Neural Computing and Applications (2025) 37:23739–23771
https://doi.org/10.1007/s00521-024-10584-6

S.I.: HYBRID MACHINE LEARNING TECHNIQUES IN MULTIDISCIPLINARY DIGITAL ORGANISATIONS (HMLT-MDO)

Development of a model for the study and measurement of consciousness in artificial cognitive systems based on the integrated information theory

Luz Enith Guerrero[1] · Jeferson Arango-López[1] · Luis Fernando Castillo[1] · Fernando Moreira[2]

Received: 21 October 2023 / Accepted: 1 October 2024 / Published online: 10 December 2024

Abstract

The study and measurement of consciousness in artificial cognitive systems have been subjects of great interest in the fields of artificial intelligence and neuroscience. This article introduces an innovative model based on the integrated information theory (IIT) aimed at addressing this fundamental issue. The IIT, proposed by Giulio Tononi, offers a robust theoretical approach to comprehend consciousness in terms of information and its integration. Our model focuses on identifying and quantifying key aspects of consciousness in artificial systems, including information complexity and the structure of artificial neural connections. We have developed a conceptual framework to assess the “integrity” of information within a system, enabling us to measure consciousness in terms of its capacity to integrate information coherently, setting it apart from other approaches. Through experiments and simulations conducted with a video game using the Go No Go task, we demonstrate how our model can be applied to artificial cognitive systems, allowing for the evaluation of their level of consciousness in different contexts. This approach carries significant implications for the advancement of more advanced and ethical artificial intelligence systems, as it provides a methodology for evaluating and comparing their level of consciousness. Ultimately, our work contributes to the progress in understanding and measuring consciousness in artificial systems, paving the way for future developments in artificial intelligence and computational neuroscience.

Keywords Automatic gait analysis · Motion capture · Computer vision · Artificial intelligence · Health information technology · Clinical decision making · eHealth

1 Introduction

This research approaches consciousness from the integrated information theory (IIT) proposed by Tononi [1]. The work carried out and aims to study and measure consciousness (understood as integrated information) using artificial systems with cognitive functions associated with consciousness using artificial intelligence techniques and computational models. This cognitive system will be reviewed using the PyPhi toolbox [2].

The search for how the phenomenon of consciousness is produced by neurobiological processes in the brain has had in recent years a very active work, although with little progress. Churchland [3], Crick [4], Penrose [5], Edelman [6], and Dennett [7], among others have recently dealt with the subject without arriving at a widely accepted theory.

✉ Fernando Moreira
fmoreira@upt.pt

Luz Enith Guerrero
luzenith_g@ucaldas.edu.co

Jeferson Arango-López
jeferson.arango@ucaldas.edu.co

Luis Fernando Castillo
luis.castillo@ucaldas.edu.co

[1] Departamento de Sistemas e Informática, Universidad de Caldas, Manizales, Colombia

[2] REMIT, IJP, Universidade Portucalense and IEETA, Universidade de Aveiro, Aveiro, Portugal

The focus of studies on consciousness in recent years has been the search for correlations between some areas of the brain and some states of consciousness, known as neural correlates of consciousness (NCC), this has allowed understanding how areas of the brain related to conscious experience such as feeling pain and work, although it is not known why. The explanations that have been given, are based on the brain, unfortunately there are not many studies that report research on subjective experience.

Currently, neuroscience has had great progress in understanding how brain mechanisms perform cognitive functions as diverse as decision making, language analysis, motor control, attention etc. However, the large number of unanswered questions about the biology of the nervous system makes it a great field for research. It is not easy to understand how an organism can function as a whole and how its most complex functions can occur.

IIT considers that the unification of consciousness should be achieved by the brain's ability to integrate information [8]. That is, conscious experiences are integrated representations of large amounts of information and is represented as Φ [9]. The degree of integrated information is related to the level of consciousness. This hypothesis has been indirectly based on experiments showing the breakdown of effective connections in the brain during loss of consciousness. Only a few studies have been able to quantify the integrated information in real neural data because of the computational costs involved. Methods to improve computational efficiency work for some measures of Φ but in not very complex systems; therefore, their use for the case of the human brain is not possible [10]. Consequently, even though there is a clear mathematical definition for Φ, in practice, it cannot be computed even in small organism systems. So, it would be very interesting to be able to quantify consciousness and thus, for example, a doctor could know the exact state in which a patient is: if he is sleeping, if he has any brain damage or is anesthetized. In this way, it would help the medical professional to make decisions regarding the attention, treatment, care, etc., of sick patients who are unable to communicate in any way.

According to Barrett [11], IIT has already inspired heuristic measures of information integration/complexity successfully applied to recorded electrophysiological data that allow distinguishing the waking state from various unconscious states, i.e., sleep and anesthesia under various anesthetics [12] and [13].

Other applications of these measures are presented for neuroimaging studies because they offer a way to quantify the complexity of brain functions that would allow identifying information-based differences between healthy subjects and patients with neurodegenerative diseases. This was to analyze them comparatively to provide a quantification of the degree of the disorder and even an analytical form of a diagnosis [14].

Cognitive science has also recently joined the existing areas that from different angles are studying the problem of consciousness. This science has as one of its goals to elaborate computational models that when implemented allow machines to be endowed with something like consciousness, which has turned out to be the "most intricate of mental phenomena" [15]. In this sense, the field of artificial consciousness has begun to be taken seriously as a way of scientifically studying consciousness. It is important to emphasize that the IIT criteria are consistent with the existence of artificial consciousness [16].

Now, through the consciousness, one can have summarized access to a great number of contents of the mind. Neurological studies have always shown that there are numerous unconscious neural processes running in parallel, but when looking at a particular moment only some of them are shown to the consciousness, this can be very interesting when applied in artificial systems [17].

This research sought to characterize aspects of consciousness according to IIT, based on cognitive functions such as attention, executive functions, and working memory, which generally have a direct association with consciousness, which generally have a direct association with consciousness, with the aim of extrapolating them to computational models that could be useful in the study of both artificial consciousness and in the understanding of human consciousness, in addition to helping the development of more efficient machines, and the refinement of IIT, since reducing the large amount of calculations required according to the guidelines established by the theory, would be very useful in improving computational efficiency and thus achieving the amount of consciousness or integrated information for the purposes of IIT.

In particular, an artificial cognitive architecture [18] was adapted to test cognitive functions commonly associated with consciousness such as attention, working memory, and executive functions [19]. These allow to study the interactions that may occur between emotions, cognitive processes, etc., and to determine how much information a system can integrate, and in the future, to obtain adaptive behaviors in complex environments.

Finally, three starting elements that motivated the research were defined: (1) according to IIT artificial consciousness is possible which means that it is possible to study consciousness in artificial systems; (2) it is evident the importance of computing Φ which, although it is a key challenge in IIT for the understanding of consciousness, can be used in many other areas such as general network analysis, physics and engineering, where notions of "dynamic complexity" are of more general importance; (3) given the importance of the scientific study of consciousness, in particular the field of artificial consciousness can present a means of experimentation that in turn serves to test the theory approaches and can then feed back into the study of human consciousness.

2 Related concepts

One of the questions that has most troubled humanity is the reason why we are conscious. In this sense, a debate has arisen that tries to explain the relationship between mind and matter, or how is it that mental or subjective states (sensations, beliefs, decisions, and memories) are related to the substances and processes of the world of objects that science studies? For a better understanding, the key concepts of the research are presented below.

2.1 Consciousness

The problem of conceptualization between mind–body and consciousness has a strong relationship. The word mind commonly refers to aspects of intellect and consciousness that manifest in some combination of thought, perception, memory, emotion, imagination, and will. Usually, the concepts mind and thought are often synonymous [20].

Phenomenal consciousness is the property that mental states, events, and processes possess when there is something that is like something for the subject who possesses them; to experience them or to be in such states. Conscious experience refers to the phenomenal character of experiences [21].

Questions about what certain areas of the brain do and what they correlate with have been called by Chalmers [22] the "easy problems[1]"of consciousness, but they do not address what he posed at the beginning, known as the "hard problem" and which can be translated into the following question: Why does every physical process in the brain have to be accompanied by consciousness, or why is it that every behavior is accompanied by a subjective experience?

Trying to answer the questions, Chalmers proposes that consciousness is a fundamental aspect, like space, time, and mass, and furthermore, that consciousness can be universal. Every system can have a degree of consciousness. Another interesting view of consciousness is given by Tegmark [23] when he hypothesizes that consciousness can be understood as a state of matter. Consciousness is a physical phenomenon that feels non-physical; consciousness would be the way information feels when it is being processed in certain complex ways. For Fekete and Edelman [24], a mathematical framework could treat experience at its base that would show how qualia[2] [25], intrinsically arise from the computational properties of their physical substrate.

For Koch et al. [26], characterizing consciousness is a deep scientific problem, which may have implications in addressing problems such as disorders in consciousness [13], state of consciousness in utero [27], and in sleep and more recently in machines [16]. Phenomenal consciousness is currently the focus of neuroscientific research, and one of its main goals is to find NCCs because (1) it defines the research target of experiments and (2) the notion of NCC determines which cases can be considered as test cases for general scientific theories of consciousness. Therefore, any general scientific theory of consciousness should involve NCCs to be testable [28].

What should be the characteristics of NCCs, if one were to think of the possibility that different species of animals could have consciousness, then perhaps the neuroanatomical characteristics should not be so relevant, and perhaps it would be more convenient to privilege other characteristics, such as neurobiological, chemical, and organizational. If the idea is to facilitate the application to AI research, perhaps the balance would be tipped toward neurofunctional or computational features, as integrated information theory does.

2.2 Integrated information theory (IIT)

IIT is an evolving theoretical framework that seeks to explain what a system requires for consciousness to arise, describes the empirical evidence, and formulates a series of predictions that can be tested; it was proposed by Tononi [1] and has been revised several times [1, 8, 29, 30].

For Tononi and Koch [31], part of the difficulty in dealing with the difficult problem of consciousness has been the way science has approached it, starting from observing the brain to locate the origin of consciousness, and if, on the contrary, the opposite approach was adopted, things could be less complicated. Thus, IIT takes the existence of consciousness for granted, i.e., it starts from experience itself, identifying its essential properties (axioms) and how they can be associated with certain physical systems, such as neurons and their connections, for which it infers what type of properties these physical systems must have (postulates) to take into account the phenomenology of consciousness [8]. The theory is derived

---

[1] So named because they seek to explain how cognitive abilities occur and are in some ways easier for science to study.

[2] They are the subjective qualities of conscious experiences. For example, the redness of red, or the painfulness of pain. For example, if you imagine a blue circle; that conscious experience has at least two qualia, the qualia of color, responsible for its sensation of blue, and the qualia of form, responsible for the circular appearance of the object in the imagination.

from mental experiments leading to phenomenological axioms and ontological postulates and tries to explain what consciousness consists of and how it can be associated with certain physical systems, i.e., it is established for such systems whether they have consciousness, the degree of consciousness and the experience they are having [1].

IIT starts from two fundamental postulates about subjective experience: differentiation, the availability of a great number of conscious experiences (so we can be conscious of many things) and integration, the unity of each of such experiences, that is why when we are conscious of something, in the mind this is given in a complete and total form, that is, it is irreducible. This unity of consciousness is due to the great number of cause-effect interactions in the brain. The more interconnected or integrated a brain or system is, the more conscious it is. According to the theory, consciousness corresponds to the capacity of a system to integrate information, which concerns the repertoire of available states of the system. Thus, information can be seen as a propagation of cause and effect within the system that can be measured from the amount of uncertainty that it reduces [1].

Now, just as the amount of information stored in a file can be measured, consciousness could also be measured, by calculating the amount of integrated information [32]. That quantity called Φ is given in bits and quantifies the uncertainty reduction (i.e., the information that is generated when a system enters a particular state through causal interactions between its parts). “Phi is the amount of causally effective information that can be integrated through the weakest informational link of a subset of elements” [1]. This means that if a system is formed by isolated elements, then its parts will not be able to share information so its phi value will be low. Furthermore, according to observations made on the disconnection of some brain areas, it is suggested that consciousness is not discrete but gradual; therefore, for an entity (human or not) to be attributed consciousness, it is required to be unitary and integrated and to have a large repertoire of differentiated states [31].

According to IIT, if a system s generates some integrated information ($\Phi > 0$), then s is phenomenally conscious, and if on the contrary $\Phi = 0$, system s in unconscious. A problem facing the theory is its current impossibility to calculate with accuracy, the value of phi Φ, not only of the human brain, because of the multiple organization of its trillions of neuronal connections, but even that of the common roundworm, with a mere 302 nerve cells. Given the evolution that IIT has had, several measures for the integrated information have been proposed but there is no consensus about it; moreover, the methods to improve the computational efficiency in the calculation of these measures, work with certain restrictions and only for not very complex systems [10, 33–37]. Thus, developing an algorithm to calculate Φ, more easily, and continuing to achieve reliable results in laboratory tests, would not only constitute a greater support for the theory, but also a way to measure the consciousness of other beings or even machines, since another interesting aspect of IIT is that it does not limit the possibility of consciousness only in humans or mammals, opening the way to artificial consciousness [31]. In this work specifically, Φ3.0, as defined in Oizumi et al. [8], was used as an integrated information measure.

2.3 Artificial consciousness

The strengthening of the study of consciousness has motivated researchers to test theories of consciousness with computational models that some believe could lead to more intelligent machines and possible phenomenal states. This type of research has come to be known as “machine consciousness,” although the term “artificial consciousness” has also been used to describe it [38]. IIT only requires a system to reach a certain level of information integration, so that it can theoretically have a non-zero “phi” value (some degree of consciousness).

Now, Searle [39] characterized strong AI by saying that “the properly programmed computer really is a mind, in the sense that computers with the right programs can literally be said to understand and have cognitive states. In contrast, weak AI posits that machines have neither mind nor consciousness and only simulate thought and understanding.

There is a difficulty in explaining how subjectivity can arise from matter, so while we know about human consciousness from a first-person perspective, artificial consciousness can be accessed from a third-person perspective. In this sense, a definition of artificial consciousness given by us will be in a third-person perspective [40].

Objections to artificial consciousness according to the “hard problem” should be taken with care about the claims made, since if the hard problem is solved, it could be discovered that machines are capable of consciousness. Furthermore, it can be argued that asking questions about phenomenal states in machines and developing models of consciousness will likely improve our understanding of human consciousness and bring us closer to a solution to the “hard problem.” Furthermore, it might be possible to create conditions that allow consciousness to emerge in a system without understanding the causes of phenomenal states [38].

How can we create a machine that has consciousness functions?

According to Gamez [41], there are four classes of artificial consciousness, which are not necessarily exclusive:

- Machines with external behavior associated with consciousness.
- Models of the correlates of consciousness.
- Models of consciousness
- Phenomenally conscious machines

On the other hand, the usual software design techniques and traditional approaches to artificial intelligence have proven to be insufficient to deal with such complex environments involving perceptual processes. In this sense, research in cognitive sciences and neurosciences is fundamental for the design of artificial cognitive systems.

2.4 Cognitive systems

For Wendin [42], cognition is the (human) mental process of acquiring knowledge and understanding via thought, experience, and senses. The development of computational modeling led to the creation of cognitive science that attempted to explain internal mental processes using ideas such as goals, memory, task queues, and strategies. This is without necessarily attempting to ground these processes in neurophysiology. Cognitive science and AI have been closely related since their inception. Cognitive science has provided clues for AI researchers, and AI has aided cognitive science with newly invented concepts useful for understanding the workings of the mind [43].

Since qualia are inherently private, determining whether a machine experiences any inner life still lacks an explanation that can be brought into a computational model, then third-person observation-based approaches can be used to advance artificial cognitive systems to bring them into bio-inspired AI. These approaches consist of examining behavior including forms of accurate reporting or communication, but one can also inspect the inner part of a creature [17].

Cognitive theories give rise to artificial systems that try to emulate human cognitive capabilities. These systems are based on software architectures that implement cognitive models and are referred to as artificial cognitive architectures, many of which have been developed very successfully by implementing techniques that typically come from AI, such as artificial neural networks. In this work, the architecture developed by Rendón et al. [18] was used.

On the other hand, it is worth considering that the relationship between cognitive functions and consciousness is a complex topic and the subject of research in both neuroscience and philosophy of mind. While there is no definitive consensus on how they are related, several cognitive functions have been identified that are often associated with consciousness. Some of these are presented below:

- *Perception:* Conscious perception involves the ability to consciously perceive and interpret the world around us. For example, conscious perception allows us to consciously see colors, hear sounds, and feel textures.
- *Attention*: Conscious attention refers to the ability to deliberately focus and direct attention to certain stimuli or tasks. It allows us to consciously select what we perceive and process at a given moment.
- *Working memory:* Working memory is a system that allows us to temporarily retain and manipulate information in the mind. The conscious version of working memory allows us to be aware of the information we are actively holding in the mind.
- *Conscious thought:* This includes processes such as reasoning, planning, and conscious decision making. It allows us to consciously solve problems and make decisions.
- *Self-awareness*: It is the ability to be aware of oneself, including awareness of our own emotions, thoughts, desires, and self-identity. This can also include the ability to reflect on our actions and motivations.
- *Time awareness:* It allows us to be aware of the sequence of events and the duration of time intervals. This allows us, for example, to remember the past, anticipate the future, and be aware of the present.
- *Awareness of the social environment*: This involves the ability to be aware of other individuals, their mental states and emotions, and the ability to interact socially in a conscious manner.

It is important to clarify that not all aspects of these functions necessarily involve consciousness, and some aspects of consciousness may be independent of these functions. Furthermore, the study of consciousness is an evolving interdisciplinary field that seeks to understand how conscious experience emerges from brain activity and how it relates to cognitive functions and perception.

3 Methodology

It should be clarified that the research had four main objectives. The following is a summary of these objectives and the methodology followed to achieve each one of them:

- Adapt an artificial cognitive architecture inspired by a computational model of consciousness with integration of different cognitive functions that considers the elements required according to IIT guidelines.

- Design a computational model that specifies conscious perception according to IIT in an artificial system.
- Design a model to characterize the artificial cognitive system that integrates artificial intelligence techniques to find some of the measures of the integrated information (Φ) in an efficient way.
- Determine characteristics of the system that allow incorporating the cognitive architecture with elements of artificial consciousness according to IIT.

3.1 Artificial cognitive architecture

In this research, the conscious perception is based on the access awareness according to [44] such as information integration or attentional focus. In addition, to reach the objectives an application environment based on artificial cognitive systems was defined, which as a serious video game interacts with young university students, in order to review certain cognitive processes commonly associated with consciousness (attention, memory, and executive functions) by performing specific tests ensuring their correct application, according to the protocols for the computerized version of the test(s) referred to. Within this environment, an artificial agent in charge of launching the tests is used, and on this agent the exploration of artificial consciousness in the environment defined for the domain of the system is addressed, since the high degree of complexity that the problem could have, had to be delimited for its implementation.

However, considering that most complex learning abilities and higher cognitive skills are mainly observed in species with higher levels of consciousness, such as humans [45]. Subsequently, taking Manzotti [19] as a reference, some of the cognitive functions associated with consciousness that would serve for the development of this model (taking as a reference what is established by some of the cognitive theories [46]) were identified. Then, a cognitive architecture was sought that, based on the selected cognitive functions, could be better adapted to the characteristics of IIT. It should be noted that with IIT, a "scale of consciousness" can be developed that indicates the level of consciousness that a system has [47]. To assess the level of awareness of a system, its complexity and how the forms of information are integrated in the system are taken as a basis [48].

Then, the cognitive tests to be applied to university students (young people between 20 and 30 years of age) without defined characteristics were to be selected. Once the available cognitive tests were reviewed, those that could be taken to a computational version, and adapted to the artificial cognitive system, were taken.

In this work, particular reference was made to the Go-Nogo test [49] with the following characteristics:

- *Participants:* By means of convenience sampling, the test was administered to students over 18 years of age. The sample consisted of a total of 20 subjects, of whom 17 were male and three were female. All subjects were presented with the information concerning the study and their rights were stated in the informed consent form. Informed consent stated their rights as participants in the research.
- *Ethical aspects:* The methodology and the handling of the information in this research were followed according to the recommendations for research found in the framework of the national and international legislation on health research (Resolution No. 008430 of 1993 of the Ministry of Health and Declaration of Helsinki).

  - *Experimental procedure:* The experiment consists of two stages where the stimuli presentation ratio proposed by Bezdjian et al. (2009) is 80% of Go stimuli and 20% of No-go stimuli. In each stage, 150 trials are presented and a total of 320 trials are presented, taking into account the test phase, as specified in Hanley, Burns, Thomas, Marstaller, and Burianová [50]. The difference between phase 1 and phase 2 is that the Go stimulus is exchanged. Before starting the experiment, the participant is presented with a series of 10 trials to understand the task in the same way in the second stage. In the training trials, feedback is presented to the participant each time he/she makes an error. A bank of images of different categories is available. Twenty images are used for the test. The exposure time is 1 s and the transition time is 500 ms. It should be noted that the experiment can be modified according to the research need. Once the participant finishes the experiment, the information is recorded in a database (MySQL).

Finally, the agent's action framework was defined in the context of the video game according to the selected tests and the chosen architecture. For this, the following aspects were considered: the cognitive model of artificial consciousness, the artificial agent, the work performed by the agent, and the agent's environment. The architecture had to articulate the processes of both perception and execution to the extent of their activation or inhibition according to the tests and the model proposed. The control functions were to be developed when the agent at time t determined what its actions would be at time $t + 1$, displaying the response at each level according to the specific test in progress and the progress of the test.

3.2 Computational model design

In the context of the artificial cognitive system, the presence of conscious perception was considered and thus, a representation of it was required. Conscious perception encompasses the ability to deliberately grasp and make sense of the environment around us. In this sense, it involves the ability to consciously observe and understand the forms and characteristics of the world around us. For example, conscious perception enables us to visualize forms consciously, appreciating the details and particularities of our environment. In this context, the fundamental premise is that perceptual stimuli act as providers of essential information, so that the agent can consciously experience them. This means that stimuli are not only perceived superficially but are interpreted and "felt" as distinctive elements of the environment in which the agent operates. This process of conscious perception implies the active representation of these particularities in the computational structures in charge of collecting and processing information, as is the case of voice and gesture representations in the field of artificial cognitive systems. The objective is that the perceptual stimuli provide the necessary information so that the agent "feels" them as those particularities of the environment in which he/she is operating, and they are represented in the computational structures that collect voice and gestures. The approach employed was based on IIT in a perspective on consciousness and how it might relate to conscious perception and other cognitive functions such as attention. IIT suggests that consciousness arises from the integration of information in a system and that this integrated information has specific properties that can characterize conscious experience. When talking about artificially conscious systems, both intrinsic and extrinsic features may be present in them. Extrinsic measurement measures how the system behaves externally because it is the only way to see phenomenal states. Intrinsic measurement measures the system in its internal state that is capable of conscious states. IIT focuses on the intrinsic characteristics of consciousness. Generally, synthetic phenomenology is measured by reviewing behavioral characteristics; however, as stated by Nazri, Abdul Ghani, Hafez, and Ng [48], evidence of intrinsic consciousness is very relevant because not all organisms and machines show their consciousness through behaviors. However, in our work, it is possible to observe the behavior of the artificial system and then to translate its representation to the transition probability matrix, TPM. That is, we initially examine the behavior of the artificial system using direct observation and from the results of the tests; we translate this observation into the changes that the system undergoes from a time t to the next time $t + 1$ and these changes are collected in the transition probability matrix (TPM).

To design the artificial system, the contents of perception and attention of the human agent were taken to advance the established tests. In addition, taking the IIT as a reference theory for the review of the information that this system integrates (or amount of consciousness), the following variables, states, and relationships relevant to the Go/No-go test were considered in order to select those that best represented the causal system.

3.2.1 Relevant variables

- *Visual Stimulus:* It represents the visual information presented to the participant during the test. Although this may not be a variable in itself, but rather an important design component of the test. It considers elements that are manipulated by the artificial agent such as stimulus type, stimulus characteristics, and stimulus presentation.
- *Inhibitory control*: According to Bezdjian, Baker, Lozano, and Raine [49], the Go-Nogo test encompasses a task that evaluates the inhibition of the motor response, in which certain stimuli must be executed or, on the contrary, inhibited.
- *Response*: This variable reflects the participant's responses to the task, such as whether a target stimulus was correctly identified, or an error was made.
- *Response Time:* The time it takes for the participant to respond to the stimuli could be a relevant variable, as it may be related to processing speed and attention.
- *Speech and gesture recognition*: The ability of the artificial agent to recognize the participant's speech and gestures is important to adapt the test and possibly make it more complex or even decrease its level of difficulty.

3.2.2 Relevant states

- *Working Memory States:* These states would represent the information that the participant is actively holding in his or her mind at any given time during the test.
- *Attention States:* Attention states could be defined to reflect the participant's concentration on the task, which may vary over time.
- *States of Consciousness*: These states could represent levels of awareness, i.e., whether the participant is experiencing conscious perception at a given time or not.

3.2.3 Relevant relationships

The relationships between above variables and states can be modeled using machine learning algorithms and neural networks. For instance, the relationship between visual stimulus and working memory could be represented by a model that updates working memory based on the presentation of new stimuli; the relationship between the participant's response and recognition of speech and gestures could be modeled using natural language processing and computer vision algorithms; the relationship between response time and attention could be associated with processing speed, which can vary as a function of the participant's attention.

In our research, a limited definition of conscious perceptions is made with the possibility of using third-person observation. In a first phase, the sensors acquire information to be understood by the artificial agent and then generate a representation of its own. In a second phase, the artificial agent processes responses of the human agent based on the current test and makes feedback that allows it to learn the behavior of the human agent mainly in terms of the evaluation of the cognitive aspect to be treated, in order to make decisions.

The following components are available for the implementation of the environment: Attention component to detect behavior of the human agent according to the specific contents that are launched by artificial agent; evaluation component, which makes review of the artificial agent's own state based on its mission; learning component to get new knowledge; decision making or management component, to establish the relevant contents and adjustment of parameters in the development of the test.

3.3 Model to characterize the cognitive system

The representation of the problem domain is focused on the characterization of the system and the integrated information structures. The causal structure can be inferred from observational data.

1. The variables defined for the system are previously discretized into two states: on (1) or off (0). This can be achieved by applying a threshold to the values, where values that exceed the threshold are assigned the "on" state and those that do not are assigned the "off" state.
2. Sequences of states must be determined. From the binarized data, sequences of states need to be determined. This can be accomplished through the identification of the sequences of state transitions that occur from the review of the stimuli applied in the test (a number of 300 stimuli were used).
3. Generation of the TPM: Once the state sequences have been identified, it is necessary to calculate the TPM, which is the matrix that describes the state transition probability from moment t to the next moment, $t + 1$. The TPM can be calculated from the frequency of occurrence of each state transition sequence in the stimulus sequence. That is, the values of each time sample are determined by the previous time values.
4. With the revision of the mathematical framework of the IIT (It was also considered relevant to address proposals for mathematical models of consciousness [51]), we started from the $\mathrm{IIT}_{3.0}$ version with a proposal to find the integrated information based on Krohn and Ostwald [52], working with EMD as a distance measure and using the PyPhi toolbox [2]. This strategy is based on a form of application of the truncated product that uses the representation of causal graphical models and is characterized by the gradual elimination of connections.

It is important to note that this process must be done for each of the participants of the go-Nogo test, since each was considered as a particular system.

3.4 System characterization

While it is true that the idea of developing cognitive power and consciousness artificially can be seen as a rather complex process, Chalmers had a slightly different idea. His proposal was because machine learning occurs solely and nothing more than by logical and mathematical calculations. In contrast, Haikonen felt that there would be no way to develop fully functional cognitive machines from mathematical computations alone [53]. In our work, we take elements from Haikonen's proposal, in the sense that machines could have perception by collecting data from the environment or from human operators. Then, this data must be processed to establish whether the system is conscious, to what degree and in what form. That is, to determine how much information the system integrates (its degree of consciousness) by analyzing the causal relationships, since according to IIT, there must be causal interactions for consciousness to exist. In summary, the integrated information of the system quantifies to what extent the system in each state has cause-effect power over itself, as a single system (i.e., irreducibly).

This model proposed for the study of consciousness in artificial systems is based on the cognitive system developed by Rendón et al. [18] to be used as a cognitive architecture and will be explained in the following section. The architecture allows the interaction of the artificial agent and a human agent and stimulates the agent's ability to adjust the execution of the test based on its learning, considering the presence of the cognitive capabilities that are analyzed in the test, as well as by reviewing the components capable of generating them, for which the system should be inspected internally, and feedback should be made based on its learning. To formalize the operation of the artificial system, according to the agent design, the agent perceives the environment using its sensors and thus takes information from its external and internal world and based on the analysis of this information, acts on the environment using its actuators [54–56]. The agent's responses, as well as the impact of its actions, can then be verified by third-person observation. Each cognitive function has defined tests for a specific analysis, to characterize the artificial agent and determine how it performs with respect to the human agent, according to the specific aspect being evaluated, the established criteria, the information that the agent takes from its inner and outer world.

The modeling of cognitive features associated with consciousness has been strongly considered in artificial consciousness, and there have been different approaches in this regard. One approach consists of simulating these cognitive features, using neural networks. These neural networks have been used, for example, to model basic forms of aspects such as volition and attention. Another approach used small recurrent neural networks and wheeled robots to simulate imagination and internal models [57].

Our work was based on the development of a system that by means of an artificial agent in a virtual environment used data from a real environment to launch different actions to a human agent performing a set of cognitive tests. The artificial agent used this model to detect where problems occurred according to the responses of the human agent and based on these responses define how to continue the test.

4 Cognitive model building

4.1 Adaptation of the cognitive architecture

The target of this work at first does not produce new rules or patterns but finds that pattern that may cause more difficulty when the human agent tests. As it was made explicit in the research objectives, a fundamental part of this project was to adapt an artificial cognitive architecture that would allow the integration of different cognitive functions according to the requirements of this research proposal. In this sense, it is convenient to highlight the following aspects:

- Cognitive science is the study of the mind—high-level processes that happen in the brain, such as attention, consciousness, perception, language, and memory.
- IIT understands the brain as a system of interacting mechanisms that can adopt different states depending on input signals.
- Cognitive architecture is the theory of the structure of mind, including human consciousness, and its raison d'être is a computational model of the essence of cognitive knowledge.

In a way, a machine could be conscious if it has learning, perception, action, and memory [53]. These architectures should contribute to the use of the knowledge obtained from the decision-making modules in such a way that cognitive behavior can be generated. Given the above and to characterize the consciousness of an agent, as established by Arrabales et al. [45], the basic components must be defined: a set of architectural modules such as sensors, internal state, sensorimotor coordination, and effectors, which are responsible for implementing processes such as perception, action, and reason. The authors state that both cognition and learning can be generated in an agent as a product of the combination of the above processes as it interacts with the external world as well as with its own interior.

An artificial cognitive system (ACS) must consider aspects of cognition such as perception, learning, and autonomy, among others. The idea is that these aspects can be included as part of a representation in an artificial system. In addition, it is important to consider cognitive control or processes to manage other cognitive processes, which includes directing attention, maintaining working memory, and inhibiting irrelevant signals, among others. This project used as a basis the SCA for cognitive training in older adults developed by Rendón et al. [18], which focused on verbal fluency in older people with mild cognitive impairment. In our research, the cognitive system developed by Rendón et al. [18] was modified to be used as a cognitive architecture; attention was chosen as the cognitive function to be evaluated, due to its relationship with consciousness, and IIT was taken as the computational model of consciousness. The original architecture of the artificial cognitive system presented in Fig. 1 was taken as a basis and modified as presented in Fig. 2.

In the original ACS (Fig. 1), the following color code is used: blue denotes the aspects related to the perception of interest (monitoring of verbal fluency, automatic speech, and gesture recognition); green denotes the aspects related to the action of the ACS (speech synthesis); red highlights the cognitive aspects with which the knowledge obtained from the user interaction modules is used in such a way that cognitive behavior can be generated.

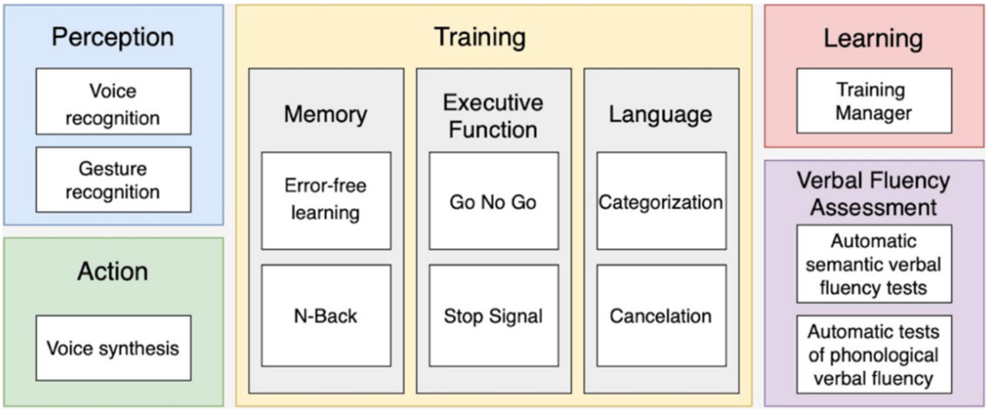
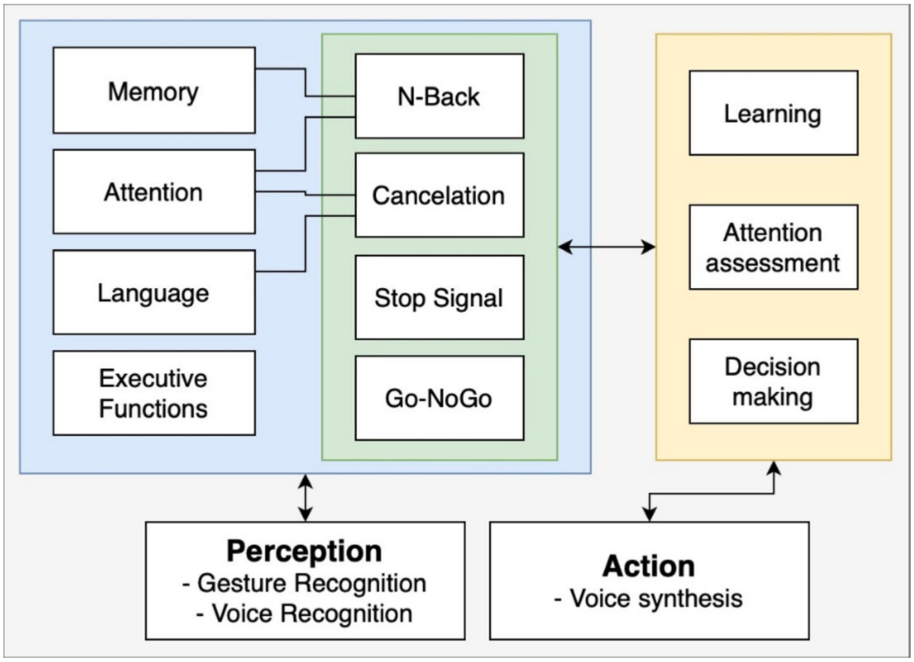


Fig. 1 Artificial cognitive system [18]

Fig. 2 New architecture of the SCA with the modifications



For the construction of the system, as mentioned above, some cognitive aspects associated with consciousness such as attention, memory, executive functions, and perception had to be selected. After reviewing the cognitive tests that fit the needs of the defined population, the available tests were analyzed and those that could be taken to a computational version for their subsequent adaptation to the artificial cognitive system were selected. The artificial cognitive system on which we worked (Fig. 1), as previously mentioned, presents a module for the perception component, another module for the aspects related to the action of the artificial cognitive system and a component for the implementation of the training to be performed according to the cognitive aspect to be studied: memory, executive functions, and attention, with which we have the versatility to include different aspects and even extend the action and perception modules as required. More specifically, by having a modular architecture that allows to separately manage the different components, perception, action, and learning, as well as a specific module for the implementation of training, according to the cognitive aspect to be analyzed. In this way, we have the flexibility to study aspects such as memory and executive functions, and even expand the modules related to the other components (perception and action) as we consider expanding the effective inputs and outputs of the system. Three tests were selected: N-Back (for “working memory”) for memory and Go-Nogo and Stop-signal (for “inhibition training”) for executive functions. A description of each test is presented in Table 1.

Table 1 Description of the tests performed following the proposed model

Test name	Description	Application	Test validity
N-Back	The *N*-back test is a useful indicator for assessing sustained attention span, concentration, and the ability to maintain relevant information in active memory during the performance of a cognitive task The updating process is closely related to the central executive system of working memory, so it is also considered a good task to optimize or measure certain executive functions It has been used in research to study brain activity related to working memory and attention	In this test the subject is presented with a sequence of stimuli, for which he/she must indicate whether it coincides with the one shown *N* steps before. The *N* factor will add complexity the higher its value since it will demand the storage and updating of *N* stimuli. The *N*-Back task can be presented in multiple ways, although the most common is to present a grid (usually 3 × 3) and indicate whether the stimulus was presented in the same position as *N* steps before. This variation is known as visuospatial *N*-Back. It is therefore a very demanding task in terms of working memory, which also requires great attention [58] The presentation of the activity is digital and the implementation of this task, in the framework of the present work, consisted of the presentation of images of different categories and the user was asked to determine whether the current image matched in category with the previous *N* image	While the task has strong face validity and is now widely used as a measure of working memory in clinical and experimental settings, there are few studies exploring the convergent validity of the *n* -back task with other measures of working memory
Go-NoGo	Go-No go activities are a type of exercise that are mainly oriented to stimulate executive functions, mainly automatic response inhibition and cognitive flexibility [59]. But it also stimulates attention, memory, and concentration This type of activity consists mainly of responding to a demand in a way that is contrary to the natural way of doing it	In this test, the subject is presented with a series of drawings. When the drawing of “a participant” appears, he or she must respond aloud “girl”; when, on the contrary, the drawing of “a girl” appears, the student must respond “participant”; and in the rest of the drawings that appear, he or she must say what he or she sees: “dog”, “sun”, etc. In this way, the participant or the girl is forced to omit an answer contrary to the natural one	From theoretical perspectives, a series of tasks aimed at measuring inhibition have also been the subject of much debate. Within this set of tasks, those with a computerized format are of great interest
Stop signal	Like go-no go, the stop-signal paradigm is one of the most widely used procedures in the study of behavioral inhibition. This task seeks to make participants react to a given stimulus in the shortest possible time by looking for two types of behavior; in the first one, subjects are asked to make a specific response to a stimulus, and in the second one, they are asked to inhibit the behavior already initiated, in the presence of specific conditions. The interesting and distinctive feature of this method or procedure is that it requires the inhibition of an ongoing behavior, i.e., of a response that has already been initiated or activated due to the presentation of one of the two stimuli	In the classical paradigm, subjects must perform two concurrent tasks: an execution task and a braking task. In the execution task, the subject must respond as quickly and accurately as possible to the execution signal, which is presented as imperative. Generally, the execution task constitutes a stop-signal choice reaction time task, the subject must perform a primary task such as, for example, pressing the left key of the computer mouse, when the letter *X* is presented and the right one in front of the letter *O*	Although this test provides a very useful measure for assessing inhibitory functioning that is not possible to obtain through other procedures, there are no studies that have implemented this paradigm in the assessment of behavioral inhibition

Table 1 (continued)

Test name	Description	Application	Test validity
Cancelation	This test evaluates an individual's ability to detect and maintain attention on specific stimuli during distractions. That is, it assesses selective attention and the ability to focus on specific stimuli while ignoring distractions. It is a test best suited to assess visual search ability and focused attention in a task that involves finding and marking specific targets amidst a series of distracters	In a typical cancelation test, the individual is presented with a sheet of paper or a screen with a series of stimuli, such as letters, numbers, or geometric shapes, arranged in rows and columns. The participant's goal is to identify and mark (or "cancel") a particular stimulus, such as a specific letter or shape, among the other distractor stimuli Scoring is based on the number of target stimuli correctly canceled and the time taken to complete the task This test is especially useful for assessing selective attention, that is, the ability to focus on a relevant stimulus while ignoring distractions	It is essential to note that the validity of the cancelation test may vary depending on the population being assessed, the context of the assessment, and the specific adaptations of the test used. Furthermore, validity is not a single attribute of the test, but rather a body of evidence that supports the interpretation of the results

4.2 Consciousness model

Consciousness can be linked to various forms of cognitive and behavioral control. The contents of consciousness are often available in ordinary waking consciousness to guide a wide variety of both cognitive and behavioral processes, e.g., those involved in reasoning, verbal report, intentional agency, attentional control, executive processing, memory consolidation, etc. "This facet of consciousness is often captured by saying that the contents of consciousness are globally available for the control of thought and action" [60].

It is necessary to be clear that attention and consciousness are intimately related processes, but they are not the same. There are computational studies that have worked on attention as computational correlates of consciousness [61]. Although it may be possible to use the cognitive aspects, previously defined to identify consciousness in artificial systems, problems may arise due to the attribution of phenomenal states. Another way to test consciousness in artificial systems arises then, and that is by using a theory of consciousness that allows predictions to be made according to the internal architecture and the states presented by the system [57]. Thus, in our research, the cognitive system developed by Rendón et al. [18] was modified to be used as cognitive architecture, attention was chosen as the cognitive function to be evaluated, due to its relationship with consciousness and that, it could be adapted to computational versions and incorporated to the SCA that took the IIT as a model of consciousness.

Software routines for attention are taken as small processes that bring items of interest to consciousness; however, our interest is focused on consciousness as integrated information according to IIT. Then to calculate Φ (integrated information) for a system, one must analyze its actual physical components, identify elements, define its cause-effect repertoires, find concepts, complexes, and determine the spatiotemporal scale at which Φ reaches a maximum [31].

4.3 Development of the computational model for the artificial cognitive system

Regarding perception and attention in the cognitive system, it can be said that perception uses prediction and expectations will affect what will be perceived [62]. In addition, attention affects perception, so perception is not a simple pattern recognition but a process that combines the effect of sensory information and information internal to the system; therefore, it is an impression that, depending on the context and other signals, can change. We focus visual attention on those things that are important for our needs or also that have some emotional significance because they are new, strange, beautiful, etc.

It can be noted that speech recognition is a particular form of auditory perception which is also often a complicated task given the many possibilities that can be presented such as frequency spectrum, timing, intonation, and rhythm. However, it has been detected that, for example, when learning a language, the expectations created by the context and the background information are very important, and that is how in speech recognition, this is not mainly supported by phoneme recognition or its combinations, but the background information is very important.

Now, the brain performs an integration of perceptions coming from different sensory modalities into a congruent representation of the world at a given moment. This integration turns out to be a complex task, since perception from a specific sensory source may affect perceptual processes occurring from another sensory modality. It remains a field of research to find out how sensory information from different kinds of senses is combined.

As stated above, the main interest in this research focuses on aspects such as perception (in the artificial agent) and attention (human agent). It is worth noting that some theories were oriented by the idea of a central unit in charge of controlling perceptions, actions etc., guiding attention. “Attention could then even be a manifestation of the controlling action of the consciousness”. It should be noted that for its part the brain with its large number of neurons and connections between them, must select from the senses what it is going to process, because of its relevance, and leave the rest in the background, and this selection process is known as attention. Basically, there are two kinds of attention, sensory attention, which has to do with the selection of information that can be obtained sensorially, and internal attention, which refers to the selection of information from those internal representations that are relevant [62].

After selecting the cognitive tests, according to the objective of this work, we proceeded to design and implement the computational adaptations of each test.

- *N-Back*: In this test, a sequence of stimuli (visual or auditory) is presented, in which the coincidence with the “N” pattern previously shown must be identified. In this computational adaptation, interspersed images of the different categories included are presented, then, the person performing the test is asked to determine whether the image he/she is currently viewing matches in category with the previous N image. When the categories of the compared images match, the trainee must say the word “equal”, otherwise he/she must remain silent. The system identifies whether the person made a hit or a miss, in the case of a hit it expresses a sentence of congratulations, otherwise it asks the person to concentrate both when the error is due to identifying as equal two images of different categories, and when not identifying the equality of images with the same category; after this message continues with the process of stimulus presentation.
- *Go-No go:* In this test, the person who performs it must be able to distinguish between frequent stimuli and not so frequent stimuli, so that highly automated behaviors are restricted. In this test, the tester is asked to keep a key category in mind and to indicate whether the stimulus presented visually is equal to the key category, if so, he/she should respond with the word equal, otherwise he/she should say nothing. In this computational version, images of three randomly presented categories were used. The system determines whether the person’s answer matches a correct or an error, if it is a correct answer the system congratulates the person by voice command, otherwise, it asks him to concentrate on the task he is performing. For this test, there is a menu that allows selecting the key category, the number of stimuli, and the time; the stimulus remains on screen and the percentage of frequent stimuli (key category) to be presented; infrequent stimuli will correspond to images of the other two categories.
- *Stop signal:* In this test, the person performing the test must be able to inhibit the appearance of the stimulus presenting the stop signal, which generally corresponds to another perceptual channel different from that of the main task. In the version of this test, images of the three categories are presented, between each stimulus a gray image is presented, and during the time that this gray image appears, the stop signal may or may not appear. The trained person must say the name of the category to which the image corresponds but must inhibit his response if prior to the appearance of the stimulus, i.e., while the gray box is presented, the signal sounds. For this training, there is a menu that allows configuring the number of stimuli to be presented, the time that a stimulus remains on the screen, the time between stimuli (gray box) and the percentage of randomness for the appearance of the sounds.
- *Cancelation:* In this test, the person performing the test is shown different stimuli at the same time and is asked to choose among them those that correspond to a certain particularity.

4.4 SCA architecture

As mentioned above, this research was based on the architecture developed in Santiago’s work focused on the evaluation of verbal fluency in older adults, which reproduces the cognitive abilities: perception, action, anticipation, autonomy, and learning.

The following is a brief description of this work in its cognitive components: For perception, regarding automated speech recognition and gesture recognition, the author took the approach of human computer interfaces (HCIs); for action, the way to instruct the user was done with the inclusion of speech synthesis by means of automatically generated human voice messages for which Microsoft’s speech recognition API was used; for anticipation, a system of rules that determine the administration of cognitive training was defined. Finally, reinforced learning algorithms were used to achieve the established objective, as part of what would be the autonomy that allows the system to pursue a defined objective, as well as learning from these same algorithms (machine learning). Thus, the system is framed within the logic of hybrid systems.

Our work, unlike that of Rendón et al [18], is aimed at the cognitive components of attention and executive functions (components commonly associated with consciousness), and the population to which the tests are oriented are young participants.

In the process of adapting the SCA (see Fig. 3), all the components of the architecture were modified, excluding all the elements that are part of the action flow corresponding to interaction through voice and gesture recognition. Therefore, it is indicated that the automatic recognition and synthesis of voice and gesture recognition have been left structured under the base design of the original SCA.

5 The software proposal

In this case, the developed software was designed according to the logic of serious video games considering the principle of Flow in the gamification theory, so that the difficulty in the game is proportional to the development of the skill of the player. Thus, people who play a game, do not enter negative states (boredom, anxiety, etc.), which can occur if, for example, the difficulty in the game is low and the person has a great skill, or if on the contrary, the difficulty is very high, and the person is not skilled enough to play the game.

Given the above, several aspects were considered for each test: configurable parameters and their default values, the number of stimuli to be presented, distracters, time that would be visible on the screen, order of presentation, the messages that would be presented to the user, punctuation rules, etc. It should be noted that for speech and gesture recognition, as well as speech synthesis, the SCA developed in [18] was used.

5.1 Requirements analysis

A psychologist with expertise in testing analysis with individuals of this type served as a developmental supervisor. With her, we began the process of gathering specific test requirements and how the chosen tests could be adapted in a computational environment. With these requirements, we were able to understand the tests and their different variants and variables to be considered. Based on the requirements, we were able to construct the test descriptions in Table 1.

Based on the requirements, we design a set of uses cases to keep a software development process with a formal documentation. The first diagram is shown in Fig. 4, which shows how the player interacts with the Videogame in order to register, sign in, and mange options before starting the game.

On the other hand, with the requirements identified for the test analysis process, the UML diagram as shown in Fig. 5 was generated. This diagram formalizes the process of select options, and later, the player can start the game, where a set of stimuli is show to the player.

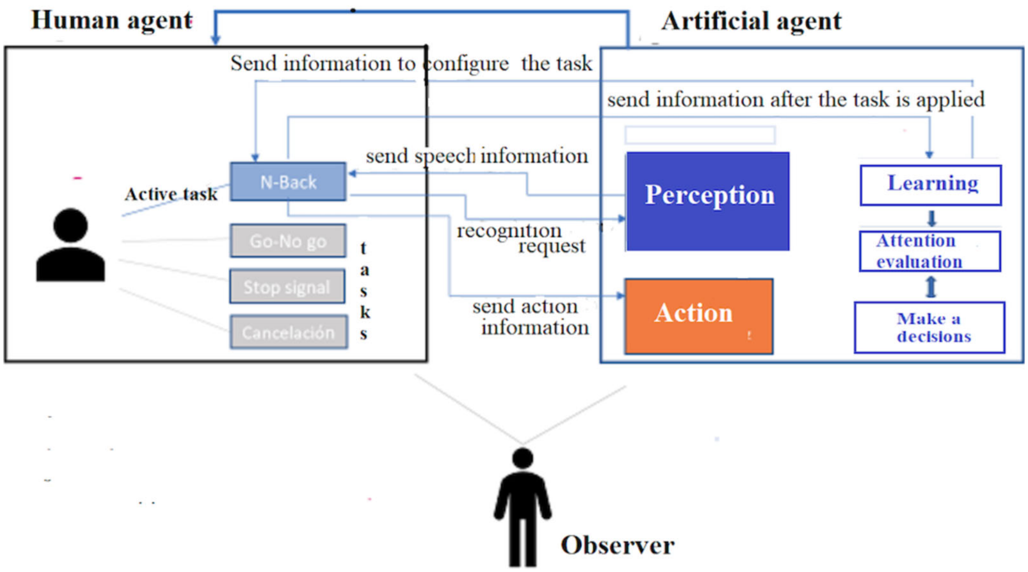


Fig. 3 New architecture of the SCA with the modifications

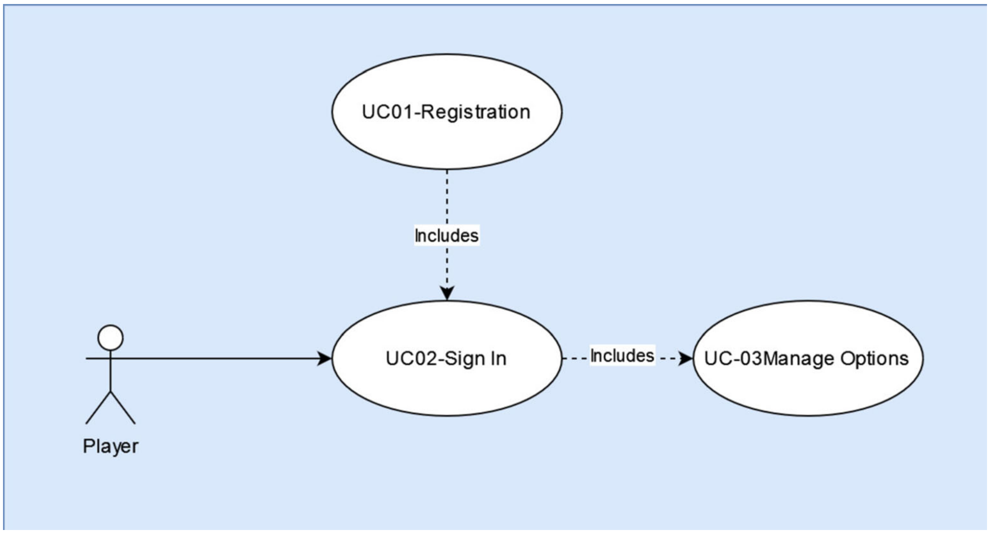
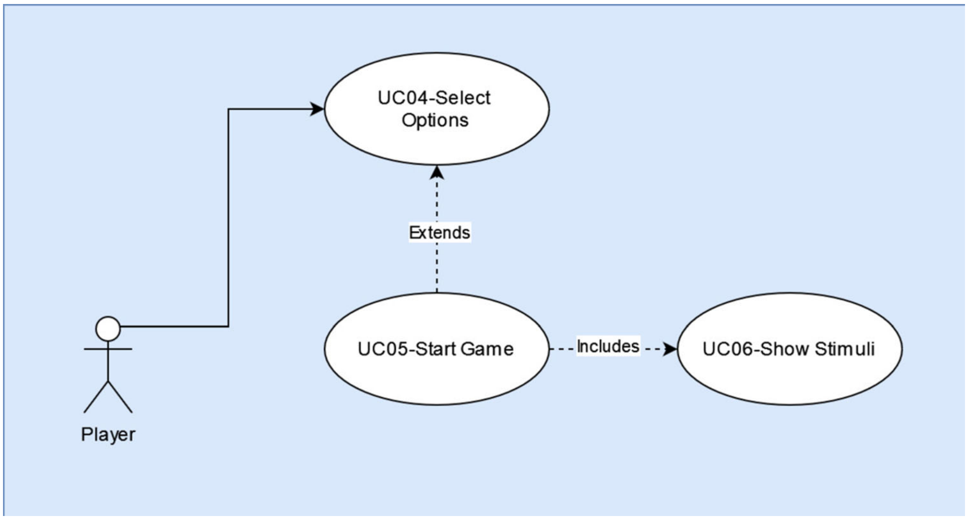


Fig. 4 Use cases from 01 to 03—interaction between player and registration, sign in, and manage options

Fig. 5 Use cases from 04 to 06—interaction between player and game



5.2 IT architecture

In today's data-driven landscape, whether processed in real time or handled asynchronously, there is an increasing demand for substantial computing power. This increase in computational demand is primarily a consequence of the sheer volume of data that needs to be analyzed to generate meaningful results to inform decision making. The goal of this work is to create a historical repository of each player's data, allowing continuous monitoring and tracking over time. This, in turn, facilitates the creation of an evolving performance profile for the player, which serves as a valuable tool for improving future evaluations.

In order to address this issue while ensuring the privacy, security, and accessibility of the data, a server-client architecture has been developed. This architecture relies on web services to facilitate communication between the client and the database. The primary goal behind employing web services is to distinctly delineate the responsibilities of each stage of the data processing workflow, starting from the initial phase and continuing through to its completion.

Figure 6 illustrates the architecture envisioned for the video game's implementation, including the backend that enables the reception of requests via the graphical user interface (GUI) containing player data. Depending on the nature of the requests, two types are involved: GET and POST requests. GET requests are responsible for retrieving and returning data, while POST requests handle the storage and retention of data.

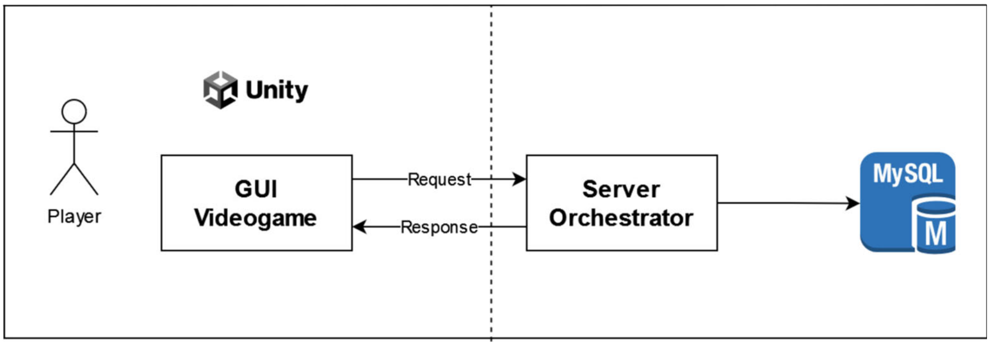


Fig. 6 Proposed IT architecture for the platform

5.3 Videogame development

As mentioned above, this video game was based on the video game application developed as part of the SCA by Rendón et al. [18]. This software was taken as a starting point and the modifications that were made are described below:

- Standardization of the database used in the original game, which implied organizing and structuring the information in a coherent and consistent manner which, in turn, allowed improving the efficiency and integrity of the data, as well as facilitating future updates and modifications.
- The visual aspect of the new video game was adjusted. To achieve this, graphic elements such as images, icons, and animations were modified to adapt them to the needs and objectives of the cognitive testing game.
- Work was done on the user interface, improving the layout of the elements and the overall design to ensure a visually appealing and user-friendly experience for the participants.
- Throughout the development process, extensive testing was also carried out to ensure the proper functioning of the video game, both in terms of game logic and visual presentation, to guarantee optimal performance and a smooth game experience.

In summary, the modifications made under the presented structure are mainly based on the process of database normalization and refactoring of the main classes under which the video game works in the Unity platform, in addition, new tests were incorporated to the game. The above process entailed organizing and structuring the information in a coherent and consistent manner, which allows improving the efficiency and integrity of the data, as well as facilitating future updates and integration processes with other applications that perform cognitive tests. Consequently, the reports generated for each game session can be easily compared with information generated by other tests and thus measure the user's evolution. On the other hand, an adjustment is made to the visual component of the entire video game, thus having the GUI aligned to the target population.

The gamification strategy used for the development of the video game is mainly focused on the user's attention and the interaction that users have with the platform. From the points system, to facing the different challenges that are proposed, all framed within the activities performed in each of the tests, which are part of the cognitive system.

The game was developed in a 2D scenario where the user must interact with graphical and auditory elements that can be interpreted to respond through a click event or by voice to perform the different activities that are proposed for each test.

Regarding the incentives or bonuses received by the user for successfully performing the different tests, the user must follow the instructions defined for each activity and perform them correctly and appropriately to obtain positive scores. In case, the user performs actions that are outside the instructions given or that do not meet the objectives of the test, this event will be recorded, to be later interpreted by the agent and generate a penalty in the result.

The determination of whether an action is right or wrong is determined by the logic of the game according to the test that is executed, as well as it can also be ratified at the end, by the observation in third person; the evaluation of the test will be given by several positive or negative points that are awarded for the different actions.

Another important element that the video game incorporates is the storage of different types of information, including data such as the date on which a user was registered, his name, and schooling, as well as essential data about the player, such as laterality, the activities performed, the points obtained, and the time; it took to perform an activity, among other statistics that could be classified by test.

To start the video game and perform the cognitive tests, the application has a user management module that allows sign in or registering new players with the characteristics (See Figs. 7 and 8, respectively).

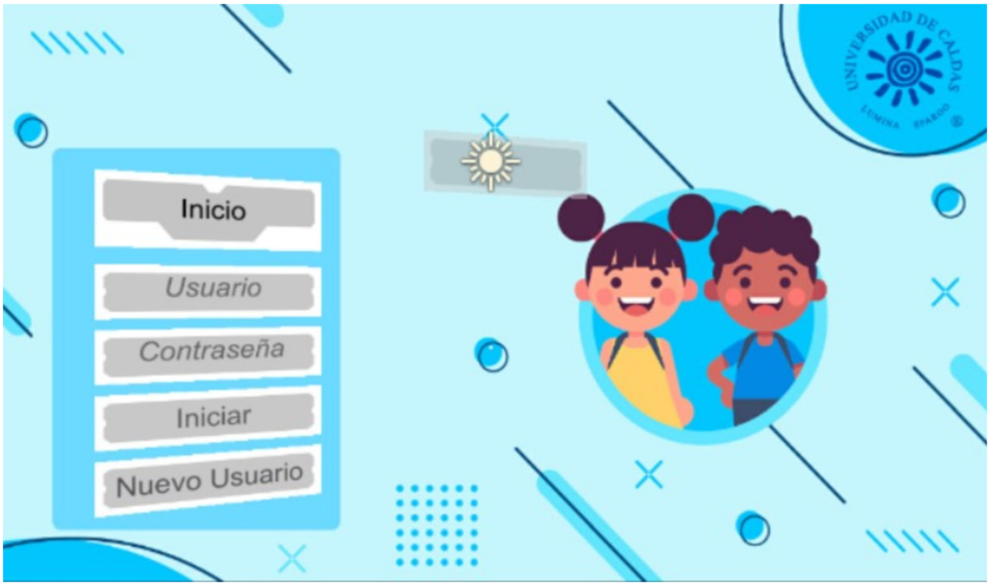
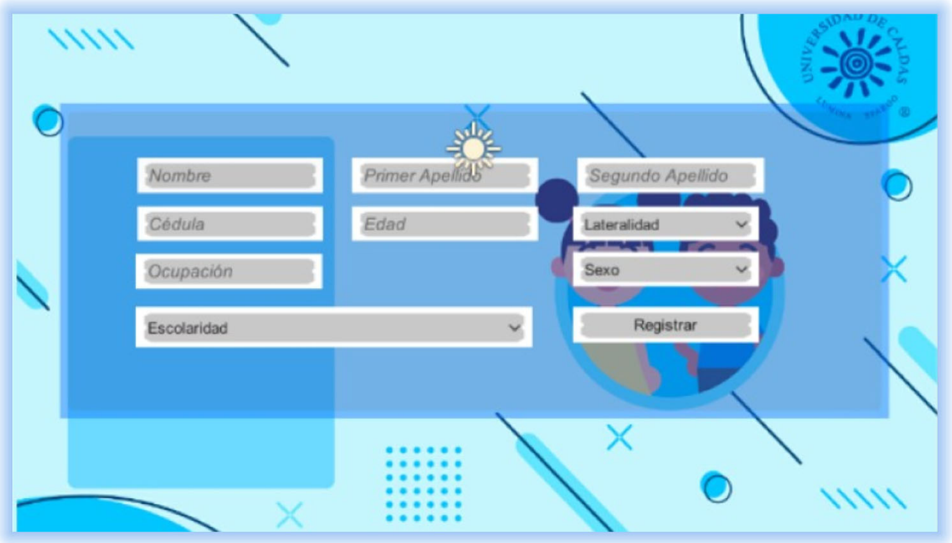


Fig. 7 Sign in form of videogame with access to register player option

Fig. 8 Register for with player data



Once logged in with the assigned credentials, you will have access to the test selection form (Fig. 9) and subsequent test configuration (Fig. 10).

Once the test is selected and the different variables for each of these are configured. The player will start the test with the stimuli (see Fig. 11). Once the process is completed, the player will be taken back to the test selection form in case, it is necessary to run a different test.

6 Validation results

In this work, the interest focused on the modeling of ACS and the measurement of its awareness in the framework of ITI. Recall that a system is just a set of interacting things. Interaction is key because according to the first axiom consciousness exists, and as science can only assume the existence of things that interact with other things; therefore, interaction is the basis of existence.

6.1 Data processing

Likewise, it can be assumed that a physical system that produces consciousness must contain interacting elements.

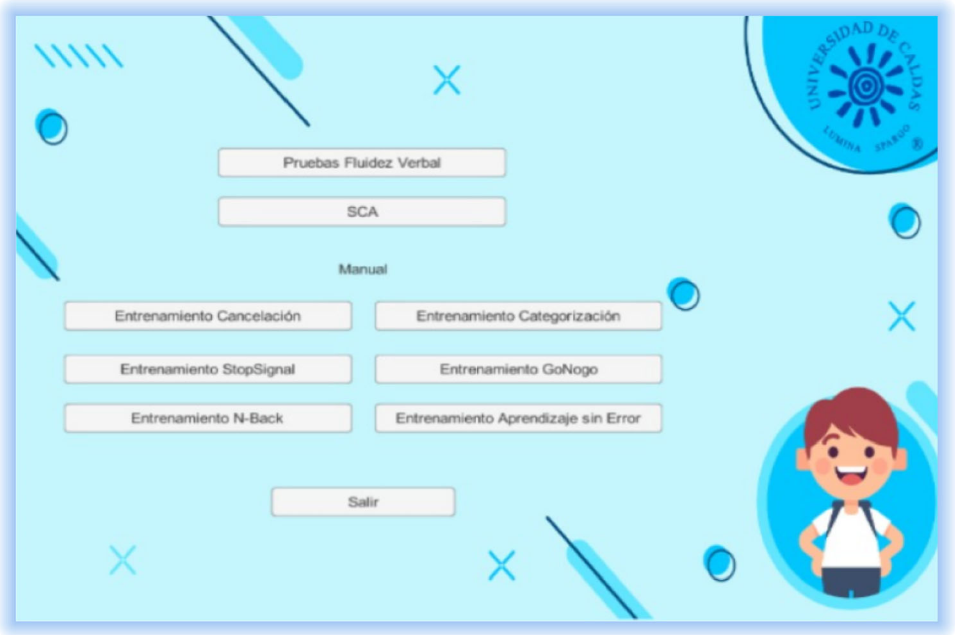
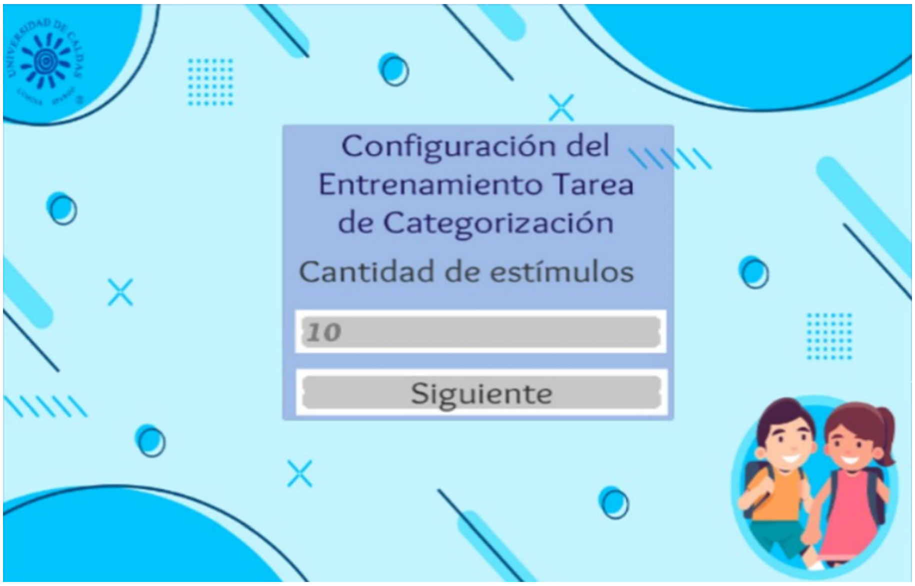


Fig. 9 Form with selection test options


Configuración del Entrenamiento Tarea de Categorización

Cantidad de estímulos

10

Fig. 10 Settings form of the selected test

It can be noted that every system consists of two or more elements, and mathematically, an element can be represented as a discrete variable Xi. The elements are defined by three properties: (1) They can have at least two states, (2) they receive inputs from other elements that can change the states, and (3) they have outputs that depend on these states. The interacting elements X1 to Xn form the system t, where t denotes the present moment (see Fig. 12).

Generally, integrated information has been demonstrated in simple systems composed of logic gates, but in our case, we worked on a nondeterministic system. Similarly, the system can be treated as a discrete dynamical system, as defined above, where each node has an input–output function that delivers the state of the node at the next time $t + 1$ given the state of its parents at the previous time t.

According to IIT for consciousness to exist there must be causal interactions. What is a causal interaction? As each element must have an output that depends on its current state, these outputs become inputs to other elements, these inputs define whether the elements change state, so the use of outputs with individual elements or a combination of elements acting in concert can change the state of other elements which constitutes a causal

Fig. 11 Settings form of the selected test

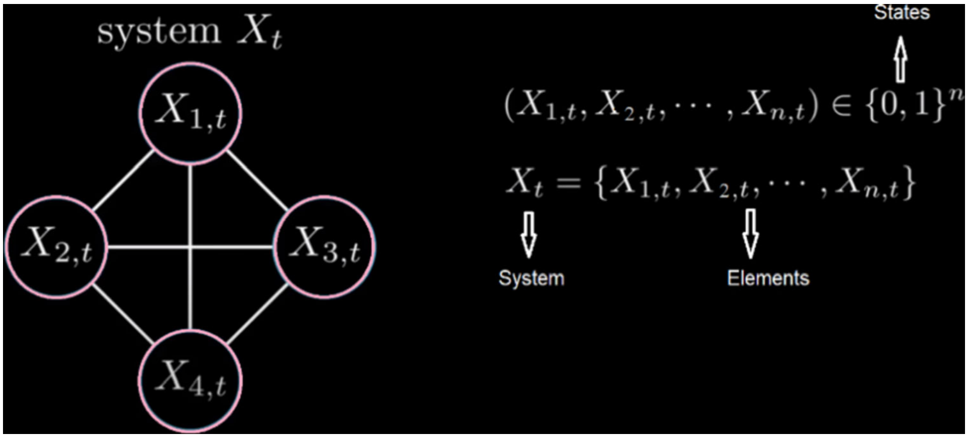


Fig. 12 Representation of a discrete dynamic system

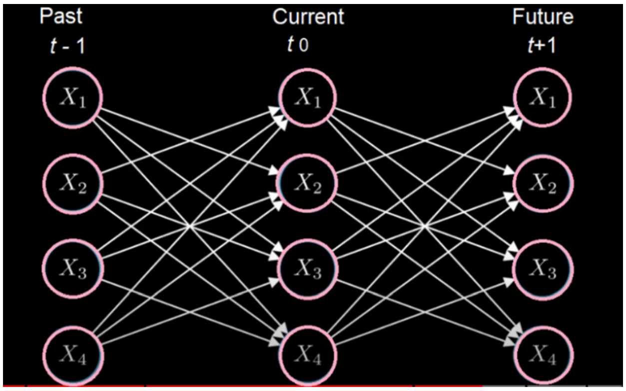


Fig. 13 Representation of the causal structure of the system

interaction. The causal structure can be inferred from observational data. In our case, with the data, we have we try to infer that causality (see Fig. 13). If a subgroup of the system, such as a single element or a combination of them, influences the overall state of the system, it is known as a causal mechanism.

Related to the above, our objective is to define a discrete dynamic system and then represent it as a dynamic causal network using tests such as Go-Nogo. Knowledge of the context allows us to identify the variables to be included, what their relevant states are and how they are related. Here, we are interested in natural and artificial systems, such as neural networks, for which detailed information on the structure of the causal network and the mechanisms of the individual elements of the system is available. These are obtained by identifying the stimulus sequences and for those states that are not achievable, the suggestion was to obtain them through extensive experiments or simulations that evaluate how the system reacts to interventions.

Then, the TPM can be calculated from the frequency of occurrence of each state transition sequence in the stimulus sequence. That is, the values of each time sample are determined by the previous time values. That is, the TPM will have the conditional probability distribution on the next state at *t + 1* given the current state at t. Thus, the cause-effect structure can be defined and the IIT formalism can be applied.

Transition probabilities can be obtained initially and then determined by perturbing the system in all possible states while holding the background variables fixed and observing the resulting transitions. Alternatively, the causal network can be constructed by experimentally identifying the input–output function of each element. However, simple observation of the system without experimental manipulation is insufficient to identify causal relationships in most cases or situations.

The variables defined for the SCA were response (correct or not), participant response time (slow or normal), difficulty setting by the artificial agent, and speech/gesture recognition (yes or no) by the artificial agent (see Fig. 14).

For the practical example shown below, the system was defined with three variables: X_1: Response, X_2: Response time, and X_3: Difficulty setting.

As explained above, the sample consisted of a total of 20 subjects. For each of them, 300 stimuli were collected, and for each stimulus, the three variables defined above were recorded. Initially, the data were collected in a database directly from the test, as it was developed. In addition to the participant's response and time spent of interest, other values of interest were recorded for use by the artificial agent (see Fig. 15).

We defined the discrete states of each variable at each time binarizing with respect to a threshold as in the case of response time. The values defined were as follows: Response = 1 if correct, 0 if incorrect; Response time = 1 (not slow), 0(slow) based on defined threshold; Adjustment difficulty = 1 (if adjustment was made), 0 if adjustment was not made. Once the data were available according to these conventions, the TPM was generated from the frequency of occurrence of each state transition sequence in the stimulus sequence. The TPM has the conditional probability distribution over the next state at *t + 1* given the current state at t, for all possible states.

To exemplify the process and make it simpler, suppose that in a trial, there are 30 stimuli, the data presented here have already been binarized (see Fig. 16).

From the above data, we generated the TPM. The TPM is generated in its State-Node form that describes how many times the current state of a system specifies the possible next states of each channel in an immediately following time (see Fig. 17). This value will basically be calculated as the probability that the next state is at a value, given that the current state is at some given value. For example, the upper left entry of the matrix is ¾, which represents the number of times variable X_1 will be at '1' at time $t_i + 1$ (for this example, (3) given that variables X_1, X_2, *and* X_3, were at '000' at time t_i (for this example, (4). Remember that in our research 300 stimuli are handled.

As the state of a system at time *t* specifies the possible future states of each variable at time *t + 1*. For example, the top left entry of the TPM is 1, which represents the probability that variable X_1 (Response) will be '1' (ON) at time *t + 1*, given that variables X_1, X_2, and X_3 were all at '0' at time *t*. The TPM was generated in the node-state form, as shown below, for one of the experiments performed (see Fig. 18).

Now, at a given state at time *t*, for example $X_1 = 0$, $X_2 = 0$,and $X_3 = 0$, the effect information specified by a subset (called "mechanism", for example X_1) on the future states of another subset (called "purview", for example X_1, X_2, X_3) is given by the probability distribution of the

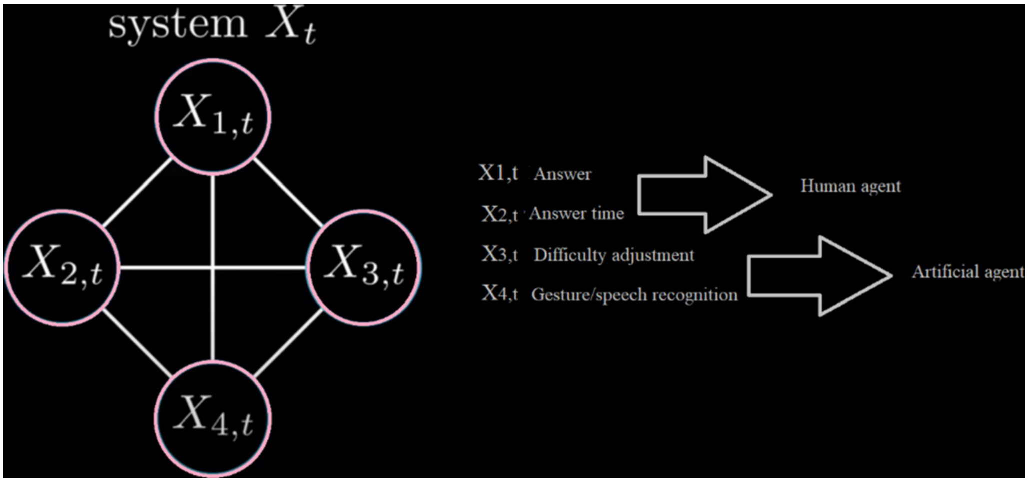
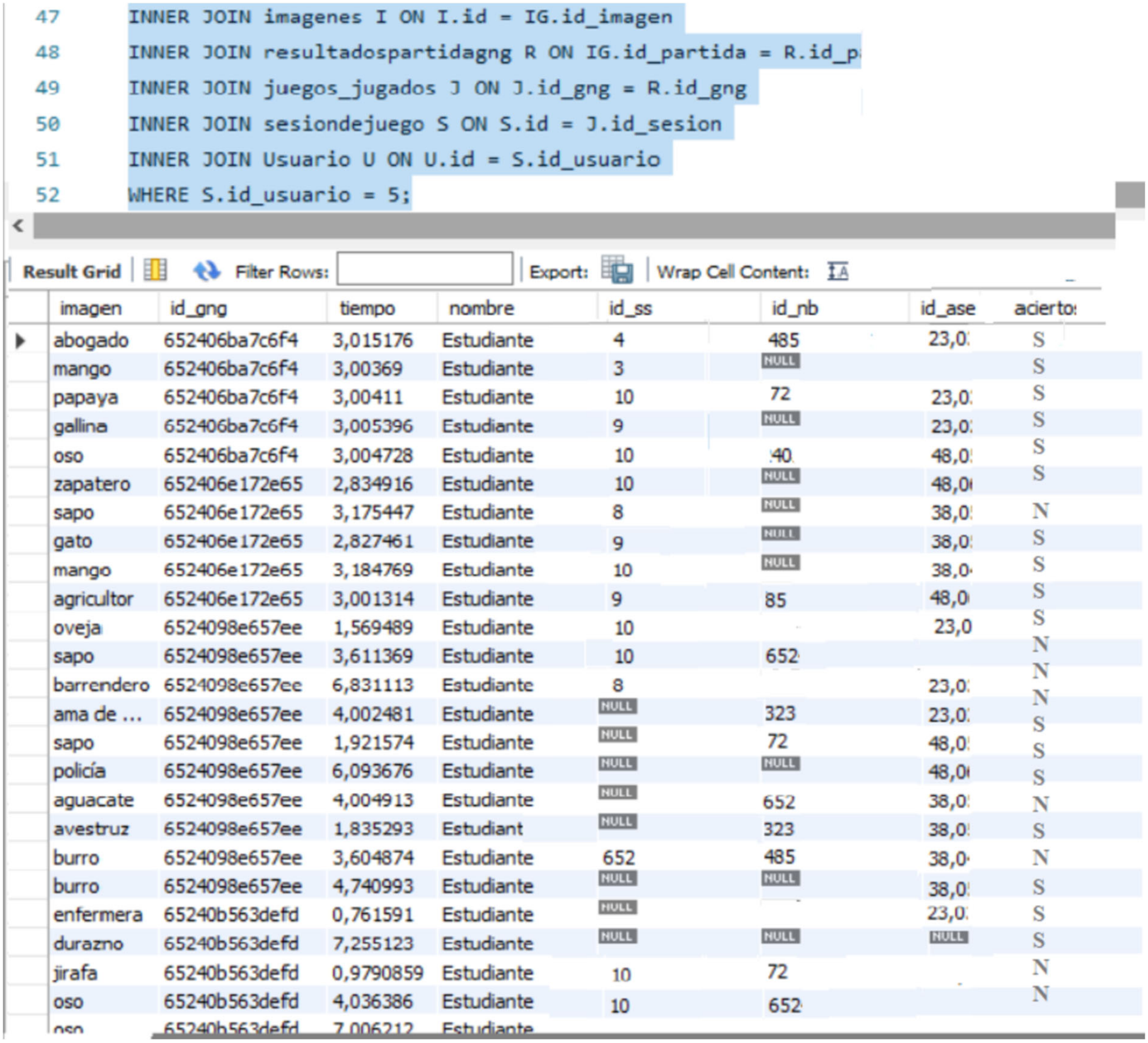


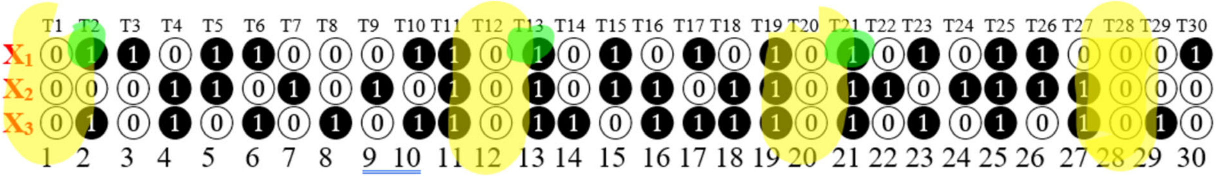
Fig. 14 Representation of the dynamic system with the variables of interest



```
INNER JOIN imagenes I ON I.id = IG.id_imagen
INNER JOIN resultadospartidagng R ON IG.id_partida = R.id_p
INNER JOIN juegos_jugados J ON J.id_gng = R.id_gng
INNER JOIN sesiondejuego S ON S.id = J.id_sesion
INNER JOIN Usuario U ON U.id = S.id_usuario
WHERE S.id_usuario = 5;
```

imagen	id_gng	tiempo	nombre	id_ss	id_nb	id_ase	acierto
abogado	652406ba7c6f4	3,015176	Estudiante	4	485	23,0	S
mango	652406ba7c6f4	3,00369	Estudiante	3	NULL		S
papaya	652406ba7c6f4	3,00411	Estudiante	10	72	23,0	S
gallina	652406ba7c6f4	3,005396	Estudiante	9	NULL	23,0	S
oso	652406ba7c6f4	3,004728	Estudiante	10	40	48,0	S
zapatero	652406e172e65	2,834916	Estudiante	10	NULL	48,0	S
sapo	652406e172e65	3,175447	Estudiante	8	NULL	38,0	N
gato	652406e172e65	2,827461	Estudiante	9	NULL	38,0	S
mango	652406e172e65	3,184769	Estudiante	10	NULL	38,0	S
agricultor	652406e172e65	3,001314	Estudiante	9	85	48,0	S
oveja	6524098e657ee	1,569489	Estudiante	10		23,0	S
sapo	6524098e657ee	3,611369	Estudiante	10	652		N
barrendero	6524098e657ee	6,831113	Estudiante	8		23,0	N
ama de ...	6524098e657ee	4,002481	Estudiante	NULL	323	23,0	N
sapo	6524098e657ee	1,921574	Estudiante	NULL	72	48,0	S
policía	6524098e657ee	6,093676	Estudiante	NULL	NULL	48,0	S S
aguacate	6524098e657ee	4,004913	Estudiante	NULL	652	38,0	N
avestruz	6524098e657ee	1,835293	Estudiant	NULL	323	38,0	S
burro	6524098e657ee	3,604874	Estudiante	652	485	38,0	N
burro	6524098e657ee	4,740993	Estudiante	NULL	NULL	38,0	S
enfermera	65240b563defd	0,761591	Estudiante	NULL		23,0	S
durazno	65240b563defd	7,255123	Estudiante	NULL	NULL	NULL	S
jirafa	65240b563defd	0,9790859	Estudiante	10	72		N
oso	65240b563defd	4,036386	Estudiante	10	652		N
oso	65240b563defd	7,006212	Estudiante				

Fig. 15 Data resulting from the Go-no go test



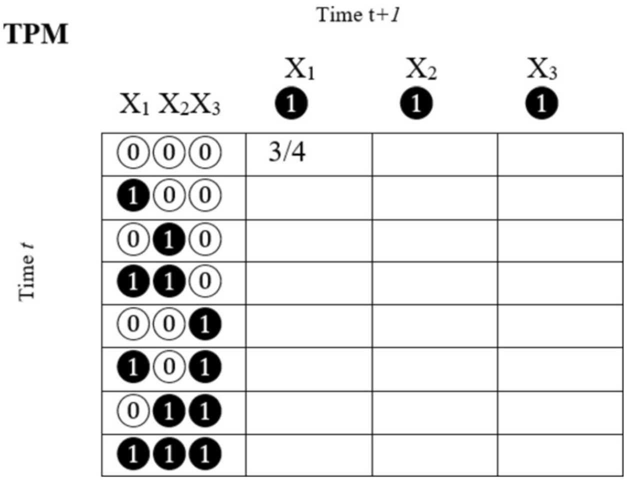
	T1	T2	T3	T4	T5	T6	T7	T8	T9	T10	T11	T12	T13	T14	T15	T16	T17	T18	T19	T20	T21	T22	T23	T24	T25	T26	T27	T28	T29	T30
X_1	0	1	1	0	1	1	0	0	0	1	1	0	1	0	1	0	1	0	1	0	1	0	1	0	1	1	0	0	0	1
X_2	0	0	0	1	1	0	1	0	1	0	1	0	1	0	1	1	0	1	1	0	1	1	0	1	1	1	1	0	0	0
X_3	0	1	0	1	0	1	0	1	0	1	1	0	1	1	0	1	1	1	1	0	1	0	1	0	1	0	1	0	1	0
	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20	21	22	23	24	25	26	27	28	29	30

Fig. 16 Representation of a set of 30 stimuli specified in a time t_i for three binary variables

purview conditioned on the current state of the mechanism (See Fig. 19, left). To calculate the amount of integrated information of a system, we first need to evaluate the information generated by each mechanism, this is done by calculating how much each mechanism affects the system. This influence is quantifiable since the current state of a mechanism limits which states of the system can occur.

Information about the past is also obtained, since only a subset of all possible states of the system can lead to the mechanism being in its current state, this means that a mechanism not only constrains future possibilities but, in a sense, also constrains the past (See Fig. 19, right).

The effect information specified by a subset (mechanism highlighted in the yellow box, in the example, X_1, on the future states of another subset, purview (in this example X_1,



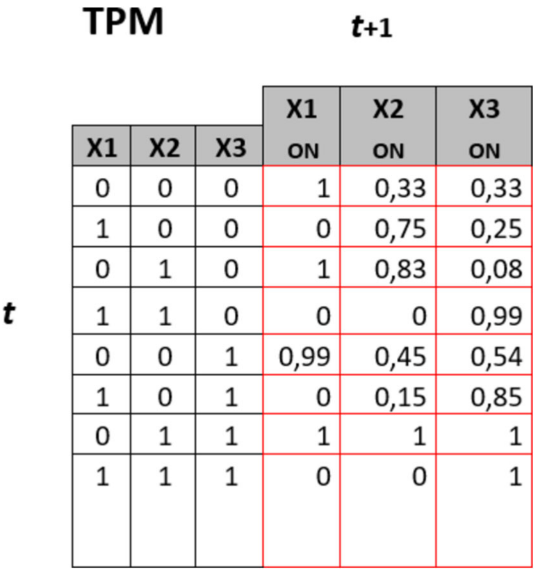
TPM

Time t+*1*

Time *t*

X_1 $X_2$$X_3$	X_1 ❶	X_2 ❶	X_3 ❶
⓪⓪⓪	3/4		
❶⓪⓪			
⓪❶⓪			
❶❶⓪			
⓪⓪❶			
❶⓪❶			
⓪❶❶			
❶❶❶			

Fig. 17 Example system TPM



TPM *t*+1

t

			X1	X2	X3
X1	X2	X3	ON	ON	ON
0	0	0	1	0,33	0,33
1	0	0	0	0,75	0,25
0	1	0	1	0,83	0,08
1	1	0	0	0	0,99
0	0	1	0,99	0,45	0,54
1	0	1	0	0,15	0,85
0	1	1	1	1	1
1	1	1	0	0	1

Fig. 18 Example system TPM

X_2 and X_3), is given by the probability distribution of the purview conditioned on the current state of the mechanism.

To calculate the integrated information effect of the mechanism X_1 on the purview, the disconnections (replacing the connections with random noise connections) between the mechanism and the purview must be found, which approximates the original probability distribution much better. This is done by comparing the disconnected probability distributions with the original distribution, using the earth mover's distance (EMD).

EMD is interpreted as the irreducible information generated over the purview by a mechanism and referred to as the φ effect. Now that it is known how to quantify information about the future that is achieved by considering a given mechanism in its current state, the same principle can be applied to past states. In conclusion, it was presented how to quantify information within a causal system that depends on or is constrained by a mechanism in its current state.

The result of this process is two numbers: the cause information and the effect information, both numbers are specific to a mechanism in a specific state. In this case, we observe that mechanism X1 is currently activated, if it were deactivated both cause and effect information would change accordingly. If you know the current state of the system you can simply repeat everything you have done so far, for each of the mechanisms each time you obtain the cause-and-effect information given the current state of that mechanism.

The final step is to decide whether the cause information or the effect information is more important to the system as it seeks to characterize the capability of a system sees the minimum of the two as a bottleneck. That bottleneck can be considered as the weakest link that breaks the chain, therefore whatever the minimum between cause-and-effect

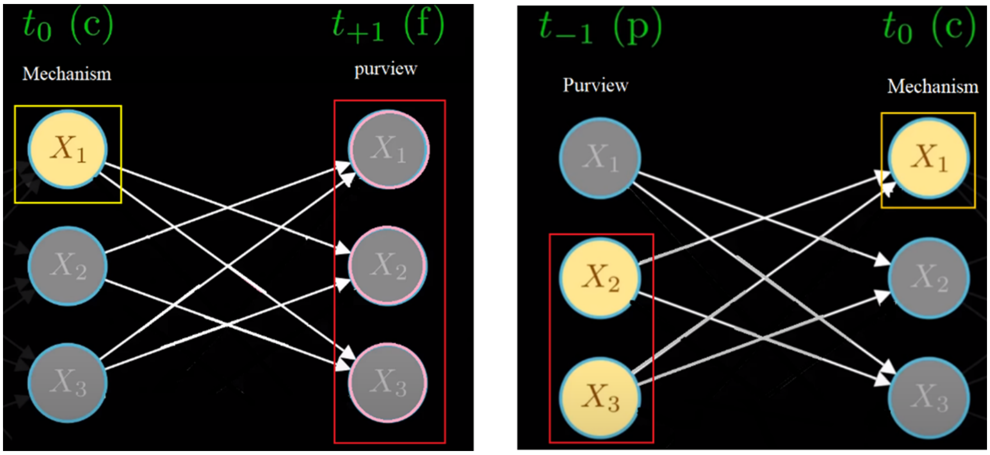


Fig. 19 Left: Mechanism on the future – Right: Mechanism on the past

information is defined as the CAUSE-EFFECT information of the mechanism that is in its current state of the system.

Now, for the integration process, the same probability distributions comparison algorithm can be used for this quantification, for this, we remove each connection between a mechanism and the rest of the system and see if that makes any difference. If the probability distributions do not change after eliminating a connection, it can be deduced that the connection was reducible and there was no integration.

In practice, this is done by a process called statistical noise or marginalization; this consists of replacing the inputs provided by our connection with a random noise distribution. Then, the distance between the partitioned and non-partitioned probability distributions is calculated. Once this has been done for all connections, we determine which elimination resulted in the smallest change. The partition that produces the smallest distance is the MIP, and the result of that calculation is the amount of integrated information or small PHI generated by the mechanism.

Now, both past and future integrated information can be calculated and if both are greater than zero, the minimum can be taken to derive the cause-effect integrated information for the mechanism. This calculation only provided the integrated information at the mechanism level. The same algorithm of replacing connections with noise can be used not only for individual mechanisms but also for complete subsystems of a system.

In the case of our example system, we obtained the values of Fig. 20 for MIP and PHI, taken directly from the PyPHI toolbox.

6.2 Results analysis

Our research did not aim to analyze differences between genders, age groups, or any other classification. Therefore, descriptive statistics (measures of central tendency and dispersion) were used to describe the values obtained by the participants, since our interest was focused on the behavior of the artificial agent in relation to its interaction in the SCA.

Male participants had a mean number of errors before No-Go stimuli of 41.36 (SD 20.56), and for females, the mean was 34.30 (SD 22.21) as shown in Fig. 21.

In reaction time, participants obtained a mean of 514.25 ms (SD 75.85) and women obtained a mean of 563.71 ms (SD 93.27) as shown in Fig. 22.

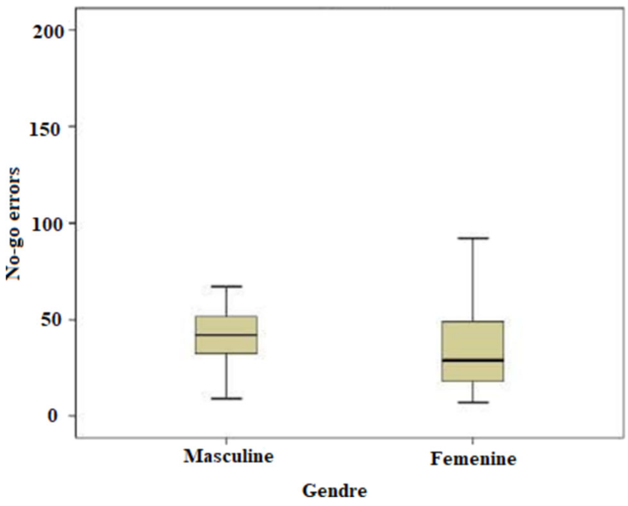


Fig. 21 Measures of errors in No-Go stimuli obtained by gender

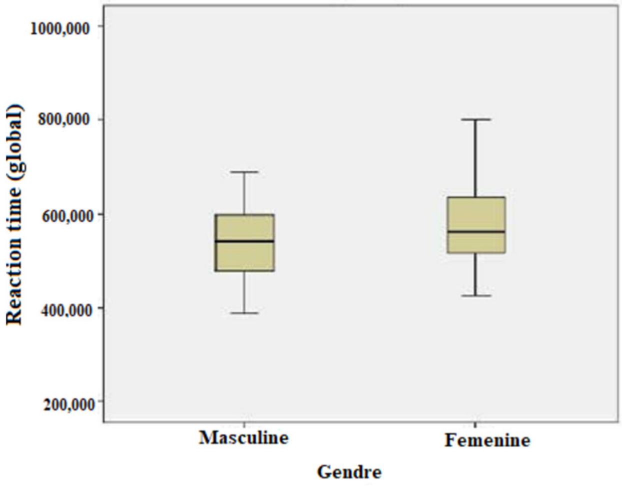


Fig. 22 Measures of errors in No-Go stimuli obtained by gender

The following are the results of the SCA execution, after having applied the test, discretized results and used the PyPhi toolbox for each of the participants. Table 2 summary is presented with the information collected from the participants once they developed the test.

Measuring the information integrated in a cognitive system, as proposed in IIT, implies evaluating the amount of information that is generated by the interactions between the components of the system. This theory suggests that consciousness arises from the integration of information in a system and that this information is greater than the sum of the individual parts. Evaluate how these components interact with each other. For example, the artificial agent

```
MIP Cut [B, C] —/ /—> [A] PHI: 0.003588 TIME: 10.57012
```

Fig. 20 MIP and PHI values in PyPHI for the example case

Table 2 Summary of the SCA execution results

TPM | **Values of PHI and MIP**

Participant 1
TPM

			X1	X2	X3
X1	**X2**	**X3**	**ON**	**ON**	**ON**
0	0	0	1	1	1
1	0	0	0,0	0,0	0,0
0	1	0	1	0	0
1	1	0	0,0	1	0
0	0	1	1,0	1	0
1	0	1	0,0	0,0	1
0	1	1	1	0	1,0
1	1	1	0,0	1,0	1,0

PHI: 0

MIP: Cut [A] ——/ /——➤ [B, C]

Participant 2
TPM

			X1	X2	X3
X1	**X2**	**X3**	**ON**	**ON**	**ON**
0	0	0	1	1	1
1	0	0	0,1	0,0	0,0
0	1	0	1	0	0
1	1	0	0,1	0,95	0,098
0	0	1	1	1	0
1	0	1	0,05	0,03	0,975
0	1	1	0,972	0	0,972
1	1	1	0,081	0,95	1

PHI: 0,040416

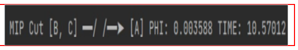
MIP: Cut [A] ——/ /——➤ [B, C]

Participant 3
TPM

			X1	X2	X3
X1	**X2**	**X3**	**ON**	**ON**	**ON**
0	0	0	1	0,33	0,33
1	0	0	0	0,75	0,25
0	1	0	1	0,83	0,08
1	1	0	0	0	0,99
0	0	1	0,99	0,45	0,54
1	0	1	0	0,15	0,85
0	1	1	1	1	1
1	1	1	0	0	1

PHI: 0,040416

MIP:



```
MIP Cut [B, C] —/ /—> [A] PHI: 0.003588 TIME: 10.57012
```

Participant 4
TPM

			X1	X2	X3
X1	**X2**	**X3**	**ON**	**ON**	**ON**
0	0	0	1	1	1
1	0	0	0,0	0,0	0,0
0	1	0	1	0	0
1	1	0	0,0	1	0
0	0	1	1,0	1	0
1	0	1	0,0	0,0	1
0	1	1	1	0	1,0
1	1	1	0,0	1,0	0,0

PHI: 0.000198

MIP: Cut [A] ——/ /——➤ [B, C]

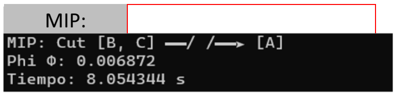
Table 2 continued

Participant 5
TPM

			X1	X2	X3
X1	X2	X3	ON	ON	ON
0	0	0	1	0	0
1	0	0	0	1	0,304
0	1	0	0,98	0,328	0,655
1	1	0	0,05	0	0,095
0	0	1	1	0	1
1	0	1	0	1	0,684
0	1	1	0,98	0,304	0,321
1	1	1	0	0,063	0,938

PHI: 0.006872

MIP:



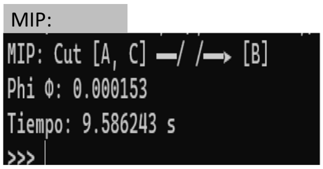
```
MIP: Cut [B, C] —/ /—→ [A]
Phi Φ: 0.006872
Tiempo: 8.054344 s
```

Participant 6
TPM

			X1	X2	X3
X1	X2	X3	ON	ON	ON
0	0	0	1	0	0
1	0	0	0	1	1
0	1	0	1	1	0
1	1	0	0	0	0,051
0	0	1	1	0	1
1	0	1	0,03	1	0,053
0	1	1	0,95	0,947	0,947
1	1	1	0	0,028	0,972

PHI: 0.000153

MIP:



```
MIP: Cut [A, C] —/ /—→ [B]
Phi Φ: 0.000153
Tiempo: 9.586243 s
>>>
```

Participant 7
TPM

			X1	X2	X3
X1	X2	X3	ON	ON	ON
0	0	0	1.0	0.0	0.0
1	0	0	0.026	1.0	0.947
0	1	0	1.0	1.0	0.0
1	1	0	0.026	0.026	0.026
0	0	1	1.0	0.0	1.0
1	0	1	0.0	1.0	0.0
0	1	1	1.0	1.0	1.0
1	1	1	0.0	0.0	1.0

PHI: 0.054429

MIP: Cut [B, C] ——/ /——➤ [A]

Participant 8
TPM

			X1	X2	X3
X1	X2	X3	ON	ON	ON
0	0	0	1.0	0.0	0.0
1	0	0	0.0	0.977	0.364
0	1	0	0.981	0.472	0.528
1	1	0	0.0	0.0	0.516
0	0	1	0.925	0.175	0.775
1	0	1	0.017	0.862	0.569
0	1	1	1.0	0.366	0.366
1	1	1	0.0	0.0	1.0

PHI: 0.004584

MIP: Cut [B, C] ——/ /——➤ [A]

Table 2 continued

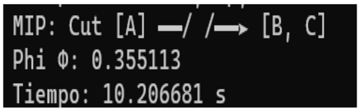
Participant 9

TPM

			X1	X2	X3
X1	X2	X3	ON	ON	ON
0	0	0	0	0	0
1	0	0	0	1	0
0	1	0	0,987	0	0,95
1	1	0	0	0	0
0	0	1	0	1	1
1	0	1	0	1	1
0	1	1	1	0	0
1	1	1	0	0	0

PHI: 0.355113

MIP:



```
MIP: Cut [A] —/ /—➝ [B, C]
Phi Φ: 0.355113
Tiempo: 10.206681 s
```

Participant 10

TPM

			X1	X2	X3
X1	X2	X3	ON	ON	ON
0	0	0	0,9000	0,9000	0,8000
1	0	0	0,4419	0,8140	0,7907
0	1	0	1,0000	0,0000	0,9565
1	1	0	0,5106	0,6383	0,2766
0	0	1	0,8000	0,5111	0,0000
1	0	1	0,3231	0,1231	0,8154
0	1	1	0,0000	1,0000	0,0000
1	1	1	1,0000	0,3158	0,0000

PHI: 0,30633

MIP: Cut [B, C] ——/ /——➤ [A]

Participant 11

TPM

			X1	X2	X3
X1	X2	X3	ON	ON	ON
0	0	0	1,0000	1,0000	1,0000
1	0	0	0,2619	1,0000	0,7381
0	1	0	1,0000	0,0000	0,9423
1	1	0	0,4615	0,6410	0,3077
0	0	1	0,7805	0,4878	0,0000
1	0	1	0,3286	0,2000	0,8000
0	1	1	0,2500	0,7857	0,0000
1	1	1	0,9444	0,2778	0,0000

PHI: 0,352329

MIP: Cut [B, C] ——/ /——➤ [A]

Participant 12

TPM

			X1	X2	X3
X1	X2	X3	ON	ON	ON
0	0	0	0	0	0
1	0	0	0	1	0
0	1	0	0,99	0	0,96
1	1	0	0,86	0,29	0,57
0	0	1	0	0	0
1	0	1	0	1	1
0	1	1	1	0,03	0
1	1	1	0	0	0

PHI: 0,280531

MIP:

```
MIP:  Cut [A, B] —/ /—➤[C]
```

Table 2 continued

Participant 13
TPM

			X1	X2	X3
X1	**X2**	**X3**	**ON**	**ON**	**ON**
0	0	0	1	0	0
1	0	0	0	1	1
0	1	0	1	1	0
1	1	0	0	0	0,05
0	0	1	1	0	1
1	0	1	0,03	1,00	0,05
0	1	1	0,95	0,95	0,95
1	1	1	0,00	0,03	0,97

PHI: 0,061698

MIP:

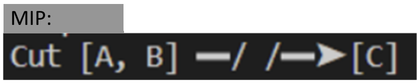

Cut [C] —/ /—➤[A, B]

Participant 14
TPM

			X1	X2	X3
X1	**X2**	**X3**	**ON**	**ON**	**ON**
0	0	0	1,00	0,00	0,00
1	0	0	0,00	1,00	0,30
0	1	0	0,98	0,33	0,66
1	1	0	0,05	0,00	0,10
0	0	1	1,00	0,00	1,00
1	0	1	0,00	1,00	0,68
0	1	1	0,98	0,30	0,32
1	1	1	0,00	0,06	0,94

PHI: 0,010005

MIP:


Cut [A, B] —/ /—➤[C]

Participant 15
TPM

			X1	X2	X3
X1	**X2**	**X3**	**ON**	**ON**	**ON**
0	0	0	1,00	0,00	0,00
1	0	0	0,00	1,00	0,30
0	1	0	1,00	0,33	0,67
1	1	0	0,05	0,00	0,10
0	0	1	1,00	0,00	1,00
1	0	1	0,00	1,00	0,68
0	1	1	0,98	0,30	0,32
1	1	1	0,00	0,06	0,94

PHI: 0,01138

MIP: Cut [A, B] ——/ /——➤ [C]

Participant 16
TPM

			X1	X2	X3
X1	**X2**	**X3**	**ON**	**ON**	**ON**
0	0	0	0,33	0,00	0,33
1	0	0	0,04	1,00	0,00
0	1	0	0,99	0,06	0,99
1	1	0	0,83	0,67	0,33
0	0	1	0,83	0,00	0,00
1	0	1	0,01	0,93	1,00
0	1	1	1,00	0,02	0,00
1	1	1	0,17	0,83	0,83

PHI: 0,174369

MIP: MIP: Cut [A, B] ——/ /——➤ [C]

Table 2 continued

Participant 17
TPM

			X1	X2	X3
X1	X2	X3	ON	ON	ON
0	0	0	1,00	0,60	0,60
1	0	0	0,32	0,49	0,45
0	1	0	0,97	0,26	0,23
1	1	0	0,33	0,60	0,05
0	0	1	0,94	0,62	0,29
1	0	1	0,21	0,31	0,67
0	1	1	0,83	0,43	0,83
1	1	1	0,29	0,61	0,84

PHI:	0,014182
MIP:	Cut [A, C] ──/ /──➤ [B]

Participant 18
TPM

			X1	X2	X3
X1	X2	X3	ON	ON	ON
0	0	0	0,94	0,83	0,78
1	0	0	0,34	0,71	0,63
0	1	0	0,94	0,10	0,84
1	1	0	0,35	0,74	0,07
0	0	1	0,81	0,52	0,08
1	0	1	0,06	0,12	0,88
0	1	1	0,20	0,84	0,24
1	1	1	0,83	0,26	0,39

PHI:	0,166328
MIP:	Cut [B, C] ──/ /──➤ [A]

Participant 19
TPM

			X1	X2	X3
X1	X2	X3	ON	ON	ON
0	0	0	0,93	0,79	0,71
1	0	0	0,42	0,65	0,55
0	1	0	0,93	0,39	0,55
1	1	0	0,33	0,44	0,44
0	0	1	0,90	0,41	0,41
1	0	1	0,12	0,34	0,64
0	1	1	0,29	0,76	0,35
1	1	1	0,79	0,32	0,53

PHI:	0,036549
MIP:	Cut [A] ──/ /──➤ [B, C]

Table 2 continued

Participant 20
TPM

			X1	X2	X3
X1	**X2**	**X3**	**ON**	**ON**	**ON**
0	0	0	1	1	1
1	0	0	1,0	0,0	0,0
0	1	0	0	0	0
1	1	0	0,0	1	0
0	0	1	1,0	1	0
1	0	1	0,0	0,0	1
0	1	1	1	0	1,0
1	1	1	0,0	1,0	1,0

PHI: 0

MIP: Cut [A] ──//──➤ [B, C]

may adjust the difficulty of tests according to the responses of the human agent. These interactions can be expressed in terms of quantitative data, such as the frequency of difficulty adjustment relative to the human agent's responses.

Interpreting the results in terms of integrated information could imply that the higher the value of Φ calculated, the greater the integrated information in the system, which could be associated with a higher level of awareness or attention in the human agent. In the data presented only two had a value of Phi = 0, for the rest of the systems analyzed, there was some degree of information integration. Once we reviewed the data to understand why these two results had a value of zero '0', we found that the initial state from which we started was never reachable in any of the application of the stimuli, and we decided not to make a change by exploring another initial state for these two results, since all the tests were done with the same initial state. In that order of ideas, our results will be referred to 18 samples corresponding to each of the participants for which the PHI value was different from zero.

Figure 23 shows the values obtained for PHI, which ranged from 0.0001 to 0.35. The high value of PHI could suggest that the system has a high degree of complexity and potentially a higher level of awareness or integration. This could allow for more efficient adaptation to the participant's needs. A significantly lower PHI value in the context of a participant's performance could indicate a lower ability of the system to adapt, understand patterns, and operate with a high level of integration. This could lead to less efficient operation and possibly less successful performance on the task, especially when difficulty needs to be adjusted according to individual performance.

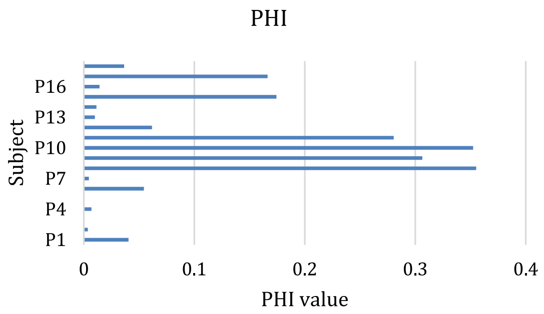


Fig. 23 Values obtained for PHI between 0, 0001 and 0, 35

Fig. 24 shows the relationship between the adjustments made by the artificial agent and each image, i.e., when the complexity of the displayed image increased, or by time, when the exposure time decreased. As the PHI values increased, the main change was in the complexity of the image.

Figure 25 shows the type of adjustments made by the agent to each test participant. In general, the preference was for adjustment for image complexity. In the cases where the adjustment preference was for time, it can be observed in the event sequences, that there is a recurrent pattern in the difficulty adjustment. The agent tends to decrease the time when in specific sequences, the participants responded relatively well to the image change.

Figure 26 shows the relationship between the value of PHI and the number of adjustments made by the artificial agent. It can be seen that in particular where higher PHI levels were obtained, the number of adjustments made by the agent was higher. When reviewing the data in more detail, we found that they corresponded to those participants who had a better performance in the tests and that the agent increased more frequently the difficulty adjustments, and on the contrary, in general, for those who had a lower performance, the number of adjustments was lower.

According to the above and considering that integrated information is used to measure the complexity of interactions in a system, for the case of an artificial cognitive system such as the one presented in this work, integrated information can help to evaluate the sophistication and processing capacity of the system. A higher level of

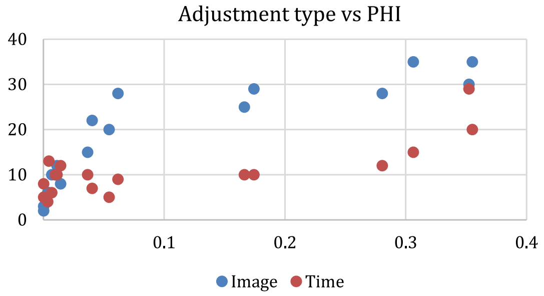


Fig. 24 Relationship between the adjustments made by the artificial agent and each image

integrated information could indicate a higher adaptive and learning capacity of the system. Thus, integrated information could be a basis for the development of more adaptive and autonomous systems.

The precedents show that the lack of adaptability of artificial cognitive systems to the conditions of each individual is one of the particularities that limit their scope, we found that by using integrated information as a reference that allows understanding how complex interactions contribute to decision making in these artificial systems, systems could be designed to better adjust to changing environments and make more informed decisions.

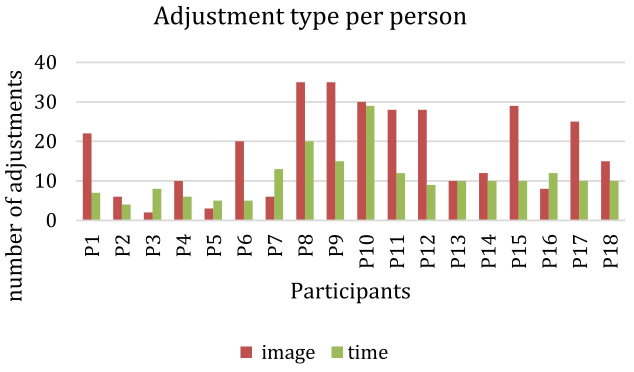


Fig. 25 Type of adjustments made by the agent to each test participant

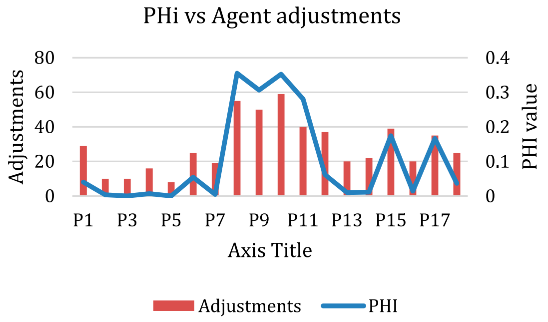


Fig. 26 Relationship between the value of PHI and the number of adjustments made by the artificial agent

7 Conclusions and future work

7.1 Conclusion

In the long term, the notion of integrated information could be relevant in the quest to create conscious artificial intelligence systems. While we are far from fully understanding consciousness, IIT provides a theoretical framework that could be relevant in exploring how consciousness might emerge in highly interconnected artificial systems. In this vein, regardless of one's beliefs about artificial consciousness, it is undeniable that advances can be made not only at the level of artificial agents with improved interaction and reasoning capabilities but also in the understanding of neural processes and understanding of human consciousness.

The measurement of integrated information could also be used to develop more adaptive and autonomous systems. By understanding how complex interactions contribute to decision making in artificial systems, we could design systems that better adjust to changing environments and make more informed decisions.

A challenge in designing artificial cognitive systems lies in the characteristics of information, as acquired from the outside world, that transcends the usual symbolic representations, in order to more closely approximate how these representations, occur in the human brain. This could even be a manifestation of experience, to the extent that meaning is constructed based on experience. Accordingly, future work would focus on the development of artificial systems that use information, with the possibility of an environment based on a trained neural network architecture using associative learning features, taking advantage of a multilayer design, as learning with multiple layers between input and output layers has multiple layered updates to minimize the error between the achieved and desired result and taking the integrated information as a measure of the work of the associative neurons.

IIT is a scientific theory of consciousness that presents an evolving theoretical framework with a mathematical model that allows one to determine whether something or someone is conscious. The idea is to use that mathematics in the form of computational algorithms and look at the relationship to neural data, and then take that mathematical formalism of IIT and apply it to data representing systems in which integrated information can be considered as, for example, a feature of the irreducibility of that system among many other aspects that could be analyzed.

In this project, consciousness was defined as integrated information, and in this sense, for a dynamic system, a model is defined with a human agent and an artificial agent on which the amount of information that the system integrates can be determined. In this sense, the artificial agent is a fundamental component of the system that in a future work could evolve to an agent that, understanding those possible problems in cognitive aspects such as attention and executive functions, helps those people with deficits in cognitive aspects to overcome them, from specific trainings that feed back to the agent and reinforce its learning capacity. In practical applications, the measurement of integrated information could improve the interaction between humans and machines. For example, in medical assistance systems, the understanding of integrated information could be used in the design of support systems that are better adapted to the patient's needs.

Finally, to develop a model of artificial consciousness that integrates different cognitive capabilities, the idea is that aspects of cognition such as perception, learning, and autonomy can be included as part of the representation in an artificial system, and consider cognitive control including attention direction, maintaining working memory, and inhibiting irrelevant signals, among others. All these aspects are brought to the architecture under the concept of modularity. By having a modular architecture that allows separate management of the different components, there is a module for aspects related to perception, another one for aspects related to action, another one in charge of learning, as well as a specific module for the implementation of training, according to the cognitive aspect to be analyzed. In this way, it provides the flexibility to study aspects such as memory and executive functions, and even expand the modules related to the other components (perception, action, etc.) as required to expand the effective inputs and outputs of the system.

7.2 Future work

In a forthcoming paper, we hope from the value of integrated information (PHI) to understand how these, and other variables are related from a causality perspective (according to integrated information theory). This could reveal whether there is some complexity in the interaction of these variables. Overall, we hope to be able to identify patterns, assess agent adaptability, and gain insight into how these variables are related in the context of different cognitive aspects commonly associated with consciousness.

Our contribution focuses on the model definition since to specify values the video game had to be applied to the target population in order to generate the TPM and calculate PHI and compare the results of different participants to obtain a more complete view of the performance of the artificial agent in the system.

Acknowledgements We thank the Vice Rector's Office for Research and Graduate Studies of the Universidad de Caldas for supporting the project "Development of a model for consciousness that allows its study and measurement in artificial cognitive systems in the framework of the Integrated Information Theory (IIT)" with code 0312321. This proposal was approved under the General Call for Research Projects 2020. Also, this work was carried out in the framework of the Project Proto-type for DDoS attack detection based on machine learning in a cloud service-oriented architecture, with code 54308, 0069122; approved in the framework of the Joint Call for Research, technological development and Innovation—2020, of the National University of Colombia and University of Caldas.

Data availability The datasets generated during and/or analyzed during the current study are available from the corresponding author on reasonable request.

Declarations

Conflict of interest The authors have no conflict of interest with this proposal.

References

1. Tononi G (2004) An information integration theory of consciousness. BMC Neurosci 5(1):42
2. Mayner WGP, Marshall W, Albantakis L, Findlay G, Marchman R, Tononi G (2018) PyPhi: A toolbox for integrated information theory. PLoS Comput Biol 14(7):e1006343
3. Churchland P (2013) Matter and consciousness. The MIT Press, Cambrigde
4. Crick F (1995) Astonishing hypothesis: the scientific search for the Soul. Scribner Reprint Edition, New York
5. Penrose R (1996) Shadows of the mind: a searchfor the missing science of consciousness. Oxford University Press, New York. https://doi.org/10.1093/oso/9780195106466.001.0001
6. Edelman GM (1992) Bright air. Basic Books, Brilliant Fire
7. Dennett D (1991) Consciousness explained: Penguin Books
8. Oizumi M, Albantakis L, Tononi G (2014) From the phenomenology to the mechanisms of consciousness: integrated information theory 3.0. PLoS Comput Biol 10(5):e1003588. https://doi.org/10.1371/journal.pcbi.1003588
9. Tononi G (2015) Integrated information theory. Scholarpedia 10:4164
10. Kitazono J, Kanai R, Oizumi M (2018) Efficient algorithms for searching the minimum information partition in integrated information theory. Entropy 20:173

11 Barrett A (2014) An integration of integrated information theory with fundamental physics. Frontiers Psychol. https://doi.org/10.3389/fpsyg.2014.00063

12 Massimini M, Ferrarelli F, Huber R, Esser SK, Singh H, Tononi G (2005) Breakdown of cortical effective connectivity during sleep. Science 309:2228–2232

13. Casali A, Grosseries O, Rosanova M, Boly M, Sarasso S, Casali K, Casarotto S, Bruno M, Laureys S, Tononi G, Massimini M (2013) A theoretically based index of consciousness independent of sensory processing and behavior. Sci Trans Med 5(198):10

14 Arsiwalla XD, Verschure PFMJ (2016) The global dynamical complexity of the human brain network. Appl Netw Sci. https://doi.org/10.1007/s41109-16-0018-8

15 Wilson R, Keil F (1999) The MIT enciclopedia of cognitive sciences (MITECS). The MIT Press, Cambridge

16 Koch C, Tononi G (2008) Can machines be conscious? IEEE Spectr 45(6):55–59. https://doi.org/10.1109/MSPEC.2008.4531463

17. Arrabales R (2011) Evaluation and development of consciousness in artificial cognitive systems

18. Rendón SM, Giraldo BS, De Mejía FR (2020) Sistema computacional para el entrenamiento cognitivo en adulto mayor (caso:fluidez verbal)," Universidad Autónoma de Manizales

19. Manzotti R A physicalist causally oriented foundation for a conscious machine based on the spread mind

20. Ivancevic VG, Ivancevic TT (2007) Introduction: human and computational mind, studies in computational intelligence (SCI)," computational mind: a complex dynamics perspective, pp. 1–269, Berlin: Springer-Verlag

21. Kriegel U (2006) Consciousness, theories of. Philos Compass 1(1):58–64

22. Chalmers DJ (1995) Facing up to the problem of consciousness. J Conscious Stud 2:19

23. Tegmark M (2015) Consciousness as a state of matter. Chaos Solitons Fractals 76:238–270

24 Fekete T, Edelman S (2011) Towards a computational theory of experience. Conscious Cognit 20(3):807–827. https://doi.org/10.1016/j.concog.2011.02.010

25 Ramachandran V, Hirstein W (1998) Three laws of qualia: What neurology tells us about the biological functions of consciousness. J Conscious Stud 4:429–457

26 Koch C, Massimini M, Boly M, Tononi G (2016) Neural correlates of consciousness: progress and problems. Nat Rev Neurosci 17(5):307–321. https://doi.org/10.1038/nrn.2016.22

27. Lagercrantz H, Changeux JP (2010) Basic consciousness of the newborn. Semin Perinatol 34(3):201–206

28. Fink SB (2016) A deeper look at the "Neural Correlate of Consciousness." Front Psychol 7:1044–1044

29. Albantakis L, Barbosa L, Findlay G, Grasso M, Haun AM, Marshall W, Mayner WGP, Zaeemzadeh A, Boly M, Juel BE, Sasai S, Fujii K, David I, Hendren J, Lang JP, Tononi G (2023) Integrated information theory (IIT) 4.0: formulating the properties of phenomenal existence in physical terms. PLOS Comput Biol 19(10):e1011465. https://doi.org/10.1371/journal.pcbi.1011465

30. Balduzzi D, Tononi G (2008) Integrated information in discrete dynamical systems: motivation and theoretical framework. PLoS Comput Biol 4(6):e1000091

31 Tononi G, Koch C (2015) Consciousness: here, there and everywhere? Philos Trans Royal Soc B Biol Sci 370(1668):20140167. https://doi.org/10.1098/rstb.2014.0167

32. Tononi G (2008) Consciousness as integrated information: a provisional manifesto. Biol Bull 215(3):216–242

33. Hidaka S, Oizumi M (2018) Fast and exact search for the partition with minimal information loss. PLoS ONE 13(9):e0201126

34. Tegmark M (2016) Improved measures of integrated information. PLoS Comput Biol 12(11):e1005123

35. Toker D, Sommer FT (2019) Information integration in large brain networks. PLoS Comput Biol 15(2):e1006807

36. Barrett A, Seth AK (2011) Practical measures of integrated information for time-series data. PLoS Comput Biol 7(1):e1001052

37. Griffith V, Koch C (2014) Quantifying synergistic mutual information," guided self-organization: inception, M. Prokopenko, eds, Berlin, Heidelberg: Springer Berlin Heidelberg

38 Gamez D, Holland O (2017) Artificial intelligence and consciousness. Reference module in neuroscience and biobehavioral psychology. Elsevier. https://doi.org/10.1016/B978-0-12-809324-5.05918-6

39. Searle JR (1980) Minds, brains, and programs. Behav Brain Sci 3(3):417–424

40 Hildt E (2019) Artificial intelligence: does consciousness matter? Frontiers Psychol. https://doi.org/10.3389/fpsyg.2019.01535

41. Gamez D (2020) The relationships between intelligence and consciousness in natural and artificial systems. J Artific Intell Conscious 07(01):51–62

42 Wendin G (2019) Can biological quantum networks solve NP-Hard problems? Adv Quantum Technol 2(7–8):1800081

43. Nilsson NJ (2009) The quest for artificial intelligence. Cambridge University Press, Cambridge

44 Block N (1995) On a confusion about a function of consciousness. Behav Brain Sci 18(2):227–247. https://doi.org/10.1017/S0140525X00038188

45. Arrabales R, Ledezma A, Sanchis A ConsScale: a pragmatic scale for measuring the level of consciousness in artificial agents

46 Baars BJ (1993) A cognitive theory of consciousness. Cambridge University Press, Cambridge

47. Kanade V (2020) A Hybrid Evolutionary Algorithm for Evolving a Conscious Machine pp. 1045–1054

48. Chandra R (2017) Towards an affective computational model for machine consciousness," neural information processing, lecture notes in computer science, pp. 897–907

49. Bezdjian S, Baker LA, Lozano DI, Raine A (2009) Assessing inattention and impulsivity in children during the Go/NoGo task. Br J Dev Psychol 27(2):365–383

50. Hanley CJ, Burns N, Thomas HR, Marstaller L, Burianová H (2023) The effects of age bias on neural correlates of successful and unsuccessful response inhibition in younger and older adults. Neurobiol Aging 131:1–10. https://doi.org/10.1016/j.neurobiolaging.2023.07.004

51 Kleiner J (2020) Mathematical models of consciousness. Entropy 22(6):609. https://doi.org/10.3390/e22060609

52. Krohn S, Ostwald D (2017) Computing integrated information. Neurosci Conscious 1:2017

53. Patnaik LM, Kallimani JS (2017) Promises and limitations of conscious machines," Self, Culture and Consciousness: Interdisciplinary Convergences on Knowing and Being, Menon S, Nagaraj N and Binoy VV, eds., pp. 79–92, Singapore: Springer Singapore

54. Huhns M, Singh M (1997) Readings in agents: Morgan Kaufmann Publishers Inc., 1997

55. Wooldridge M, Rao AE (1999) Foudations of Rational Agency Kluwer Academic publishers

56. Wooldridge M (2002) An introduction to multiagent systems

57 Holland O, Gamez D (2009) Artificial intelligence and consciousness. Encyclopedia of consciousness. Elsevier, pp 37–45. https://doi.org/10.1016/B978-012373873-8.00004-9

58 Gajewski PD, Hanisch E, Falkenstein M, Thönes S, Wascher E (2018) What does the n-back task measure as we get older? Relations between working-memory measures and other cognitive functions across the lifespan. Frontiers Psychol. https://doi.org/10.3389/fpsyg.2018.02208

59. Ramos-Galarza C, Jadán-Guerrero J, Ramos D, Bolaños M, Ramos V (2017) Evaluación Neuropsicológica Del Control Inhibitorio Y El Control De La Interferencia: Validación De Tareas Experimentales En El Contexto Ecuatoriano. Revista Ecuatoriana de Neurología 26:27–34

60. Bayne T, Hohwy J, Owen AM (2016) Are there levels of consciousness? Trends Cogn Sci 20(6):405–413

61 Reggia JA, Katz G, Huang D-W (2016) What are the computational correlates of consciousness? Biol Inspir Cognit Archit 17:101–113. https://doi.org/10.1016/j.bica.2016.07.009

62 Haikonen PO (2015) The cognitive approach to conscious machines. Imprint Academic, Exeter