proposed in the global workspace theory.

We have already described how we have modeled the theater of the mind, where the information is processed, but we also need to model how the decisions made by the potentially aware robot influence the external behavior of the robot and how the agent gets information from its environment. We gain the inspiration for our modeling decisions regarding these components from the five main types of processors connected to the global workspace described in a neuronal model of a global workspace in effortful cognitive tasks [15]. Concretely, these modules are long-term memory and an attentional system, that we have described in the previous paragraph and that are directly connected to the decision-making process, a perceptual system, a motor system, and an evaluative system. As a consequence, we model the information coming from the senses to the mind via a perceptual module that is directly connected to the tuples of the attentional system, potentially selected by the decision-making process. We also model the actions performed by the body using a motor module, that acts according to the winner tuple according to our global workspace implementation. Finally, we design an evaluation system that simulates how the autonomous agent perceives the environment by modifying the happiness degree, a perception that changes the whole behavior of the global workspace model and, hence, it is potentially perceived by the autonomous agent.

Our proposal is not the only implementation of the global workspace theory, consciousness is hardly referenced in cognitive architectures but there exist a few pointers in the literature regarding this concept. Related work includes an architecture that adopts a model of information flow from global workspace theory that incorporates approximations to the concepts of consciousness, imagination, and emotion [46]. This system is capable of generating a cognitively enhanced motor response to an ongoing situation based on imagining how a future situation might affect the system. The difference between this proposal and our work is that we put the focus on the bottleneck of consciousness and the attentional system and not on imagination. A future line of research will be to propose a hybrid model that combines imagination and attention. Another interesting cognitive computational model that simulates the global workspace is the LIDA systems-level cognitive model [22]. In particular, LIDA proposes a big set of memories including features that are not studied in our proposal such as a whole set of different memories (spatial, transient episodic, perceptual associative, procedural...). However, the basic components are also covered in our proposed

architecture and the LIDA proposal is conceptual and partly computational. Our proposed autonomous agent has been completely implemented on software and we have run a full set of experiments to add empirical evidence of the usefulness of a global workspace design. Another advantage of our proposal is that all the behavior of the agent relies on the end on functional expressions and tensors of weights, which in our main further line of research, can all be modeled by deep neural networks and generative models that will learn from experience. Now that all of the modules of the system have been explained from a global workspace theory point of view and linked to the concepts of global workspace theory, we provide in the following section a detailed technical description of these modules of our proposed autonomous agent.

2.2 Modules of the cognitive architecture of the robot

Having described the environment, the global workspace theory that serves as an inspiration for this implementation, and how the robot moves, we describe the modules of our proposed implementation of a global workspace model architecture and how they are related to the global workspace theory. These modules are an attention process that filters information that is not necessary for the robot to make a conscious decision, the long and short term memories, the sensors that the robot uses to detect information from the environment, the evaluation system, the motor system, the degree of happiness that the robot possesses and the most important process, the decision-making process. The whole architecture can be seen as a global workspace theory implementation, as the tuples of the attentional process are only raised to a potentially conscious state according to the information of all the modules and the decision-making process also decides the next action as a function of all the information coming from the system.

We define that a robot has a set of feelings. In this model, we only consider one type of feeling that corresponds to the happiness degree. It is represented as $h \in \mathbb{R} : h \in [0, 10]$ and can be initialized, for example in $h = 5$ in such a way that higher values are good and lower are bad. If happiness degree decreases too much passing a threshold τ, for example, $\tau = 2$, the robot deactivates, simulating death. As explained below, in this setting, the objective of the robot is to optimize its happiness degree h w.r.t a single step in the grid $\mathbf{G}$. In other words, the decision x^* that the robot takes follows $x^* = \arg\max_{x \in \mathcal{X}} f(x, t|h, \mathbf{\Theta}, \mathbf{p})$, where $\mathcal{X}$ is the set of possible actions that the robot may take in time t conditioned to its happiness degree h and in multiple flows of information $\mathbf{\Theta}$, which are the subconscious importance that the robot gives to every possible action x, and the short and long term memory about the actions that the robot is aware of. Lastly, $\mathbf{P}$ is a matrix of preferences that describes, for every sensor s_k of the grid $\mathbf{G}$, the values of the sensor that make the objective h of the robot change. For the sake of simplicity, in this implementation, we have used a single objective $f(\cdot)$ (with the degree of happiness h) but in a more realistic situation, this should be represented by a set of objectives $\mathbf{f}$. The problem that appears is that, for any time t, we can only evaluate f for a single action x. As it is going to be explained

later, this function f is computed by the evaluative system and f is estimated for $\mathcal{X}$ by the attentional system $a(\mathcal{X}|\theta) \approx f$ where θ is the set of weights of the attentional system and also estimated for the most relevant actions according to the attentional system by the decision-making process $g(a(\mathcal{X}|\theta)|\mathbf{\Theta}) \approx f$, employing all the information in $\mathbf{\Theta}$, that includes weights for the most relevant actions decided by the attentional system (subconscious decision) $a(\mathcal{X}|\theta)$, the short $\mathbf{s}(t)$ and long term memory $\mathbf{l}(t)$ and the current state of the objective that we are maximizing $h(t)$. The robot hence wants to maximize, for every step t, the feeling h, emulating happiness, through $f(\cdot)$ that can only be executed once for every t by the evaluative system. In order to make a proper decision to maximize $h(t)$, it maximizes $a(\cdot)$ and $g(\cdot)$ which are approximations to $f(\cdot)$ based in the cognitive information learned from the environment by the robot.

Each robot has its preferences $\mathbf{P}$ that represents how good a certain value given by a sensor s_k observed in the environment grid $\mathbf{G}$ is for the robot. Each position g_{ij} of the grid contains magnitudes observed by sensors. Preferences are functions of the magnitudes that affect how the situation is evaluated by the robot. The evaluative system conditions the evaluation of a position and magnitude $(g_{ij}, p(m_k))$ to the preferences of that sensor of the robot $\mathbf{p}_k$. Each robot can have different preferences. More details about the parametric expressions of these functions are given in the experiments section.

A belief is a conscious conclusion inferred by the robot that can be stored in the memory as experience to be used later as a tuple $(r|m_l, s_k, t)$, that is, a movement m_l (up, down, left, right) exploring a sensor s_k at time t that has been evaluated with reward r. This conclusion is given by the evaluative system as a result of evaluating $f(x^*, t|h, \mathbf{\Theta}, \mathbf{p})$. The information that we need to store and that conditions all the weights of the system is basically how good or bad is moving in a certain direction m_l. These beliefs tuples, B, are going to be stored in the long term memory $\mathbf{l}$ if a threshold β_l is reached when we evaluate an action by the evaluative system $|f(m_l, s_k)| > \beta_l$ or in the short term memory with another threshold. Beliefs are ranked in the memory according to their relevance $f(\cdot)$. To simulate forgetting beliefs, we use the parameter l that represents the maximum number of beliefs in the memory. In the long-term memory, less relevant beliefs are deleted first. When beliefs have the same degree of relevance, the oldest beliefs are deleted first. In the short-term memory, new beliefs substitute older beliefs independently of their degree of relevance.

The reactive attention expresses how the robot pays attention [21] to observations of sensors s_k performed on each position of the grid g_{ij}. The attentional system represents the subconscious decision about the importance of all the inputs given by the perceptual system of the robot. The subconscious system of the robot pays attention to all the information of every position that the robot performs in time $t+1$ and of all the sensors that the robot can evaluate in time $t+1$. That is m possible movements multiplied by k sensors generating mk possible actions that the robot can perform at time t. The consciousness bottleneck limits the amount of information that a human being can pay attention to in a certain time

t, but a human brain is processing all the information given by the human senses, that is, the perceptual system of the robot. This is done in the subconscious mind, here the attentional system of the robot. This attentional system then is going to select A possible actions from the available mk, for example, $A = 2$ that is going to transfer to the decision-making process. In order to select which are the most important actions x, every possible mk action is represented by a weight $w_i \in [0, 1] : \sum_{i=0}^{mk} w_i = 1$, that is, a multinomial distribution of weights that is initialized randomly or uniformly. The decision about which is the most important action is done probabilistically according to the weights and to the inputs given by the perceptual system x_{mk}. The attentional system does not have access to the evaluation system because the action has not been performed and the robot has not taken a conscious decision about the next step.

We sample A indexes from that multinomial distribution and those indexes determine the actions that are going to be transferred to the decision-making process $a \sim M(\mathbf{w})$. These weights are also going to be conditioned by the happiness degree of the robot to represent the survival instinct of consciousness. If the robot has a value for h far from the death threshold, then, exploration of new sensors and movements is encouraged and the selection is done probabilistically as described. However, if the robot feeling value is close to death, $h < \tau + \epsilon$ then the selection will be done with a ranking of the weights. The idea is to exploit the actions that have been good in the past.

Another mechanism that needs to be described of the attentional system is how are the weights $\mathbf{w}$ adapted from the evaluations done by the evaluation system $f(\cdot)$. These adaptations δ of the weights are going to be determined by the quality of the evaluation $f(\cdot)$ and the feeling objective value h, that is $\delta = e(f(\cdot), h)$. For $e(\cdot)$, we propose the following function, if the robot is not close to death $h \geq \tau + \epsilon$, then, we propose a learning rate sampled from a Gaussian distribution $r \sim N(0.05, 0.001)$. If the evaluation done by the evaluative system f has lower value than the short and long term memory thresholds and is positive $f(\cdot) > 0$, the selected weight is incremented by $w_* = w_* + r$ and the rest of the weights are normalized $\mathbf{w}_n = \mathbf{w}_n - \frac{r}{mk-1}$ where $\mathbf{w}_n$ does not contain w_*. If it is negative $f(\cdot) < 0$, the selected weight is decremented by $w_* = w_* - r$ and the rest of the weights are normalized $\mathbf{w}_n = \mathbf{w}_n + \frac{r}{mk-1}$ where $\mathbf{w}_n$ does not contain w_*.

If the evaluation done by the evaluative system has involved storing a belief in the short-term memory $|f(\cdot)| > \beta_s$, then the weights are varied with a higher learning rate, for example, $r \sim N(0.1, 0.001)$. If the evaluation is extreme, involving the long-term memory, the learning rate is also higher, for example, $r \sim N(0.2, 0.001)$. Lastly, a special update process occurs when the robot is close to death, modifying drastically the weights to not continuing a bad behavior or encouraging good behaviour. In this scenario, the weights are modified by an extreme learning weight which is $min(1, N(1 - (h - \tau), 0.001))$.

The selected tuples (m_i, s_k) fed to the decision-making process are weighted for this system by a quantity $w_a \in [0, 1] : \sum_{a=0}^{A} w_a = 1$ proportional to the sum of

the weights of the selected tuples. More formally, the computation that needs to be done is: $w_a = \frac{w_a}{\sum_{a=0}^{A} w_a}$.

The goal of the robot is to move in the environment to optimize its feelings. In our case, it corresponds to maximize the degree of happiness h as a single objective f. In a more realistic scenario, the robot would use a set of feelings $\mathbf{f}$ which correspond to multi-objective optimization.

At every time step t, the robot updates the happiness by the evaluation of a grid position and a sensor. The current state of happiness $h(t)$ is updated by the reward obtained by the evaluation system r as $h(t+1) = h(t) + r$. In addition, at every time step t, the robot loses a fixed quantity of happiness, for example $c \sim N(0.01, 0.001)$. The complete operation over the objective at every time t is $h(t+1) = h(t) + r - c$. To avoid death, the robot must ensure that its happiness is above a configurable threshold τ. That is the reason why, to survive, the robot must be always moving.

The perceptual system gets information of the grid g_{ij} and the sensors s_k in the form of observations of physical magnitudes given by the sensors of the robot and fed to the attentional system (subconscious mind) and the evaluation system.

The decision-making process makes a conscious decision about the action that is going to be performed at time t concerning the A tuples (m_i, s_k) given by the attentional system based on the importance given to those tuples by the attention system $a(\mathbf{x}|\mathbf{w})$, the short term memory $s(\mathcal{B}_s)$, where $\mathcal{B}_s$ is the set of short term beliefs, the long term memory $l(\mathcal{B}_l)$ and the degree of happiness h. The phenomenal experience (or subjective experience) is the following decision problem: which action to take $(up, down, left, right)$ based on the a tuples about observed physical magnitudes, or sensors s_k and directions m_i served by the attention module. We do only select A tuples here to emulate the bottleneck of consciousness, although we are experiencing a bubble of data at each time t, we are only aware of a subset of this information, represented here by these tuples.

To make the decision of what is the tuple that I am most interested in, that is, the action that I want to make at time $t+1$ and the action that needs to be ordered to do to the motor system, the decision-making process assigns weights to every module involved in the decision, representing the importance that we consciously give to our instincts, represented by the attentional system $a(\cdot)$ and weighted as w_a; to the short term memory beliefs (if that tuple is contained in the set of short term memory beliefs) w_s and to the long term memory w_l these weights must sum 1.

Every action $(m_l, s_k)_a$ will be assigned the following punctuation $g(a(\cdot)|a(\mathbf{x}|\mathbf{w}), s(\mathcal{B}_s), s(\mathcal{B}_l), h(t), t) = f_h(w_a a(\cdot) + w_s s + w_l l|h)$, where $a(\cdot)$ is the importance given to the tuple by the attentional system, $s = 1$ if the short term memory has determined that the tuple is positive and $s = -1$ if it is negative, $l = 2$ if the long term memory has determined that the tuple is positive and $l = -2$ if it is negative. The function f_h depends on the result of this operation

and the degree h of happiness. If the robot is in a good state, for example $h > 5$, scores are not modified to encourage exploration. If it lies in a risk state $h < 5$ and $h > \tau + \epsilon$, then the scores are modified by doubling the best score and decrementing to a half the worst score. If the situation is critical $h \in [\tau, \tau + \epsilon]$ the score given to the best tuple is 1 and the other tuple recieves value 0 just to focus on the best tuple.

The final decision is also done probabilistically, simulating rational doubts. Scores of the A tuples are normalized with respect to its sum are recomputed as: $w_a = \frac{w_a}{\sum_{a=0}^{A} w_a}$. Then, we select at random from the multinomial distribution given by those weights.

In the beginning, each module is initialized a weight uniformly. The decision-making process then outputs the selected tuple to the perception system for it to inform the values of the sensor to the evaluation system and the motor system to move to the new position m_l and measure the sensor s_k.

The weights of every module w_a, w_s, w_l are updated depending on the reward given by the evaluative system. If the criterion chosen by the attention was positive, the attention weight is incremented by 0.05 and the others decremented and normalized as in the attention module. The contrary operation is done if the reward is negative. If stored beliefs corresponding to the selected tuple exist in the memories and they have the same criterion as the evaluative system, the memories are updated by the same mechanism as in the previous weight and vice-versa. The update procedure is done iteratively and in a random order to not give preference to any particular module.

Finally, the goal of the evaluative system is to evaluate the current information fed by the perceptual system that the global workspace implementation via all the modules of the system has selected $f(g(\cdot)|\mathbf{p})$, this evaluation modifies the degree of happiness, which is the feeling that is being optimized and it can generate beliefs that can be stored in the short-term and long-term memories. Once an action is decided to be taken by the global workspace, the observed physical magnitude that has been used in this decision is sent to the evaluative system to be potentially remembered as experience if the evaluation $f(\cdot) > \beta_s$ and $f(\cdot) > \beta_l$ for the short and long term respectively. Also, depending on $f(\cdot)$ the weights of the attentional system and the decision-making process are modified as it has been described in their respective sections of this manuscript.

The operation performed by the evaluative system is the following one. First, we apply the preference function that corresponds to the sensor s_k that the perceptual system has measured, that is $r = \mathbf{p}_k(s_k(g_{ij}|t))$, obtaining the reward r that modifies the happiness degree.

3 Cognitive architecture

As we have seen by the modules involved in the architecture, our proposed cognitive architecture is heavily inspired by the global workspace model [15].

The robot takes a decision based on several modules: long-term and short-term memory, an attentional system that filters information that is not relevant, and its degree of happiness. The attentional system receives a big set of information coming from the sensor of the agent in an analogy of the information coming from the senses that enter our brain. Despite the big amount of information that the senses perceive we are not conscious of all the information in each instant of time. We make our conscious decisions based on the information that we consciously perceive, which is a subset of the information processed by the brain. The unconscious mind filters the information that it considers to be useless for the conscious mind. Figure 1 shows the structure of the cognitive architecture. We would like to clarify that the proposed architecture makes the robot able to

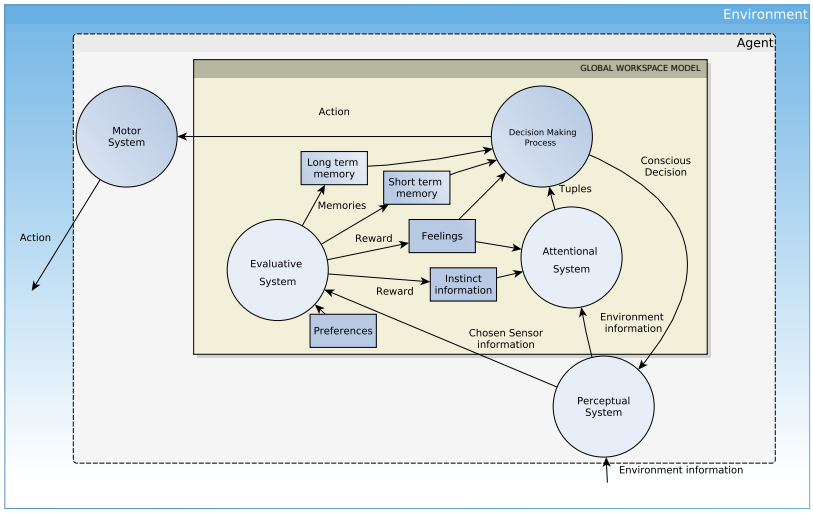

Fig. 1. Architecture of the models described in the Model section

adapt to an environment that suffers changes and it is an exploration-exploitation trade-off. We can assume an environment that changes concerning time in a smooth manner. For example, the temperature, the light coming from the sun or the humidity changes smoothly concerning time. By optimizing the happiness measure, the short-term memory can learn to determine the best actions when the environment changes whilst the long-term memory may only store the most important information that may not change so drastically according to the smoothness assumption. Hence, if the environment does not change the proposed cognitive architecture will, in the end, exploit promising solutions after exploring thanks to the f_h function, that encourages exploration by not considering to perform the best action according to the modules of the cognitive architecture if happiness is not low, but if the environment changes the cognitive architecture

will always tend to perform an exploration and exploitation trade-off, trying to position the robot and neighborhoods where it is safe according to the different modules of the system. Instead of simply considering a reward that modifies the behavior of the agent, optimizing the happiness degree modifies the whole set of parameters and the behavior of our cognitive architecture. For example, if the robot has a critical happiness degree, it performs lower risk decisions by only considering the best action, evaluated by the f_h function, according to the cognitive architecture. On the other hand, the robot encourages exploration by deciding randomly according to a multinomial distribution given by the weights $\mathbf{x}$ which is the next action to perform. We can see how the robot happiness metric tries to emulate our emotions, completely modifying, or more specifically conditioning, the way that we consciously perform actions and not just being a reward system that memorizes promising areas. In section 5 we will add empirical evidence that supports the described claim.

4 Philosophy of mind connections

Artificial consciousness is a sub-field of computer science that makes assumptions disputed by the philosophy of the mind community, since the potential success and implications of the implementations provided by the field regarding consciousness depend on the assumptions made by, mainly, the functionalism perspective, one philosophy of mind theory that is described on this section. However, artificial consciousness rarely ever discusses these concepts although other fields such as neuroscience do so, as we have described in the introduction. Depending on the philosophical perspective, several models coming from the machine consciousness community could be proposed. For instance, if one assumes the principles of naturalistic dualism, that will be described in this section, a proper modelization of the Cartesian theatre and how all the other computational modules interact of the implementation will be critical to study, from the machine consciousness perspective, how to attract consciousness to the autonomous agent. However, in this paper, we assume the principles from the functionalist perspective, which will be explained in the next subsection, so we do not need to model the Cartesian theatre nor to assume its implications, changing our proposal. This is the reason why we have emulated happiness as a quantitative measure that conditions the behavior of our whole global workspace model implementation, making the behavior change according to the value of happiness and, hence, making the robot potentially conscious of how its behavior changes according to the happiness metric. Our implementation hence depends on functionalism and does not make sense from a dualist point of view. Hence, it is important to review some philosophy of mind ideas and how the model proposed in this work, or another machine consciousness works, is related or could be a proper implementation according to these ideas. We begin the section with a description of functionalism and how it is related to this work. We then introduce a subsection describing other philosophy of mind theories and how these theories could be targeted with new models coming from the machine consciousness community.

4.1 Functionalism and related theories

If we assume that consciousness can emerge independently of the materials and reactions but dependently on the physical states and the functional processing of information, then, artificial consciousness could be implemented in computers. This is basically what functionalism defends about artificial intelligence [34]. In functionalism, the mental state is defined by its position on a causal chain and its dependence on external stimuli, previous behavior, and mental states [29]. The mental states are functional, transformations performed over representations such as how our global workspace implementation changes its behavior according to the happiness measure. In this case, our robot has as mental states the ones given by its position, its memories, its happiness state, and the weights given by the attentional system. Pain and other emotions could as well be simulated with other thresholds that can vary through the information received by sensors, emulating our senses.

A critical assumption of functionalism, which in particular is especially relevant for this work and machine consciousness models in general, is multiple realizability. Concretely, we refer with multiple realizability that a potentially conscious state, like the perception of happiness in our proposed implementation, can be multiply realized in different material substrates, or by many distinct physical kinds. Hence, our autonomous agent can potentially experience happiness, or other experiences, if it shares the same functional connections like the ones that the brain of human beings do. In this work, we assume multiple realizability. As a consequence, if the global workspace theory is a valid model of the brain and we implement it in software, if we assume multiple realizability, the autonomous agent could approximately, as obviously, we do not perfectly know the functional connections of the brain, be potentially conscious of the happiness perception. We encourage the machine consciousness community to assume multiple realizability in their proposals to add theoretical evidence of the potential of their models. In particular, any global workspace model implementation that assumes multiple realizability could potentially arise consciousness in autonomous agents if it emulates perceptions such as, in this case, happiness.

Related to functionalism, physicalism defines that the mental states are a direct physical representation of our mental states or the experiences that we feel [37]. If mental states are not only correlated but completely explained by physical states we would be assuming that the identity theory is true [47]. Hence, if we could program those physical states in a machine, mental states may automatically arise in response to them. The states may be programmed, as in this proposal, as the level of an indicator and the weights of an adaptive system. These states activate in response to environmental stimuli and in doing so they causally impact our behaviors [31]. We could argue to these theories that biological beings can only be the ones that develop consciousness, but an objection resides in the fact biological beings are developed by DNA, which is essentially a code. Ironically, we can, then, see biological, and concretely human, beings as machines made up of machines that are codified by DNA, as Daniel Dennett has popularized [17]. If

we are seen as machines [44] being developed by DNA [17], why would not an agent as the one described in this work be able to develop consciousness as well if it is also essentially a code?

Functionalism can be seen as an extension of logical behaviorism, which states that the mental states correspond directly to the behavior of the agent. For example, an agent walking across a room corresponds directly to the mental state of walking across the room [43]. The moves of the autonomous agent presented in this work correspond directly to the value of all the parameters of the global workspace implementation and external stimuli. In other words, the behavior of walking across a room is not only correlated with mental states but it is a direct effect of the mental state of walking across a room. The mental state of walking across the room is the direct cause that an agent walks across the room. This thesis is the base of emulating consciousness in computers, by simulating the mental states with software and translating them into actions performed by an agent [35]. In this work, the robot performs movements according to its memories and the level of the happiness indicator. Hence, we could predict how happy is the robot based on its decisions. This is an example of how mental states, that are modeled here by simple combinations of memories and levels of an indicator, could be a mirror of external behaviors. The movements performed by the robot are also rational [28] in the sense that they are a logical implication of internal and external outputs, as in conscious beings.

Possible enhancements of our agent reside in being able to process abstract complex data representations as in the consciousness prior architecture [4]. In this architecture, an attention neural network [51] sums up the information that the sensors of a robot perceive into a complex abstract representation of it that is processed by another deep neural network which entries are the output of the attention neural network and information coming from memories and the representations of the network at the previous instant of time. We agree in modeling cognitive architectures following these principles and it would be interesting that our robot would also learn these abstract representations. Another example is a machine consciousness architecture using deep learning and Gaussian process to emulate behaviors seen in humans when they are conscious [35].

We could only assume that the proposed model or a generalization replacing all the adaptive systems based on weights by deep neural networks could potentially make consciousness emerge if we agree, and assume the functionalist arguments. However, other philosophy of mind theories, that we briefly describe in the following subsection, are not agree with functionalism. In particular, we can find in the literature which are the main objections to the functionalism theory [5,18].

4.2 Other theories and their relation with machine consciousness

Folk psychology believes in an immaterial soul, a substance that resides in a separate realm that interacts with the brain via an unknown communication system and that is basically what it is now defined as phenomenal consciousness. This

view is defined as interactionist dualism, and it was first considered by Descartes [41]. As it is also called, interactionism believes that the mind is a separate entity, or substance, than the body. In the classical dualism theory, the mind is a witness of the information that the senses coming from the physical world through the senses of the body give it. The mind would be conscious of the information coming from the senses and of internal perceptions via the Cartesian theatre, a place where all the information is displayed and the mind can perceive it. Our model is not based on this theory, as we assume that consciousness or perceptions may arise from functional processing. However, machine consciousness could propose a model that emulates the Cartesian theatre as internal representations computed by generative neural networks such as autoencoders or generative adversarial networks that process information coming from our senses. That model would assume that the internal representations generated by the deep neural models would be perceived by a mind that would be theoretically attracted in an unknown way to that information processing. However, and although being certainly a critical historical curiosity, few members of the philosophy of mind community take this sort of dualism seriously anymore. Nowadays the naturalistic dualism theory is the main exponent of this view. In particular, naturalistic dualism [7] claims that consciousness is a fundamental and irreducible property or force of nature, and hence that it cannot be reduced to other physical properties or forces. The previously described machine consciousness model would also be compatible with this view.

If we consider, as in dualism in the extreme position of defining consciousness as a process that belong to a non-physical realm [10] or that it is an irreducible property or force of nature [7], that the experiences that we feel are related with psychical, or non-physical properties that just correlate with mental, or brain, processes, we can define these properties as qualia [30]. Concretely, some examples of these experiences are being able to perceive the redness of the red color or being able to feel the taste of a cookie. In all these cases, we are the subject of mental states that have a particular subjective set of properties. In other words, we feel different subjective sensations in each of these mental states that are ineffable to describe objectively to an entity that is not a subject of them. We hence refer to as qualia the introspectively accessible, phenomenal aspects of our mental lives [50]. Qualia will be theoretically experimented by any machine consciousness system in ways that human beings can not describe, as we do not experience those phenomenal sensations.

Another discussed theory for machine consciousness, apart from the global workspace theory, that could make machines conscious is the integrated information theory [49]. This theory is based on the following axioms of consciousness: It has intrinsic existence, my experience exists from my intrinsic perspective; it has structure, it has to process a certain amount of information, it is integrated and excludes information. This amount of information is called the integrated information that the system must process to be conscious. The system would need to autonomously process cause and effect relations, which are related to in-

trinsic perspective. These relations must affect the whole system to be structured, excluding non-necessary information and being irreducible. Our proposed method is not based on this theory but in our robot the decisions taken have their cause on the happiness degree and the environment, having an intrinsic perspective. Then the robot makes a movement based on that decision. We exclude from the decision taken by the type 2 process the information that the attentional or unconscious system qualifies as not important giving its weights. Following this reasoning, but excluding the idea of the minimum amount of information processed by an agent to be conscious, we also have influence from the integrated information theory.

Finally, to conclude by giving a summary about the main philosophy of mind perspectives on artificial intelligence and studying briefly its views about the role of artificial intelligence in consciousness, we would like to mention in this section an alternative point of view of consciousness that is given by the phenomenology perspective, that studies the structures of consciousness as being experienced by a subject, from its private and observer relative point of view and its intentionality [23]. From this perspective, several philosophers as Dreyfus uses the ideas from phenomenology to support the hypothesis of denying the possibility of rising consciousness in artificial intelligence or Husserl also adopts a skeptical point of view about the idea that consciousness can be treated by mathematical analysis. However, other phenomenologists like Gallagher and Zahavi are more optimistic, adapting and revising phenomenology, about the role of artificial intelligence in cognition and consciousness. Hence, the opinions in this field regarding the role of artificial intelligence in cognition and consciousness are divided, giving us enough evidence that our studies could also be supported by these views.

We believe that we have provided critical arguments about the importance of specifying the assumptions taken from the philosophy of mind when a new machine consciousness model is proposed. In this case, our model assumes functionalist ideas. These assumptions have conditioned our modeling decisions and hence our implementation of the global workspace model. As a direct consequence, these modeling decisions also condition the results given by the experiments, which would be different if we had followed ideas coming from the dualist perspective, where we would need to implement internal representations computed by generative models and shown in a Cartesian theatre. As we do not need to implement these representations, but only perceptions, as our happiness metric, that change the behavior of the whole global workspace, making the agent potentially conscious about the happiness perception according to functionalism, the modeling changes and, most critically, the results of the experiments, that are shown in the following section, change. As a corollary, we emphasize that machine consciousness should, at least, cite the assumptions that it makes from the philosophy of mind perspective, as these assumptions condition the performance given by the autonomous agent in any set of experiments.

5 Experiments

In this section, we provide a big set of experiments to provide empirical evidence of the advantages that cognitive architectures such as the one described in this work could give to agents.

We first define the environment that our model will interact with is represented as a list of m magnitudes each of one can be represented as a bi-dimensional grid of size $N \times N$. Also, one condition that was established is that there should not be high differences in the magnitude value of any pair of adjacent cells. Each position of the grid holds several values m that every sensor of the robot measures. A tri-dimensional grid is instantiated with size $N \times N \times m$. Each cell of the environment grid is initiated with a random real number, which is calculated using a normal distribution $\mathcal{N}(0, 1)$. For each magnitude, all the values assigned to its bi-dimensional grid are normalized so their values lay within the interval $[-\alpha, \alpha]$. We generate multiple random environments with a different size of the grid $N = [20, 100, 200]$, whose first magnitude can be seen on Figure 2 and number of magnitudes $m = [5, 20, 100]$. These environments are identified by ESxMy. Where x is the grid size and y is the number of magnitudes. In order to interpret the magnitude values into preferences, it is

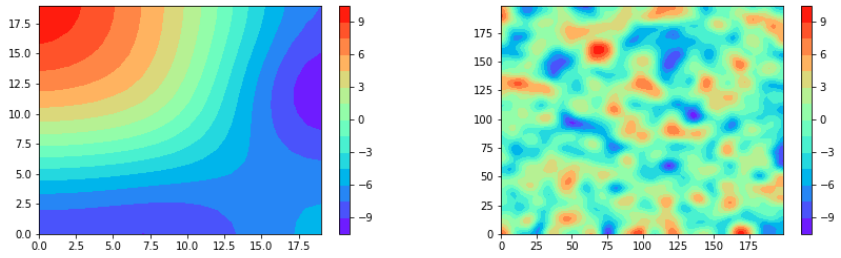

Fig. 2. Heat map of the first magnitude values of the environment ES20M5 and ES200M5.

necessary to define the preference functions related to each magnitude. All of the defined functions satisfy that any magnitude value within $[-10, 10]$, returns a preference value within $[0, 10]$, with a preference of 0 being the worst possible result for the agent and viceversa. The analytic expression of these functions are $f(x) = |10 \cdot \sin\left(\frac{x}{10} \cdot \frac{\pi}{2}\right)|$, $f(x) = |10 \cdot \cos\left(\frac{x}{10} \cdot \frac{\pi}{2}\right)|$, $f(x) = 5 + 5 \cdot \sin\left(\frac{x}{10} \cdot \frac{\pi}{2}\right)$ and $f(x) = -\cos\left(\frac{x-1}{10} \cdot \frac{\pi}{2}\right)$.

The absolute sine function gives a higher preference to extreme values while giving a lower preference to magnitude values close to zero. The absolute cosine function

has the opposite behavior to the absolute sine function. This function is good for mapping most of the physical magnitudes that can be found in real life, for example, temperature, since extreme values are, by definition, quite extreme and bad for both life and computational agents. The sine function returns a higher preference for higher magnitude values and the additive inverse sine function, which acts as the opposite of the sine function. The preference functions are mapped to magnitudes randomly.

To be able to determine how good our model is performing, it is necessary to first define the evaluation metrics that are going to be used: FH: Final happiness. MH: Mean happiness. HSD: Happiness standard deviation NI: Number of iterations. WGFP: Weighted goodness of final position. MWG: Mean weighted goodness. MTI: Meantime per iteration (milliseconds). The final happiness can tell us how good was the agent at the moment of reaching the end of its path. The mean and standard deviation happiness are calculated with all happiness values obtained at each iteration of the simulation. Another goal of the model is to design agents that can localize and travel to the best possible position inside its environment, which is defined by the position that returns the best values of goodness for all of the magnitudes of the environment, using their corresponding preference functions. The global goodness of any position will be equal to the mean of all good values for each different magnitude. The sum of weights assigned to all four observations for one magnitude, one for each possible action, is considered the importance that is given to such magnitude at any specific iteration. Therefore, this value will be used as a weight for that magnitude at the moment of calculating the weighted goodness of the position of the agent. The weighted goodness of the final positions is a good metric to evaluate how effective was the agent navigating into a position that satisfies its own idea of what is a good position.

The presented model has multiple variables that need to be defined for each simulation. Each parameter is tuned individually in all the environments according to several metrics. The initial parameters values considered are shown bellow: Number of iterations = 100. Initial degree of happiness = 5. Fatigue = 0.1. Death threshold $\tau_d = 2$. Critical near death threshold $\tau_c = 3$. Risk threshold $\tau_r = 4$. Attentional conscious limit = 2. Short-term memory capacity positions = 3. Short-term memory threshold = 2. Long-term memory capacity = 2. Long-term memory threshold = 4. Adaptation weights = $[0.05, 0.1, 0.2]$.

After initializing all variables with the values described, the model was tested against all the nine environments defined before. The obtained results can be observed at the table 1. Multiple conclusions can be made after analyzing these results. First, it can be seen that out of all nine environments that were used to test the agent, it was able to survive in eight of them, only dropping its happiness under the death threshold in one environment, ES100M5. The overall mean happiness is also pretty high.

The weighted goodness values of final positions can also be considered good. We can see that the highest MWG was the one obtained in the only simulation that

ENV	MH	MWG	FH	HSD	NI	WGFP	MTI
ES20M5	8.3425	3.2180	8.0033	2.0070	100	2.4248	0.3700
ES20M20	8.5641	2.3737	10.0000	2.0624	100	2.6874	0.6400
ES20M100	7.8666	3.3172	6.3860	1.7848	100	3.6951	2.3700
ES100M5	3.4002	6.5684	1.9108	1.0258	18	6.3144	0.3333
ES100M20	4.9274	3.6557	6.8798	0.9812	100	2.8410	0.7500
ES100M100	7.9255	2.0716	9.6474	2.4830	100	1.1165	2.7100
ES200M5	8.4133	1.6965	10.0000	1.8782	100	0.0932	0.4500
ES200M20	8.8552	1.6618	10.0000	1.9101	100	0.4386	1.4000
ES200M100	8.5203	2.6202	10.0000	2.0038	100	2.8214	2.5200
MEAN	7.4239	3.0203	8.0919	1.7929	91	2.4924	1.1125

Table 1. Result of evaluating the model with the initial set of parameters

ended with the agent death, which is the simulation performed in the environment ES100M5, with an MWG value of 6.5684. Overall, all values of MWG are under 4 for the eight surviving agents, and in two simulations the agent was able to achieve a WGFP under 2.0, which is a very good result. There is a clear relation between the MTI and the number of magnitudes. This relationship is caused by the additional effort that has to be made by the attentional system to select the observations that will be passed to the global workspace. Even better results might be obtained after the calibration of all parameters of the model. Figure 3 is represented the average value of happiness over time for one hundred iterations in all environments. It can be observed that the agent, on average, maintained

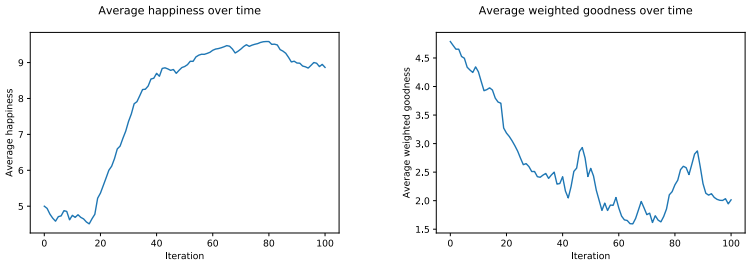

Fig. 3. Average happiness (left) and weighted goodness (right) over a hundred iterations for the initial set of paremeters tested against the nine different environments

its initial happiness, 5, for around twenty iterations before its happiness started rising rapidly, up to iteration sixty, when the agent maintained a constant value near 9. Also, in Figure 3 we can see the average weighted goodness over time.

We can see a constant improvement of the average weighted goodness, which starts at an initial value of near 5 and progresses over time up to an average

of near 2. There are some ups and downs, but overall the graph shows that the progression follows a linear improvement over time. Given this progression, it is interesting to see if a lower value of weighted goodness could be achieved with a higher number of iterations. We repeat this simulation with a maximum number of iteration set to 500, obtaining a similar set of results as in the case of 100 iterations. We observe that one more agent died before reaching the 500 iterations, the one related to environment ES100M20, at the iteration number 275. We plot the average happiness and weighted goodness in Figure 4. It can be

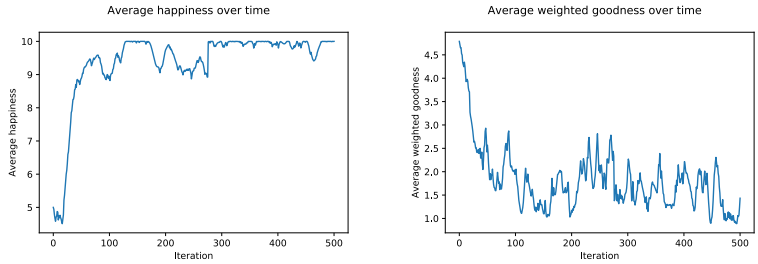

Fig. 4. Average happiness (left) and weighted goodness (right) over a five hundred iterations tested against the nine different environments

seen that there is a fast improvement during the first 100 iterations and after this, however, agents seem to maintain on average a weighted goodness value between 1.0 and 3.0, and a happiness value between 9 and 10, converging. An interesting fact is that the direct relationship between happiness and weighted goodness: when the latter goes down, so does the former. This can be observed in the section between the iterations 200 and 300. Therefore, the maximum number of iterations was set to 500 for all the following simulations.

We now tune four happiness-related parameters that exist in the model: fatigue, initial degree of happiness, and two of the three happiness thresholds: critical near-death threshold τ_c, and risk threshold τ_r.

The first happiness-related parameter to be tuned was fatigue, which is a constant number that is substracted to the happiness of the agent in each iteration. We do not include more numerical results in tables for these parameters but talk about the conclusions extracted from the experiments. In terms of mean happiness, a tendency to obtain higher happiness with lower values of fatigue can be noticed. In terms of mean weighted goodness, lower fatigue returns better results for all values. Still, we do not want a fatigue value that is too low, since we still want to punish agents that are performing badly for a long period. Therefore, it was decided for this analysis that the best value of fatigue was 0.05.

The next parameter that was analyzed was the initial degree of happiness. As it was expected, a higher initial of happiness results in higher mean happiness.

The third happiness-related parameter that was tested was the risk threshold, τ_r. There is not a noticeable change in performance when comparing the first two tested values, 4.0 and 4.5. The value for the risk threshold that was chosen due to resulting in a better performance was 4.5.

Finally, the last happiness-related parameter that is tested is the critical near-death threshold, τ_c. The results for $\tau_c = 3$ and $\tau_c = 3.5$ does not have a noticeable difference. Using MWG as the decisive metric, the chosen for the critical near-death threshold is $\tau_c = 3.5$. The attentional limit also takes an important role in the behavior of the agent. It represents the number of magnitudes that are chosen by the attentional system in each iteration, which are later sent to the decision-making process at the moment of making a decision about which action to execute. Multiple simulations were executed testing five different values for the attentional limit, whose results can be seen in the table 2. The attentional

attentional limit	MH	MWG	FH	HSD	NI	WGFP	MTI
1	7.9394	2.9516	7.2807	1.3820	373.0000	3.1601	1.0873
2	8.7627	2.2434	8.1974	1.2392	431.4444	1.8643	1.1254
3	8.6832	2.9691	8.0908	1.1957	425.5556	3.6208	1.1141
4	8.4873	2.6407	7.4097	1.3014	444.2222	3.1423	1.1427
5	8.8075	2.5964	8.1871	1.1160	435.3333	3.4292	1.1777

Table 2. Mean results for different attentional limit values

limit = 2 has the best performance. Its value of mean weighted goodness, 2.2434, is the lowest out of all, and its mean happiness, 8.7627, is also very high, only improved by the mean happiness obtained with attentional limit = 5. We set the attentional limit to 2.

Each of the memories of the agent has two parameters to be tuned, their capacity and the threshold. The first parameter that is analyzed is the short-term memory capacity. It can be seen that MWG-wise, the best result is obtained with a capacity of five beliefs. We then proceeded with the other parameter of the short-term memory: its threshold, τ_{stm}. In this case, it seems that decreasing the initial parameter from $\tau_{stm} = 2$ to $\tau_{stm} = 1.5$ produces slightly better results in terms of both mean happiness and mean weighted goodness, so the change is made permanent. We perform analogously with the long-term memory. Here, it seems like there is a decent improvement of the model performance when increasing the long-term memory capacity from 2 to 6, going from mean weighted goodness of 2.1758 to 2.1025. Regarding the long-term memory threshold, it seems like changing the initial values decreases the performance of the model. Therefore, the initial value of $\tau_{ltm} = 4$ stays intact.

Finally, the only parameters left to tune are the learning rates: the low learning rate, λ_l, the medium learning rate, λ_m and the high learning rate, λ_h. Regarding the low learning rate, λ_l. It can be seen that changing the learning rate produces a

higher difference in performance out of all the tested parameters, with a deviation of mean weighted goodness of almost 1 when comparing some of the simulations. The best performance is found when the low learning rate is set to its initial value, 0.05.

We set 1.5 for the final value of the medium learning rate. Lastly, regarding the high learning rate, we can appreciate that the lowest value of MWG is obtained with $\lambda_h = 0.4$, but as we want the agent to be alive in critical situations, we set it to 2.

The results obtained for each of the nine tested environments, using the tuned set of parameters, are displayed in Table 3. To analyze how well the final parameters

ENV	MH	MWG	FH	HSD	NI	WGFP	MTI
ES20M5	9.6633	0.7233	10.0000	1.0984	500	0.0396	0.3760
ES20M20	9.4097	2.4639	10.0000	1.4519	500	1.8189	0.5760
ES20M100	9.6779	1.1043	10.0000	1.0570	500	0.0292	2.0100
ES100M5	3.2294	6.5653	1.9887	0.8067	20	6.5903	0.3000
ES100M20	9.5266	1.6654	10.0000	1.4315	500	3.0864	0.7060
ES100M100	9.5062	2.3734	10.0000	1.3311	500	1.1592	2.0680
ES200M5	9.5654	2.1843	10.0000	1.2407	500	2.2895	0.3300
ES200M20	9.7824	0.6477	10.0000	0.9607	500	2.6468	0.6760
ES200M100	9.7408	0.8837	10.0000	1.0270	500	0.5856	2.6000

Table 3. Results of evaluating the model with the final set of parameters

perform compared to the initial set, both the average happiness and average weighted goodness over time have been plotted in Figure 5. On average, the final

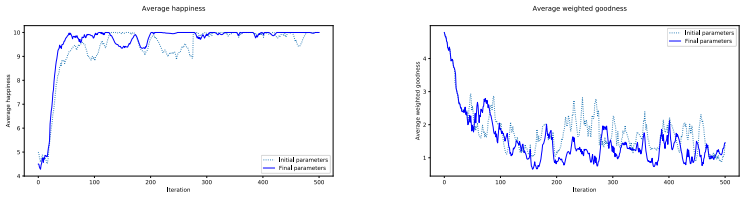

Fig. 5. Average happiness (left) and weighted goodness (right) over a five hundred iterations comparing the final (solid line) and initial parameters (dotted line)

set outperforms the initial in both metrics. We can observe that peak performance is achieved near the iteration 150 when an average weighted goodness of near 0 is achieved. The average happiness also shows good performance, keeping a value of near 10 at the iteration 100, and maintaining such value for later iterations.

Recall that any value of goodness lower than 5 is considered good. In this experiment, agents have shown to be able to maintain average goodness values between 0 and 2, which by the definition given, are very good values. Therefore, we can deduct that the model is well constructed and the agents are achieving the goal that was set for them.

We perform more experiments in five different simulations. In each simulation, nine new environments are randomly generated and each magnitude of each environment is assigned one of the four defined preference functions. The average results per simulation can be found in the table 4. The average happiness and

Simulation	MH	MWG	FH	HSD	NI	WGFP	MTI
1	7.9834	2.8295	7.2776	1.2497	349.6667	2.8194	1.1530
2	8.9011	2.4198	8.1681	1.1666	442.7778	2.9534	1.1381
3	8.7009	2.9118	7.2381	1.1820	390.5556	3.9969	1.0910
4	9.4744	1.5601	9.1094	1.4773	481.6667	2.0276	1.1219
5	8.7284	2.4389	8.2148	1.2845	416.4444	1.9841	1.1809

Table 4. Evaluation of the final model in 5 different simulations, for each simulation showing the average result of 9 different environments.

average weighted goodness are also shown over time in the plots of Figure 6. We

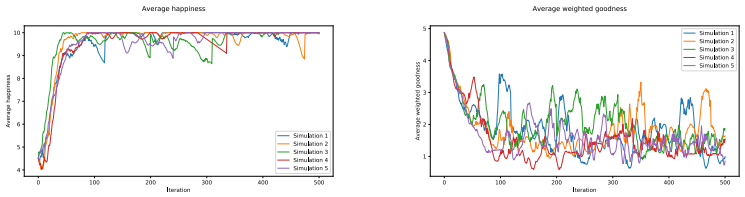

Fig. 6. Average happiness and weighted goodness of the final model over five hundred iterations, in 5 different simulations

observe that all the results are similar: they all have a starting learning curve in the first 100 iterations, and then a constant performance is maintained during later iterations. It can be seen that in none of the simulations the mean weighted goodness has a value higher than 3, and, since we consider goodness values under five to be good, given the definition of goodness, we can determine that the model is having a good performance in all five simulations. Overall, this experiment proves the capability of the model to generalize to different environments.

We can conclude the evaluation of the presented model can satisfy its two main goals: achieving and maintaining high values of happiness and low values

of weighted goodness. In particular, we have seen how the happiness measure achieves good values thanks to the behavior of different modules of our global workspace model implementation such as weights, memories, varying the way the agent makes decisions, and the consciousness bottleneck. If we for example made the agent consider only the best decision in all the situations, we would only encourage exploitation of promising results and not exploration by considering several ideas. But if we get rid of the consciousness bottleneck the exploitation exploration tradeoff will be more focused on exploration (since we would consider all the possible actions) and the computational time would be higher. Hence, we have found the consciousness bottleneck to be useful empirically as we clearly show in all the experiments that we have carried out, where the happiness metric is stable.

6 Conclusions and further work

Along with this manuscript, a prototype of a potentially conscious agent with an implementation of the global workspace model has been proposed which has shown to create autonomous agents able to navigate through environments. The work presented in this paper assumes the ideas coming from the functionalism philosophy of mind theory, where the mind is seen as a calculating machine. From the functionalism perspective, a mental state of a particular type, like happiness in this paper, does not depend on its internal constitution, but rather on the way it functions, or the role it plays, in the system of which it is a part [32]. Most importantly, we assume multiple realizability, which states that a potentially conscious state is independent of its material substrate but dependant on functional connections, enabling artificial intelligence to create potentially conscious machines. Following this reasoning, we simulate a measure of happiness that conditions the whole behavior of the agent concerning its environment and it is also affected by the evaluations of the environment, assuming that the robot is potentially conscious of the happiness perception. The results of executing multiple simulations with different sets of parameters and different environments allow us to determine that the presented model shows proper behavior, being capable to reach and maintain high values of happiness. We have modeled the bottleneck of consciousness for two reasons in this work. First, our main purpose of the paper was to test the hypothesis that computational models could benefit from our aware rational decision-making process. Biological beings are not aware of all the information and inference that our brain is making in every moment, but are only aware of the information that our brain, unconsciously, considers to be more important, an idea modeled by access consciousness. We have modeled this process via the attentional system and the type 2 process that integrates the information, only receiving the tuples that are considered to be more important from the global workspace model perspective (including the happiness measure). Second, we assume that if we make decisions that way, this is an evolutionary advantage, since we have evolved to adapt to the environment and are nowadays the "general artificial intelligence" that better adapts to new

environments and problems. Hence, by the described two points, we hypothesize that if we simulate our process of making decisions an artificial autonomous agent will deliver better results in different environments. And to support this claim empirically, we have designed the experiments section. We see there how simulating a consciousness bottleneck and making decisions according to internal preferences such as happiness that condition the whole system (emulating a global workspace model) enhances the performance obtained by an autonomous agent, adding empirical evidence to support our claims. We show how the possession of feelings has proven to be a powerful tool that grants the system multiple beneficial abilities. Feelings are a great representation of the conscious state of an agent and could serve as an example of a self-model representation. Two main advantages are provided by such a self-model. First, it gives the agents the capability to report their conscious state at any moment. Secondly, it allows the agents to have an understanding of their inner state and act according to to multiple situations. Attention filters the information that is way too big to process as a whole, deciding which observations are relevant enough to become conscious and which do not. The memory system, used to store past experiences and beliefs, also reinforces the adaptation ability that makes the agents able to learn from previous mistakes and successes. Using two different memories gives a higher degree of depth to this system. Finally, the global workspace, where decisions are made, has the information processed by the rest of the modules is integrated. If consciousness could exist within this model, it would lay here.

As further work, we can implement more complex environments where the robot would need to interact in complex ways with the environment to remain alive. We will also implement deep neural networks for all the adaptive systems and memories of the robots, expanding and implementing the model in the same fashion as in the consciousness prior. As we have seen, the model has several parameters but we do not have an analytic expression to optimize them, and computing the mean results can be expensive. Bayesian optimization has been used with success in similar problems [25,11]. In particular, we could optimize several metrics simultaneously under the presence of constraints such as the robot not dying with a constrained multi-objective Bayesian optimization approach [19,26]. Finally, we will implement several levels of feelings in the robot, for example: hungry, social needs, or need to dream. The robot actions will be a function of those feelings along with happiness in a multi-objective problem where the robot will need to keep the level of all its indicators above some threshold.

Compliance with Ethical Standards

Not applicable.

References

1. Alpaydin, E. *Introduction to machine learning.* MIT press, 2020.

2. Baars, B. J. The global workspace theory of consciousness. *The Blackwell companion to consciousness* (2007), 236–246.
3. Baars, B. J. The global workspace theory of consciousness: Predictions and results. *The blackwell companion to consciousness* (2017), 227–242.
4. Bengio, Y. The consciousness prior. *arXiv preprint arXiv:1709.08568* (2017).
5. Block, N. Troubles with functionalism.
6. Block, N. Some concepts of consciousness. *Sciences 18*, 2 (1995), 1–28.
7. Chalmers, D. Naturalistic dualism. *The Blackwell companion to consciousness* (2017), 363–373.
8. Chalmers, D. J. Philosophy of mind: Classical and contemporary readings.
9. Chella, A., and Manzotti, R. *Artificial consciousness*. Andrews UK Limited, 2013.
10. Clarke, D. *Descartes's theory of mind*. OUP Oxford, 2003.
11. Córdoba, I., Garrido-Merchán, E. C., Hernández-Lobato, D., Bielza, C., and Larranaga, P. Bayesian optimization of the pc algorithm for learning gaussian bayesian networks. In *Conference of the Spanish Association for Artificial Intelligence* (2018), Springer, pp. 44–54.
12. Creswell, A., White, T., Dumoulin, V., Arulkumaran, K., Sengupta, B., and Bharath, A. A. Generative adversarial networks: An overview. *IEEE Signal Processing Magazine 35*, 1 (2018), 53–65.
13. Crick, F., and Koch, C. Consciousness and neuroscience. *Essential sources in the scientific study of consciousness* (1998).
14. Dehaene, S. *Consciousness and the brain: Deciphering how the brain codes our thoughts*. Penguin, 2014.
15. Dehaene, S., Kerszberg, M., and Changeux, J.-P. A neuronal model of a global workspace in effortful cognitive tasks. *Proceedings of the national Academy of Sciences 95*, 24 (1998), 14529–14534.
16. Deng, J., Dong, W., Socher, R., Li, L.-J., Li, K., and Fei-Fei, L. Imagenet: A large-scale hierarchical image database. In *2009 IEEE conference on computer vision and pattern recognition* (2009), Ieee, pp. 248–255.
17. Dennett, D. C., and Dennett, D. C. *Darwin's Dangerous Idea: Evolution and the Meanins of Life*. No. 39. Simon and Schuster, 1996.
18. Eliasmith, C. The myth of the turing machine: The failings of functionalism and related theses. *Journal of Experimental & Theoretical Artificial Intelligence 14*, 1 (2002), 1–8.
19. Fernández-Sánchez, D., Garrido-Merchán, E. C., and Hernández-Lobato, D. Max-value entropy search for multi-objective bayesian optimization with constraints. *arXiv preprint arXiv:2011.01150* (2020).
20. Floridi, L., and Chiriatti, M. Gpt-3: Its nature, scope, limits, and consequences. *Minds and Machines 30*, 4 (2020), 681–694.
21. for Research into Nervous, A. A., Diseases, M., Posner, M., and Rothbart, M. Attention, self–regulation and consciousness. *Philosophical Transactions of the Royal Society of London. Series B: Biological Sciences 353*, 1377 (1998), 1915–1927.
22. Franklin, S., Madl, T., Strain, S., Faghihi, U., Dong, D., Kugele, S., Snaider, J., Agrawal, P., and Chen, S. A lida cognitive model tutorial. *Biologically Inspired Cognitive Architectures 16* (2016), 105–130.
23. Gallagher, S., and Zahavi, D. *The phenomenological mind*. Routledge, 2020.
24. Gamez, D. *Human and machine consciousness*. Open Book Publishers, 2018.

25. Garrido-Merchán, E. C., and Albarca-Molina, A. Suggesting cooking recipes through simulation and bayesian optimization. In *International Conference on Intelligent Data Engineering and Automated Learning* (2018), Springer, pp. 277–284.
26. Garrido-Merchán, E. C., and Hernández-Lobato, D. Parallel predictive entropy search for multi-objective bayesian optimization with constraints. *arXiv preprint arXiv:2004.00601* (2020).
27. Goyal, A., Didolkar, A., Lamb, A., Badola, K., Ke, N. R., Rahaman, N., Binas, J., Blundell, C., Mozer, M., and Bengio, Y. Coordination among neural modules through a shared global workspace. *arXiv preprint arXiv:2103.01197* (2021).
28. Harman, G. Rationality. *International Encyclopedia of Ethics* (2013).
29. Hierro-Pescador, J. *Filosofía de la mente y de la ciencia cognitiva*, vol. 9. Ediciones AKAL, 2005.
30. Jackson, F. Epiphenomenal qualia. *The Philosophical Quarterly (1950-) 32*, 127 (1982), 127–136.
31. Kim, J. *Philosophy of mind*. Routledge, 2018.
32. Levin, J. Functionalism, stanford encyclopedia of philosophy.
33. Li, Y. Deep reinforcement learning: An overview. *arXiv preprint arXiv:1701.07274* (2017).
34. Lycan, W. G. Mental states and putnam's functionalist hypothesis. *Australasian Journal of Philosophy 52*, 1 (1974), 48–62.
35. Merchán, E. C. G., and Molina, M. A machine consciousness architecture based on deep learning and gaussian processes. *arXiv preprint arXiv:2002.00509* (2020).
36. Murphy, K. P. *Machine learning: a probabilistic perspective*. MIT press, 2012.
37. Nagel, T. Physicalism. *The Philosophical Review 74*, 3 (1965), 339–356.
38. Newman, J., Baars, B. J., and Cho, S.-B. A neural global workspace model for conscious attention. *Neural Networks 10*, 7 (1997), 1195–1206.
39. Pipitone, A., and Chella, A. Robot passes the mirror test by inner speech. *Robotics and Autonomous Systems* (2021), 103838.
40. Reggia, J. A. The rise of machine consciousness: Studying consciousness with computational models. *Neural Networks 44* (2013), 112–131.
41. Rozemond, M., and Rozemond, M. *Descartes's dualism*. Harvard University Press, 2009.
42. Russell, S., and Norvig, P. Artificial intelligence: a modern approach.
43. Ryle, G. *The concept of mind*. Routledge, 2009.
44. Searle, J. R. Is the brain's mind a computer program? *Scientific American 262*, 1 (1990), 25–31.
45. Seth, A. Explanatory correlates of consciousness: theoretical and computational challenges. *Cognitive Computation 1*, 1 (2009), 50–63.
46. Shanahan, M. A cognitive architecture that combines internal simulation with a global workspace. *Consciousness and cognition 15*, 2 (2006), 433–449.
47. Smart, J. J. The identity theory of mind. *Stanford Encyclopedia of Philosophy* (2009).
48. Soto, F. M. M. Artificial conciousness: an approach to autonomous agents based in a. *Science 385*, 6362 (2017), 486–492.
49. Tononi, G., Boly, M., Massimini, M., and Koch, C. Integrated information theory: from consciousness to its physical substrate. *Nature Reviews Neuroscience 17*, 7 (2016), 450–461.

50. Tye, M. Qualia, stanford encyclopedia of philosophy (revised 31 july 2007).
51. Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, Ł., and Polosukhin, I. Attention is all you need. In *Advances in neural information processing systems* (2017), pp. 5998–6008.
52. Zimmermann, A. Philosophers on GPT-3. `https://dailynous.com/2020/07/30/philosophers-gpt-3/`, 2020.