Consiglio Nazionale delle Ricerche
Istituto di Calcolo e Reti ad Alte Prestazioni

A Conscious Architecture: An Initial Description

M. Cossentino, G. Pilato, G. Averna, M. Mylopoulos, J. Mylopoulos

Rapporto Tecnico N.: 2
RT-ICAR-PA-24-02 **ottobre 2024**

Consiglio Nazionale delle Ricerche, Istituto di Calcolo e Reti ad Alte Prestazioni (ICAR)
– Sede di Cosenza, Via P. Bucci Cubo 8/9C, 87036 Rende, Italy, URL: *www.icar.cnr.it*
– Sede di Napoli, Via P. Castellino 111, 80131 Napoli, URL: *www.na.icar.cnr.it*
– Sede di Palermo, Via Ugo La Malfa 153, 90146 Palermo, URL: *www.pa.icar.cnr.it*

Consiglio Nazionale delle Ricerche
Istituto di Calcolo e Reti ad Alte Prestazioni

A Conscious Architecture: An Initial Description

M. Cossentino [1], G. Pilato [1], G. Averna [1], M. Mylopoulos [2], J. Mylopoulos [3]

Rapporto Tecnico N.:2 ***Data:***
RT-ICAR-PA-24-02 **ottobre 2024**

---

[1] Istituto di Calcolo e Reti ad Alte Prestazioni, ICAR-CNR, Sede di Palermo, Via Ugo La Malfa 153, 90146 Palermo.
[2] Carleton University, 1125 COLONEL BY DRIVE, Ottawa, Canada.
[3] University of Toronto, 40 St George , Toronto, Canada.

Contents

Abstract

One of the most interesting questions in the field of artificial intelligent agents is whether it is feasible to design and implement software agent architectures that can emulate consciousness characteristics. The paper illustrates an agent architecture that builds on top of the BDI paradigm a consciousness feature by exploiting and embedding the Global Workspace and the Executive System Theories in an architecture that we named CARA (Conscious Agent Reasoning Architecture). The whole architecture is described together with the prototype implementation of a case study inspired by the "Ticket to Ride" table game.

A Conscious Architecture: An Initial Description

October 14, 2024

1 Introduction

The question of what mechanisms underlie consciousness is one of the most important in the field of cognitive science. [13]. Relatedly, one of the most pressing questions in Artificial Intelligence is if it is possible to design and implement architectures for intelligent artificial agents that are capable of emulating conscious properties [1] [10].

There are several competing theories of consciousness, and as of yet, no consensus as to which framework is the most successful has been reached [9]. For the purposes of this paper, we remain neutral on this question. The main aim of this paper is to propose a conscious agent architecture with specific features that some individual theories have taken to be necessary for consciousness. Our long-term aim is to propose an agent architecture with specific features that implement key aspects of leading theories. In particular, in this first implementation step, we adopt insights from Baars' Global Workspace Theory (GWT) of consciousness [4] [3]. To this, we include some additional functions inspired by Buehler's proposal that the executive system constitutes an agent's capacity to guide their actions [8], [7]. Therefore, in the current version of the proposed architecture, an agent is conscious of some information insofar as the agent attends to it, and that information enters into the global workspace, thus making it available to a range of psychological functions. We will show how this mechanism can be implemented by exploiting a Belief-Desire-Intention approach [12] [6].

This approach is justified by the fact that Baars treats consciousness as requiring an attention mechanism that constantly shifts an agent's focus from some active goals to others and introduces new ones [4] [3]. According to Baars' theater metaphor, [2], in our minds, there are two main parts: what we're currently focused on (the *bright spot*) and everything else around it (the *fringe*). The bright spot is like a spotlight that shines on something specific. Our brain takes information from this bright spot and uses it to guide other processes, some of which happen unconsciously. These unconscious processes can be divided into two types and characterized using the following metaphors: an audience in a theatre that receives information from the bright spot and acts consequently, or who works behind the scenes, who makes the events in the

bright spot according with the context and the background information, such as assumptions that we make, memories and spatial awareness.

Implementing the presence of the audience and of someone behind the scenes can be fruitfully done by adding to the architecture a set of functions arising from Buehler's approach to agentive guidance that draws on the functions of the executive system [8], [7]. This theory highlights the role of higher-order cognitive functions in regulating and coordinating mental activities, which play a key role in guiding behaviour and in decision-making and problem-solving tasks. It includes processes such as planning, attention, working memory, cognitive flexibility, and inhibitory control, which facilitate goal-directed behaviour. The executive system integrates different kinds of information, making the agent capable of self-monitoring and capable of adapting to complex scenarios.

Considering the diffusion of the Bratman's BDI paradigm [5][6] in cognitive agent architectures and their successful adoption even with consciousness features [12], we decided to adopt such a paradigm and incorporate that in one high-order executive function concerned with the (practical) reasoning of the agent.

In fact, BDI is an ideal candidate component for an architecture aimed to implement some sort of consciousness mechanism that includes the Global Workspace Theory and the Buehler's approach towards achieving the completion of the actions and accomplishment of the obejctive.

Following these research directions, this paper presents *Conscious Agent Reasoning Architecture* (CARA): an innovative agent architecture built on top of the BDI approach and embedding Global Workspace Theory and Executive System Theory. This integration can lead to more robust and adaptable BDI agents capable of complex reasoning and improved performance in uncertain situations. This approach makes CARA a first step towards an agent being capable of making conscious and rational decisions while pursuing multiple goals simultaneously, even under time and resource constraints. Moreover, it aims to allow for the consideration of trade-offs in achieving multiple goals, mimicking human behaviour when a compromise is necessary for partial goal satisfaction.

The CARA architecture has been fully implemented, and a first experimental setup has been realized by exploiting a scenario inspired by the 'Ticket to Ride' table game. This paper also reports the proposed solution's UML class diagram.

The remaining part of the paper is organized as follows: section 2 provides the research baseline; section 3 illustrates the requirements of the proposed architecture; section 4 describes the proposed architecture; and section 5 reports on the implementation of the CARA architecture. Conclusions and future work are given at the end of the paper.

2 Research Baseline

The proposed architecture is based on three key elements: the Baar's theory on the Global Workspace, the Buehler's Executive System functions, and the Bratman's Belief-Desire-Intention (BDI) model. In this section we briefly sum-

marize these theories and models constituents of the proposed architecture, that we named CARA.

2.1 Belief-Desire-Intention model of Bratman

The BDI model of Bratman [6] is an architecture for practical reasoning for an agent. In this model, three concepts are the core of practical reasoning: beliefs, desires, and intentions. A belief represents the state of anything the agent knows about the world; its value is related to some real-world property and can change over time with that. A desire is a change in the state of something in the world that the agent wants to achieve. Not all desires may be pursued at the same time so the agent deliberates to pursue some of them by promoting them to intentions. Intentions are a set of actions (plans) that an agent voluntarily decides to enact to achieve the desired change in the state of the world.

The BDI model proposed by Bratman includes several components, among them: the Means-End Reasoner, the Opportunity Analyzer, the Filtering Process, and the Deliberation Process. The Means-End Reasoner uses the agent's beliefs and desires to retrieve an existing plan from its repository or to conceive a new one if needed. These plans constitute the options the agent has to fulfil its desires. These options will be provided to the Filtering Process, which rejects all options that could clash with the intentions (and related plans) that are in execution but also permits a revision of the current decisions to meet the changes perceived in the environment. The Opportunity Analyzer elaborates on the agent's desires and is ready to catch opportunities arising from the state of the world to improve the current intentions or consider new ones. Finally, the Deliberation Process considers all the filtered and surviving options and deliberates one or more useful options for promotion to intentions. The selected intentions will then be enacted by executing the actions specified in their options.

2.2 The Global Workspace Theory of Baars

Baars's Global Workspace Theory [3, 4] plays a key role in our architecture. It is a shared global memory where all the knowledge is stored. The Global Workspace (GW) forwards incoming information to the psychological functions that may be interested in it. In this sense, it allows one to focus attention on specific information. Using a metaphor, Baars refers to this as a spotlight that illuminates a portion of memory, leaving the rest dark and obscured. Only this bright portion of the memory is conscious; the rest isn't. The knowledge contained in the conscious (enlightened) part of the GW is sent to the other mind's modules; the rest of the memory isn't conscious. The GW has a limited memory space, so any stored information is subject to atemporal decay, which allows the less or unused information to be replaced or eliminated. Conversely, the most used or recalled information is reinforced and moved to longer-term memory.

2.3 Buehler's Executive System

The third key element that we consider in our architecture is the is based on Buehler's discussion of the executive system and its role in agentive guidance [8],[7]. According to this view, an agent's psychology includes an executive system that manages several subsystems or subfunctions assigned to different tasks. These subsystems are four: the *Executive Inhibition Function System*, the *Executive Switching Function System*, the *Executive Resource Allocation Function System* and the *Work Memory Managment Function System*.
The Executive Switching Function initializes the phase of intention development and allows to alternate attention between conscious (*endogenous*) stimuli and conscious (*exogenous*) stimuli. The Executive Inhibition Function System inhibits everything that can interfere with the agent's goal, playing, together with the Executive Switching Function System, a key role in attention control. The Executive Resource Allocation Function System performs every action useful for the satisfaction of the agent's intentions. Finally, the Executive Working Memory Maintenance Function System takes care of the transfer and maintenance of information in memory between the Long Term Memory and the Global Workspace.

3 The Proposed Architecture

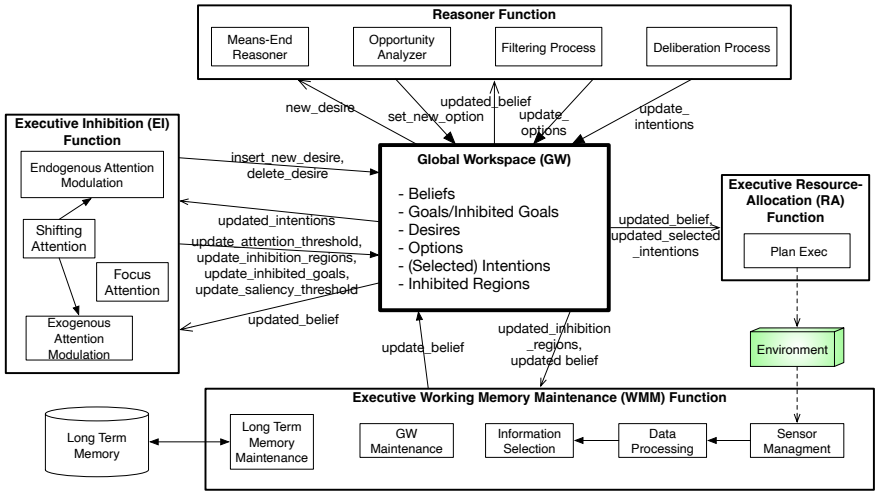

Figure 1: The Proposed Cognitive Architecture

The proposed architecture (see Fig. 1) supports the attention focus mechanism [8] in both the endogenous and exogenous attention modulation aspects. It is worth noting that in Fig. 1, relationships terminating with an open arrow represent an event, while relationships with filled arrows represent messages

transporting some kind of information (a new desire, an updated belief,...). In the following, for clarity, we describe the specific meaning of concepts like *goal*, *desire*, and *intention*, as well as the use we make of them. For these concepts, we were inspired by Bratman's concepts [6], and one of our previous papers [Citation removed for double-blind review].

A **(functional) goal** is an intended state of the world; for instance, g_0:='*Be in TownY*'. Usually, higher-level goals are decomposed into smaller ones, thus constituting a goal-tree whose satisfaction is an agent's desire. When the entire goal-tree cannot be satisfied, the agent may accept trade-offs that may result in the partial satisfaction of some goals. A goal also has a pre-condition defining the context where the goal is to be pursued and a saliency expressing the relevance of the goal. It is worth noting that a goal becomes part of the agent's will to pursue only when it is selected as a desire (see what follows).

A **desire** is a goal that has passed the agent's saliency and attention thresholds filter (more on that in subsect. 3.6).

The Reasoner Function evaluates the desires and generates one or more alternative plans for each of them, thus creating the options that could be used to fulfil the desire. The plan adopted as an option may also come from the previous agent's experience. In fact, our architecture supports the plan repository we proposed in [Citation removed for double-blind review].

Finally, when the Deliberation Process selects the best option, the pair $< Desire, Option >$ becomes a new intention that the Executive Resource-Allocation Function will execute. In other words, an **intention** is a desire the agent will pursue by enacting some optimal option that the agent puts into action.

There are other types of goals we explicitly consider in our architecture, more specifically: *epistemic* goals, *quality* goals, and the already cited *green* goals.

Epistemic goals are related to the need of the agent to update its knowledge. For instance, g_1:='*There is ice on the road*'. They may be triggered by a stimulus coming from some perception that motivates the agent to explore the part of the environment that generated the perception. Epistemic goals are fulfilled by options (i.e. plans) exactly as functional goals are.

Quality goals represent the conventional concept of a quality property conditioning the life of the agent; more specifically, a quality goal applies to a functional or epistemic goal and has a primary role in the selection of the best option for achieving that, when more than one is available. For instance, a constrained version of goal g_0 is: g_0':= *'Be in TownY by 8pm'* where the functional goal g_0 is constrained by the quality goal g_2:='*By 8pm*'. Quality goals participate in the trade-offs the agent will conceive to maximise the satisfaction of its goal-tree. In fact, an agent may relax the quantitative constraint imposed by the quality goal, thus partially achieving its objective.

Finally, **green goals** represent the constraints the agent complies with in terms of green policies during the pursuit of its goals. For instance g_3 := '*Minimize* CO_2 *emissions while fulfilling* g_0'. These are a special kind of quality goal since they have some kind of normative legitimation that forbids the acceptance of any trade-off about their accomplishment.

According with our approach, the agent's knowledge is in the Global Workspace (as prescribed by Baars), in form of Belief. The agent's mind behaviour is implemented by some of the Executive Functions discussed by Buhelers[8] and Diamond [11], namely: the Inhibition Function, the Resource-Allocation Function, and the Working Memory Maintenance Function. We also introduced one higher-level function (the Reasoner) to support some of the reasoning functionalities proposed by Bratman. Each Function is decomposed into several modules each one implementing specific portions of its behaviour.

In the next subsections, we will discuss these executive functions and their (sub-)modules.

3.1 The Global Workspace

The Global Workspace (GW) is a shared memory accessible to all executive functions within the proposed architecture. As described by Baar's GW theory, it is pivotal for implementing consciousness.

We will now propose a description of a working loop starting from perception and arriving to action execution that shows the role of the GW.

The Executive Working Memory Management (WMM) Function is responsible for transferring into the GW all pertinent information perceived by sensors. Raw data from sensors are processed, and the extracted information is stored within the GW as a belief update and made available to any function. Specifically, the GW acts as a publish-subscribe dashboard; it processes incoming messages (such as belief updates) and generates outgoing events (see Fig. 1) that notify the modules registered for that specific piece of information. Each function can then access the GW and retrieve the updated belief when needed.

Let us suppose a new belief represents the perception of some dangerous situation (the Data Processing and Information Selection modules cooperate to define the saliency of each perception). The inhibition Function accesses the new belief and compares its saliency with the current attention threshold. If the belief's saliency overcomes the saliency threshold, then the Function evaluates whether it is appropriate to generate a new epistemic goal that monitors the situation. The new epistemic goal is successively evaluated for promotion to desire so that the cause of the belief may be investigated. The function also revises the current set of desires, for instance, deleting one of them if some condition prevents its pursuit (for instance, pre-condition no more valid, clashing with current environment condition or other desires).

The new desire is processed by the Reasoner Function, which looks for new options to pursue and stores these options in the GW. In the meanwhile, it reasons on the opportunity offered by the updated belief, and if the case, it revises the current options (options related to currently pursued intentions) and intentions (for instance, adding the new desire to the set of intentions selected for execution).

The GW notifies the Resource Allocation Function of the updated beliefs and intentions so that this Function may execute the new intentions while considering the updated state of the world. The actions generated by this Function will

alter the environment, generating new perceptions that the Working Memory Maintenance Function will process; thus, the loop restarts.

As we can see, the GW plays a central role in the execution loop supported by the proposed architecture. Consciousness features (like the attention and focusing mechanisms) will be introduced in the next subsections.

3.2 The Executive Inhibition Function

The Executive Inhibition Function has two main duties: it implements the attention modulation mechanism and generates inhibition regions that limit the environment areas where the agent focuses its perception.

Attention Modulation is a fundamental feature of the GW theory since it allows the agent to enlighten the part of its knowledge it is conscious of at a certain moment. As Baars states [3]:
Only the bright spot is conscious, while the rest of the theater is dark and unconscious.

In our architecture, we consider two different directions for attention propagation (and related modulation mechanisms): Endogenous Attention Modulation is top-down attention related to a focusing effort guided by the rational will of the agent to pursue its goals. At the same time, Exogenous Attention Modulation is bottom-up attention driven by perceptions that could overcome the inhibition barriers raised by the current attention threshold and require the agent to focus on a new stimulus that has a high saliency [17].

The Endogenous Attention Modulation process generates new agent desires. The agent has some goals and would like to pursue all of them, but in many cases, this is not practically possible, and therefore, only a few of them will be selected as intentions.

The Exogenous Attention Modulation process involves how attention is captured by external stimuli, like brightness or movement, which are processed through belief updates. This bottom-up attention is quite automatic and involuntary.

When something significant happens, the agent directs its attention to processing the new stimulus, for instance, creating a new epistemic goal (and desire) to investigate the specific area of the environment that generated the perception. Additionally, inhibitions can be applied to reduce the sensitivity to these external events, acting as a filter that allows only important stimuli to pass.

The articulation of the cognitive areas of the agent's mind in executive functions allows us to split the competence of different functions into the different stages of evolution of a goal (something the analyst wants the agent to pursue), to a desire (something the agent wants to pursue) and, finally, to an intention (something the agent will actively pursue).

When the Executive Inhibition Function receives an update of a belief from the GW, it evaluates if the belief's saliency is greater than the current saliency/attention threshold (more details about that in subject. 3.6). If it is, the agent considers whether the new perception requires revising the current set of desires (bottom-up attention modulation).

Belief updates may also represent the achievement of some intention. This triggers the top-down attention modulation mechanism that revises the attention threshold and, if necessary, promotes some goal to new desire.

New desires will be posted to the GW, which notifies the Reasoner Function for the generation of options and the deliberation about intentions.

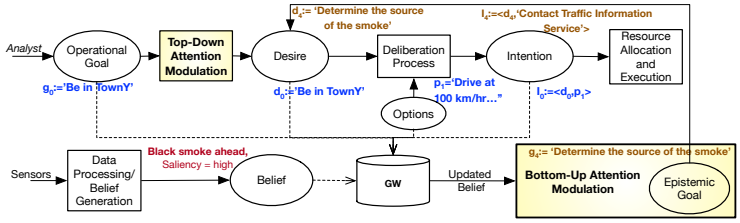

Figure 2: Consciousness and the Attention mechanism in CARA.

3.3 The Reasoner Function

The Reasoner Function is a higher-level function [11] inspired, in the CARA architecture, by the Bratman's reasoner [6]. This function includes four modules: The first, the **Means-End Reasoner**, processes the new desires and searches the Plan Library for plans that can satisfy them. If no suitable plan exists, it invokes the Planner sub-module to create a new plan and stores it in the Global Workspace (GW) as a new option for pursuing that desire. If one or more plans are found in the library, the Reasoner evaluates their quality and publishes in the GW those that satisfy the quality desires of the agent.

The second module is the **Opportunity Analyser**. As suggested by Bratman [6], this considers the current state of the world and the agent's state looking for better opportunities to achieve the current desires. For instance, it may find out, that in a specific situation, a good option arises for pursuing some desire that is not currently in the selected intentions. Suppose the agent is driving an autonomous vehicle from A to B but it also desires to pass from C (that is not in the best route connecting A to B). If some road blockage forces the vehicle to take another route, this may offer the opportunity to visit C without a significant effort.

The **Filtering Process** is the third module. This removes the previously generated options that new environment conditions make incompatible with the current intentions or the agent's green goals.

Finally, the **Deliberation Process** decides which is the best option for satisfying each desire and if it may be adopted in the current agent's state. Therefore this module decides which agent's desires are actively pursued through intentions. The Deliberation Process also implements the trade-off capability, which consists of the algorithmic evaluation of degraded levels of qualities for the functional goals or even the relaxation of some parts of the goal formula (in terms of predicates or temporal constraints). It is worth noting that the

current version of the CARA implementation does not yet support the trade-off capability, which is still under development.

3.4 The Executive Resource-Allocation (RA) Function

The Executive Resource Allocation (RA) Function executes the Intention's plan. This involves orchestrating the various actions that make up the plan and ensuring proper management, distribution, and equilibrium of resources. The current implementation performs the simple execution of the list of actions composing the plan, i.e. it invokes the corresponding agent's behaviours. In future work, we plan to adopt some kind of workflow engine so that complex plans involving a parallel flow of actions may be supported.

This Function cooperates with the WMM Function in the estimation of the quality of the result obtained after the execution of some plan. This is a complex issue that, by now, is implemented referring to the observation of simple environmental parameters that are used to estimate the quality metrics. For instance, in the example of an autonomous vehicle, the quality metrics of the plan used to go from town A to town B may be the travel time and the consumption of fuel. This latter may be constrained by some green goal thus its observation ensures the abidance of the goal.

3.5 The Executive Working Memory Maintenance (WMM) Function

The Executive Working Memory Maintenance (WMM) Function is responsible for managing, maintaining, and updating data within both the Long-Term Memory and the Global Workspace. It deals with sensor management addressing perceptions according to the prescriptions of the inhibited regions generated by the EI Function. This means that if a camera is looking at the road in front of an autonomous vehicle, this function removes from the processing area all the parts of the image that are inhibited (for instance, the sky that is not significant for driving the vehicle). Indeed, we have already discussed that bottom-up attention modulation allows for properly processing salient stimuli coming from the GW.

3.6 An Example of Attention, Focusing and Consciousness

In this subsection, we will provide a quick example of how the proposed architecture realizes the attention modulation mechanism and focusing behaviour. The example refers to Fig. 2.

Let us suppose the cognitive agent, an autonomous vehicle (AV) starts in TownX with goals g_0:=‘ *Be in TownY*’, g_1 := ‘*Fulfill* g_0 *by 6 pm*’, g_2 := ‘*Minimize* CO_2 *emissions while fulfilling* g_0’ and g_3 := ‘*Fulfill* g_0 *safely*’. The plan p_0 for fulfilling g_0 is provided by Google Maps. This is also good for g_1 because

the chosen route is the fastest way to get to TownY under current traffic conditions. As regards g_3 this is fulfilled by the plan p_1 := '*Drive at 100 km/hr on the highway, if you have trucks near you drive at 80 km/hr*'. We can note that 100 km/hr is also an optimal speed for minimizing carbon emissions, and therefore, it fulfils goal g_2. So, the AV proceeds with plans p_0 and p_1 that fulfil all its goals. As it is driving along the highway, the AV sees a tall stack of black smoke, an exceptional sensation. So, it adds an epistemic goal g_4 := '*Determine the source of the smoke*' with the highest saliency, so all other goals are put aside. As it comes closer to the smoke stack, it sees that the smoke resulted from a collision of two trucks, and the road ahead is closed. This belief is added to its workspace and becomes available to the active plans being pursued, i.e., the AV is conscious of the collision. As a result, the AV concludes that p_0 is no longer viable. However, there is an exit before the two disabled trucks. So, the AV gets another plan from Google Maps for fulfilling g_0, say p_2, which suggests a minimal path to TownY through country roads. Unfortunately, travelling through country roads means that the AV will be travelling at a speed that is less than 50 km/hr and will not reach TownY by 6pm. Moreover, it will not meet its green goal because it will be driving longer at sub-optimal speed for carbon emissions. Now the AV considers two plans: p_{0+}:= '*Wait till the autostrada reopens, then continue with* p_0 *and* p_1'. But how long will it have to wait? The AV adds another epistemic goal: to get a good estimate from the traffic information service. It finds that it will take about an hour because the police have to arrive, and reopen all the highway lanes. This means that plan p_{0+} will result in a 7 pm arrival in TownY, whereas plan p_2 will result in a 7:30 arrival and more carbon emissions because of lesser travelling speed and longer drive. So the AV revises g_0 to g_{0+}:='*Be in TownY by 7:00pm*' and keeps p_{0+} and p_1 as its active plans for the rest of the trip.

4 The CARA Implementation

The described architecture is currently implemented in Java; we have not implemented any Message Transportation Service (and related features) for that since we plan to embed our work as the reasoning part of a SARL agent [14].

We implemented the two main features related to consciousness behaviour supported by Baars' and Bueheler's theories: the spotlight movement (or focusing) and the attention modulation mechanism. Moreover, we aim to exemplify how an agent, built upon this architecture, reacts to a significant stimulus that can perturb its plans.

Our experimental setup implements a game inspired by the 'Ticket to Ride' table game with a few ad hoc variations of the rules. The playfield represents a railway map connecting several European towns. Each player receives one or more goals. Each goal requires the player to connect two railway stations, following a path composed of a sequence of routes, each route connecting two stations; each route has a different number of steps of the same colour. We suppose that each colour defines some route features: *panorama*, *transit speed*,

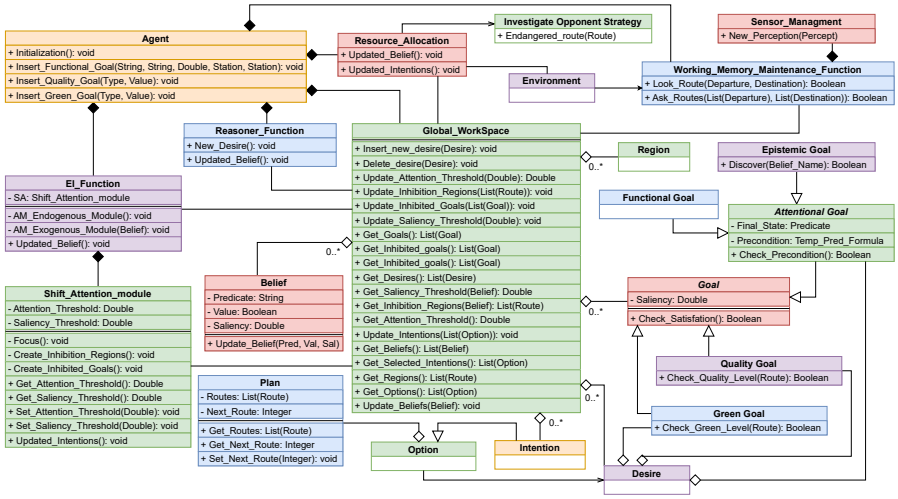

Figure 3: The Class Diagram Implementing the Proposed Architecture for the Ticket to Ride Case Study

and the specific *amount of pollutant* released to pass through that route step. A route can already be acquired by another player or randomly interrupted by unknown rail damage. We introduced this last feature to test the agent's reaction to new events. At each turn, each player receives two Train Cards, each one in a different colour (corresponding to the route features). When the player has enough train cards of the same colour, she may buy a route on the map with steps of that same colour. At the beginning of the game, the agent is injected with a set of *functional goals* (like '*Connect Lisboa to Berlin*') so that it can deliberate on its initial desires.

Moreover, each functional goal is constrained by a quality goal, such as how panoramic the path should be or the expected average speed of the path. Also, a green goal may be specified, constraining the average pollutant emission allowed for each step of the path. An agent can decide to follow one path rather than another one because of its green and quality goals.

Each functional goal also has a priority that the agent perceives as a saliency or urgency to achieve. By now, we suppose the agent will pursue only one functional goal at a time, usually selecting the one with the highest saliency. Future implementations will support multi-objective planning.

In the following subsections, we will detail the functioning of the implemented architecture referring to the CARA's class diagram reported in Fig.3. The Figure illustrates the class diagram implementing the proposed architecture for the discussed Ticket to Ride case study. To limit the diagram's dimension and improve its readability, we omitted the less significant information (attributes and methods).

4.1 Triggering a New Epistemic Goal

The Global_Workspace class (Fig. 3) implements the central dashboard as proposed in the Baars' theory. It exposes the methods that can be invoked by the classes implementing the executive functions.

The WMM Function, after receiving and processing some perception, invokes the *Update_Beliefs(Belief)* method; the Global_Workspace class reacts by notifying the event to the other executive function classes, for instance invoking the *Updated_Belief()* method of the EI_Function class. This class reacts by requiring an update of the existing beliefs to the GW (method *Get_Beliefs ()*). The new beliefs (in general, more than one may have been updated since the last get method invocation) are processed considering their saliency. The *AM_Exogenous_Module (Belief)* method is invoked to compare the current attention threshold with that assigned to the new perception by the WMM Function. If the saliency is greater than the attention threshold, this belief triggers an epistemic goal.

For instance, in the case study, any time a user moves into an area near a route of the agent's intended path, or a user acquires a critical route for the agent's intended path, an epistemic goal is raised and soon promoted to desire, so the *AM_Exogenous_Module* method invokes the *Insert_new_desire(Desire)* method of the Global_Workspace class. The objective of this epistemic goal is to understand whether the opposing player may, in the future, be interested in acquiring routes that are part of the agent's intended path.

When the goal is promoted to desire, the Reasoner Function processes the desire and finds a proper option (we prepared a specific plan which includes the *Endangered_route(Route)* method of the Investigate_Opponent_Strategy class). This method evaluates if the opponent is taking a dangerous strategy.

The Reasoner class inserts the new intention in the GW, which informs the Resource_Allocation (RA) class (by invoking its *Updated_Intentions()* method). The RA function executes the plan by invoking the previously cited method.

Understanding another player's strategy from her moves is a very challenging task, but that is not in the scope of our current setup. We currently implemented a very simple strategy that notifies that the opposing player threatens the agent's route. This is done by sending a new perception to Sensor_Management. The WMM Function processes this and updates the belief regarding that route in the GW. The new belief is, in turn, processed by the Reasoner Function that may deliberate to change the plan (for instance, by selecting another option to replace the one that contains the endangered route).

4.2 The Attention Modulation Mechanism

The Top-Down (or Endogenous) attention modulation mainly deals with the promotion of goals to desires. For a better introduction to the process, we should refer to the two different concepts of saliency threshold and attention threshold. The two threshold values coincide when the agent is not focused on any specific activity. Vice versa, when the agent is already focused on pursuing

some goal, the saliency threshold comes from the saliency value of the current desires, while the attention threshold is modulated as a higher value that should filter perceptions that are not relevant to the current flow of action.

A relevant mechanism involves the inhibition of goals; a goal is inhibited when its precondition is not verified or when its saliency is lower than the agent's saliency or attention thresholds.

The Endogenous module continuously revises the agent's desires and goals. Regarding desires, the ones whose goal is in the set of not inhibited goals are compared with the current saliency threshold. If the desire's goal is in the list of inhibited goals, its saliency is compared with the attention threshold. In both cases, if the goal's saliency falls below the considered threshold, the desire is deleted from the GW.

Goals are processed in two different ways according to their belonging to the inhibited goal set or not. The first step in processing goals that are not inhibited is checking their precondition. If verified, the goal's saliency is compared to the saliency threshold, and if it exceeds that, the goal is promoted to desire and inserted in the GW. The process for inhibited goals is quite similar, but the goal's saliency is compared with the attention threshold, meaning that this goal should have a very high saliency to be promoted to desire. In this approach, we suppose that goals' saliency may change at runtime because of external interventions (by the analyst) or because of environmental changes over time.

4.3 The Global Workspace Spotlight

The GW spotlight illuminates the portion of the agent's knowledge that is concerned with the prosecution of the current desires and intentions. The behaviour of this mechanism matches what is reported in Fig. 1. Initially, the *Focus()* method of the Shift_Attention_ Module computes the saliency and attention threshold according with the selected intentions already stored in the GW and updates these thresholds to the GW. Hence, if the attention threshold is higher than the saliency threshold (this means the agent is focused on achieving some intention), this method creates the inhibition regions according to the selected intentions. These regions have a twofold purpose: they are used to inhibit all the goals that are not related to the current intentions, and they mask the non-relevant beliefs in the GW. This way, the agent is 'conscious' of the relevant part of its knowledge, but new relevant stimuli may still be perceived if their saliency overpasses the attention threshold.

5 Conclusions and Future Works

We have illustrated an artificial intelligent agent architecture that we named CARA (Conscious Agent Reasoning Architecture). The architecture supports *goal-oriented reasoning*, accepting *trade-offs* in pursuing multiple goals, providing the agent with *consciousness* features, which enable advanced reasoning. Furthermore, a *normative dimension* could condition the agent's decisions and

trade-offs with a specific regard for green policies. The work takes inspiration from the Global Workspace Theory [3], thereby enabling an integrated implementation of key Executive Functions (specifically, Inhibitory Control, Resource Allocation, and Working Memory Maintenance) of the Executive System Theory [11]. The architecture has been described in each component, it has been fully implemented and the UML diagram of the classes, which implements the proposed architecture for a specific case of study, has been provided.

Future work will consider using the proposed architecture in specific domains to determine its effectiveness and capabilities compared to other traditional approaches and experimentally validate the approach. Furthermore, we plan to enlarge the scope of this architecture to become a multi-tool that can support more than one approach to consciousness. In this direction, we plan to implement aspects of Rosenthal's higher-order theory of consciousness [15], [16], according to which an agent is in a conscious state in virtue of being aware of itself as being in that state by way of a higher-order representation, where this is typically accompanied by the ability to report the state that it is in. We have already done a preliminary study in that direction in [citation removed for double blind-review].

6 Acknowledgements

We acknowledge the support of the PNRR project FAIR - Future AI Research (PE00000013), Spoke 9 - Green-aware AI, under the NRRP MUR program funded by the NextGenerationEU.

References

[1] Raúl Arrabales, Agapito Ledezma, and Araceli Sanchis. Criteria for consciousness in artificial intelligent agents. In *ALAMAS&ALAg Workshop at AAMAS 2008*, pages 57–64, 2008.

[2] Bernard J Baars. *In the theater of consciousness: The workspace of the mind*. Oxford University Press, USA, 1997.

[3] Bernard J Baars. Global workspace theory of consciousness: toward a cognitive neuroscience of human experience. *Progress in brain research*, 150:45–53, 2005.

[4] Bernard J Baars. The global workspace theory of consciousness: Predictions and results. *The Blackwell companion to consciousness*, pages 227–242, 2017.

[5] Michael Bratman. *Intention, Plans, and Practical Reason*. Harvard University Press, 1987.

[6] Michael E Bratman, David J Israel, and Martha E Pollack. Plans and resource-bounded practical reasoning. *Computational intelligence*, 4(3):349–355, 1988.

[7] Denis Buehler. *Psychological Agency-Guidance of Visual Attention*. PhD thesis, UCLA, 2014.

[8] Denis Buehler. Agential capacities: a capacity to guide. *Philosophical Studies*, 179(1):21–47, 2022.

[9] Antonio Chella. Robots and machine consciousness. 2022.

[10] Antonio Chella, Riccardo Manzotti, et al. Artificial intelligence and consciousness. In *Association for the advancement of Artificial Intelligence Fall Symposium*, pages 1–8, 2007.

[11] Adele Diamond. Executive functions. *Annual review of psychology*, 64(1):135–168, 2013.

[12] Michael Georgeff, Barney Pell, Martha Pollack, Milind Tambe, and Michael Wooldridge. The belief-desire-intention model of agency. In *Intelligent Agents V: Agents Theories, Architectures, and Languages: 5th International Workshop, ATAL'98 Paris, France, July 4–7, 1998 Proceedings 5*, pages 1–10. Springer, 1999.

[13] Marek Pokropski. *Mechanisms and consciousness: Integrating phenomenology with cognitive science*. Routledge, 2021.

[14] Sebastian Rodriguez, Nicolas Gaud, and Stéphane Galland. Sarl: a general-purpose agent-oriented programming language. In *2014 IEEE/WIC/ACM International Joint Conferences on Web Intelligence (WI) and Intelligent Agent Technologies (IAT)*, volume 3, pages 103–110. IEEE, 2014.

[15] David Rosenthal. Higher-order awareness, misrepresentation and function. *Philosophical Transactions of the Royal Society B: Biological Sciences*, 367(1594):1424–1438, 2012.

[16] David Rosenthal. Consciousness and confidence. *Neuropsychologia*, 128:255–265, 2019.

[17] Wayne Wu. We know what attention is! *Trends in Cognitive Sciences*, 28(4):304–318, 2024.