UNIVERSIDAD CARLOS III DE MADRID

Escuela Politécnica Superior

DOCTORAL THESIS

Evaluation and Development of Consciousness in Artificial Cognitive Systems

AUTHOR

Raúl Arrabales Moreno

ADVISORS

Araceli Sanchis de Miguel
Agapito Ledezma Espino

COMPUTER SCIENCE DEPARTMENT

Leganés, February 2011

UNIVERSIDAD CARLOS III DE MADRID

Escuela Politécnica Superior

TESIS DOCTORAL

Evaluación y Desarrollo de la Conciencia en Sistemas Cognitivos Artificiales

AUTOR

Raúl Arrabales Moreno

DIRECTORES

Araceli Sanchis de Miguel
Agapito Ledezma Espino

DEPARTAMENTO DE INFORMÁTICA

Leganés, Febrero de 2011

DEPARTAMENTO DE INFORMÁTICA

Escuela Politécnica Superior
Universidad Carlos III de Madrid

Evaluación y Desarrollo de la Conciencia en Sistemas Cognitivos Artificiales

AUTOR: Raúl Arrabales Moreno

DIRECTORES: Araceli Sanchis de Miguel
Agapito Ledezma Espino

TESIS DOCTORAL

EVALUATION AND DEVELOPMENT OF CONSCIOUSNESS IN ARTIFICIAL COGNITIVE SYSTEMS

Autor: Raúl Arrabales Moreno

Directores: Araceli Sanchis de Miguel

Agapito Ledezma Espino

Firma del Tribunal Calificador:

Firma

Presidente: Pedro Isasi Viñuela

Vocal: Rafael Martínez Tomás

Vocal: Juan Manuel Corchado Rofríguez

Vocal: Antonio Chella

Secretario: José Manuel Molina López

Calificación:

Leganés, 7 de Febrero de 2011.

A mi hermano Jesús

"Nuestra capacidad de empatía nos diferencia de los animales, pero lo que realmente hace grandes a las personas es saber comprender, aceptar y enmendar los errores que todos cometemos".

Pilar Moreno Carbonero.

Agradecimientos

A hombros de gigantes. Como no podría ser de otro modo, esta tesis no hubiera sido posible si no hubiera mirado al mundo desde la altura de los hombros de unos gigantes a los que admiro profundamente. No me refiero sólo a los gigantes académicos, sin los que sin duda este trabajo tampoco existiría, sino que también me complace agradecer la incalculable ayuda que he recibido de unos gigantes extraordinarios, mi familia, que me han dado lo más importante, su apoyo incondicional. Esta tesis es el resultado de cuatro años de trabajo en los que, durante demasiadas horas al día, he disfrutado de la ayuda de colegas, compañeros y amigos, y también he privado a mi familia y a Diana de mi ayuda y compañía. Quiero agradecer a todos ellos su ayuda, su paciencia y su comprensión.

Gracias a mis directores de tesis, Araceli y Agapito, por su apoyo para la realización de una tesis tan particular y recordarme de vez en cuando la necesidad de descender del limbo de la filosofía al mundo real de la ingeniería.

También me complace agradecer su ayuda a numerosos expertos en diversas áreas de las ciencias cognitivas, que desinteresadamente han dedicado parte de su escaso tiempo disponible para ayudarme con la tesis. Gracias a Pentti Haikonen por su constante apoyo y su inestimable ayuda durante mi estancia en el Nokia Research Center de Helsinki. Gracias a Anil Seth y Owen Holland por su ayuda durante mi estancia en la Universidad de Sussex, que finalmente me permitió ver el final de esta tesis. Gracias a Ricardo Sanz, Manolo González Bedia, Carlos Hernández, Jaime Gómez, Xabier Barandiarán, Gal Kaminka, David Gamez, Hugo Gravato, Adam Barret, Trung Doan, Ron Chrisley, Pierre Bonzon, Juan Escasany, Stan Franklin, Antonio Chella, Kevin O'Regan, Stuart Hameroff, Richard Stanley, Igor Aleksander, Philip Hingston, Gregor Schöner, Ben Goertzel y Nick Mayer por invitarme a participar en los debates más interesantes y descubrirme nuevos puntos de vista desde los que estudiar la conciencia.

Durante el desarrollo de esta tesis, también he usado código de terceros y he recibido gran ayuda con cientos de problemas técnicos, de diseño y de desarrollo software. Gracias por esta ayuda a Francisco Javier González López, Trevor Taylor, Ben Axelrod y especialmente a Jorge Muñoz, cuyo esfuerzo y destreza fueron clave para conseguir la victoria en la competición 2K BotPrize 2010. Gracias a las empresas y organizaciones que de una forma u otra han apoyado la realización de esta tesis doctoral: UC3M, Nokia Reseach Center, Sigma Data Services, 2K Marin, Retecog, EuCogII, Nexociencia y Arrabales Motor.

Por último, estoy muy agradecido a todos mis compañeros del grupo CAOS que hacen posible un extraordinario ambiente de trabajo y que hacen que estar en el despacho 2.1.B14 sea todo un privilegio, que incluye la compañía, paciencia y ayuda de José Antonio, las amenas y entretenidas visitas esporádicas de Juan, la compañía de JMAW cuando nos quedamos a trabajar demasiado tarde, y entre otras muchas más cosas agradables, ser el punto de encuentro no oficial del grupo. Debo agradecer también a Paz, Jorge, Javi, Carlos, Germán, y ahora también Andrés, que vengan a despegarme literalmente de mi sitio cada día para ir a desayunar y a comer. No sé qué sería de mí sin vosotros, ni sé que sería de mi cabeza sin poder desahogarme durante las comidas. Santa paciencia la de Bea, Paula y Paz cuando cuento esos chistes que sólo hacen gracia a JP.

Acknowledgements

On giant's shoulders. As it is always the case, this thesis wouldn't have been possible if I hadn't stared at the world from the shoulders of giants I profoundly admire. I'm not only talking about giants from academia, whose help has undoubtedly made this work possible, but I'm also glad to thank the inestimable help I got from some extraordinary giants, my family, who gave me the most important thing, their unconditional support. This thesis is the result of four years of work during which, too many hours a day, I enjoyed the help of colleagues and friends, and I've also deprived my family and Diana from my own help and company. I want to thank them all for their help, patience, and understanding.

Thanks to my thesis advisors, Araceli and Agapito, for their support for the development of such a particular thesis as this one, and reminding me from time to time of the need to come down from the limbo of philosophy to the real world of engineering.

I am also grateful to many experts in diverse areas of cognitive science, who selflessly have dedicated part of their little available time to help me out with the thesis. Thanks to Pentti Haikonen for his constant support and inestimable help during my stay at Nokia Research Center in Helsinki. Thanks to Anil Seth and Owen Holland for their help during my stay at the University of Sussex, which finally allowed me to glimpse a way to the end of this thesis. Thanks to Ricardo Sanz, Manolo González Bedia, Carlos Hernández, Jaime Gómez, Xabier Barandiarán, Gal Kaminka, David Gamez, Hugo Gravato, Adam Barret, Trung Doan, Ron Chrisley, Pierre Bonzon, Juan Escasany, Stan Franklin, Antonio Chella, Kevin O'Regan, Stuart Hameroff, Richard Stanley, Igor Aleksander, Philip Hingston, Gregor Schöner, Ben Goertzel, and Nick Mayer for inviting me to participate in the most interesting debates and showing me new points of view to study consciousness.

During the development of this thesis, I've also used code from others and I've received great help with hundreds of technical problems, and design and development issues. Thanks for this help to Francisco Javier González López, Trevor Taylor, Ben Axelrod, and specially Jorge Muñoz, whose effort and skills were key to win the 2K BotPrize 2010 competition. Thanks to the companies and organizations that some way or another supported this doctoral thesis: UC3M, Nokia Reseach Center, Sigma Data Services, 2K Marin, Retecog, EuCogII, Nexociencia and Arrabales Motor.

Lastly, I'm very grateful to all my colleagues at CAOS group, who make possible an extraordinary work atmosphere and the big priviledge of being in the 2.1.B14 office, including the company, patience and help of José Antonio, the funny and amusing visits of Juan, the company of JMAW when we stay working late, and amongst many other nice things, being the unofficial meeting point of the group. I'm also indebted to Paz, Jorge, Javi, Carlos, Germán, and now also Andrés, for literally unstick me from the chair everyday for breakfast and lunch. I don't know what would be of me without you, nor what would be of my mind without the chance to unburden myself during lunch time. Thanks for the great patience of Bea, Paula, and Paz when I tell that kind of jokes that are only funny for JP.

Resumen

Históricamente el fenómeno de la conciencia humana ha sido en buena medida apartado del debate científico, siendo su estudio relegado casi exclusivamente al ámbito de la filosofía. Sin embargo, durante las últimas tres décadas se ha producido un interés creciente por el problema de la conciencia en diferentes disciplinas como la filosofía de la mente y la psicología cognitiva. Esta tendencia también se ha producido paralelamente en el ámbito multidisciplinar de las neurociencias. De hecho, la aparición de nuevas técnicas de diagnóstico por imagen ha propiciado que actualmente la mayoría de investigadores considere que la conciencia es susceptible de estudio científico.

La formulación de nuevas teorías, tanto biológicas como psicológicas, acerca de la producción de la conciencia en los humanos, ha dado lugar a que se retomen algunos de los objetivos originales de la Inteligencia Artificial. Concretamente, se ha empezado a reconocer el campo de la Conciencia Artificial como una disciplina científica seria que se encarga del estudio y la posible construcción de máquinas con diferentes tipos y niveles de conciencia.

En el ámbito de la Conciencia Artificial, la presente tesis pretende contribuir al conocimiento científico de la conciencia por medio de dos líneas de investigación relacionadas: la primera consiste en la concepción y aplicación de un método inédito para la medición y caracterización del nivel de desarrollo de la conciencia en un agente artificial, la segunda se basa en una arquitectura cognitiva artificial cuyo diseño se ha inspirado en diversas teorías de la conciencia.

La utilización del método de medición propuesto permitirá analizar en detalle cuál es el nivel de desarrollo actual de máquinas conscientes y establecer cuáles son los aspectos que no se ha conseguido explicar o implementar hasta la fecha. Además, la escala propuesta se podrá utilizar como hoja de ruta para identificar cuáles son las habilidades cognitivas cuya implementación es necesaria para construir máquinas que muestren comportamientos equivalentes a los típicamente humanos.

La aplicación de la arquitectura cognitiva propuesta como parte fundamental de los sistemas de control autónomo de agentes artificiales permitirá la experimentación con diferentes funciones cognitivas asociadas a la conciencia. Se estudiarán por tanto las interacciones entre diferentes capacidades como la atención, las emociones o la predicción sensorial, intentando descubrir las sinergias que potencialmente dan lugar a comportamientos complejos y adaptativos. Adicionalmente, usando el modelo computacional de conciencia implementado, se aplicará un enfoque de fenomenología sintética, que consiste en el modelado del contenido de la experiencia consciente. Se comparará la experiencia consciente descrita por un ser humano expuesto a ciertos estímulos perceptivos con el contenido explícito que la arquitectura cognitiva es capaz de generar al enfrentarse a los mismos estímulos.

Los resultados de estas líneas de investigación proporcionarán información valiosa acerca de la validez de las teorías de la conciencia analizadas y de las diferencias encontradas entre los procesos cognitivos naturales y los generados artificialmente. Asimismo, se explorarán posibles áreas de aplicación práctica de la arquitectura cognitiva implementada, como por ejemplo, la creación de agentes artificiales cuyo comportamiento sea percibido por los usuarios como comportamiento humano.

Abstract

Historically human consciousness has been rather excluded from the scientific debates, being philosphy the most important perspective for its study. However, over the last three decades different research disciplines such as philosophy of mind and cognitive psychology have shown a growing interest on the problem of consciousness. This trend has taken place also in the multidisciplinary context of neuroscience. In fact, recent advances in neuroimaging techniques have led most researchers to consider consciousness as a subject for scientific study.

The development of new biological and psychological theories on the production of consciousness in humans has revived the original challenges of Artificial Intelligence. Specifically, the field of Machine Consciousness is becoming a rigorous scientific discipline aimed at studying and potentially creating machines with different types and levels of consciousness.

In the context of Machine Consciousness, this thesis aims at contributing to the scientific knowledge about consciousness by means of two interrelated research lines: the first one consists of the conception and application of a novel method for the measurement and characterization of the level of development of consciousness of an artificial agent; the second one is based on an artificial cognitive architecture inspired on several theories of consciousness.

The application of the proposed measuring technique will permit the detailed analysis of the current level of development of conscious machines and to identify what are the aspects that have not yet been achieved. Furthermore, the proposed scale will be used as a roadmap to identify what are the key cognitive skills that need to be implemented in order to create human-like machines.

The application of the proposed cognitive architecture as a fundamental component of the control system of different artificial agents will permit the experimentation with different cognitive functions associated to consciousness. The interaction between different capabilities like attention, emotions or sensory prediction will be analyzed, looking for potential synergies that produce complex and adaptive behaviors. Additionally, using the proposed computational model of consciousness, a synthetic phenomenology approach will be adopted based on the modeling of the contents of conscious experience. The conscious experience reported by a human subject when confronted to certain stimuli will be compared with the explicit content that the cognitive architecture is able to generate when confronted to the same stimuli.

The results obtained from these research lines will provide valuable information about the validity of the theories of consciousness that have been analyzed as well as the differences between natural and artificial cognitive processes. Besides, possible areas of application of the proposed cognitive architecture will be explored, such us the creation of artificial agents able to develop believable human-like behaviors.

Índice general

Índice de figuras

Índice de tablas

Índice de Ecuaciones

1 Introduction

Understanding human consciousness is considered one of the greatest challenges of modern science. In fact, the challenge is not only in the study of human beings, but also in the study of other potentially conscious entities such as other animals and even machines created by man. The present doctoral thesis focuses on the paradigm of Machine Consciousness, a research field dedicated to the scientific and engineering study of consciousness in machines.

According to the Merriam-Webster dictionary consciousness is *"the quality or state of being aware especially of something within oneself"*. As an additional definition consciousness can be also regarded as *"the state of being characterized by sensation, emotion, volition, and thought"*. Although former definitions might seem to indicate that consciousness is a clearly understood and identified process, a slightly deeper analysis reveals the great variety of opinions and schools of thought that have arisen around consciousness since ancient times. This research spans multiple knowledge areas, from theology to neurobiology, including areas such as quantum physics and computer science.

Although the curiosity of man about his own consciousness is as old as humanity, Machine Consciousness is a very young research filed and subject of great controversy. Therefore, carrying out a research work in this field not only requires a fundamental positioning, but also the capability of continuous revision, about the nature of such an intricate phenomenon as consciousness.

Taking the challenge of the scientific study of consciousness implies reviewing the basis of the scientific method. The study of the purely subjective and immaterial seems to be unreachable for any classical empirical approach. The study of consciousness is intimately related with the so-called Mind-Body problem, which historically has been addressed as a duality. However, modern science is trying to explain consciousness without the need of a separated – independent of the body – mental entity (Bunge 2002).

The problem of explaining the apparent duality of body and mind is known as the explanatory gap (Levine 1983). Specifically, the term "hard problem" is used in consciousness studies to refer to the search of an explanation for subjective experience. In contrast, the "easy problems" are related with cognitive skills like attention, which are easier to study scientifically (Chalmers 1995). Addressing the problem of consciousness from the perspective of engineering, typically separated from philosophical issues, seem to enlarge the size of this quest. Nevertheless, this thesis aims at addressing the problem of Machine Consciousness from a pragmatic point of

view, trying to provide new solutions and knowledge through the use of computational models. As argued by Seth (Seth 2009), the application of synthetic models of the mechanisms that can potentially produce consciousness might provide valuable information about the development of conscious minds.

The intrinsic difficulty in the study of consciousness comes primarily from two sources: the multidisciplinary approach required for its analysis and its complex character. These features of the scientific approach to consciousness have led to a generalized confusion regarding the real meaning of the typical terms used by researchers. Given this situation, a set of working hypotheses need to be established in order to both limit the scope of the thesis and clarify the concrete meaning of the terms used in the manuscript. Consciousness has been studied from disciplines so disparate as quantum physics, neurobiology, neurophysiology, psychology, philosophy, mathematics, computer science, etc. That is the reason why there exists confusion regarding basic concepts, like for instance *representation*, which can be assigned quite different meanings depending on the specific academic environment in where it is used. This issue usually comes together with the problem of the proper definition of consciousness. As indicated by Block (Block 2001), Consciousness is a "mongrel" concept, which comes from the breeding of a number of concepts referring to different phenomena. Dealing with all these different concepts as if they were just one is a typical but problematic trend in the study of consciousness.

As mentioned above, the definition of a set of working hypotheses about the possibility of implementing or modeling Machine Consciousness can be useful as the main tool to circumvent these sorts of problems (the proposed set of working hypotheses are specified in Section 1.1). Furthermore, the specific definition of these hypotheses will permit to draw conclusions about the validity of the proposed assumptions at the end of the thesis. It is important to remark that some of the hypotheses put forward in this work are the subject of great controversy, mainly because of the great difficulty associated to their verification. Therefore, it is expected that not all the proposed hypotheses can be irrefutably tested. However, the research conducted in this thesis aims at shedding some light on key aspects related with Machine Consciousness. This will contribute to the advancement of a field that has been historically excluded from scientific research.

As it is the case for other abstract terms like intelligence, the concept of consciousness may comprise a great number of meanings and perspectives. The position adopted in this thesis is based on the initial consideration of functional aspects of consciousness, subsequently moving onto implications related with the phenomenal dimension (e.g. conscious sensation and subjective experience).

Biological inspiration is common in many areas of Artificial Intelligence research, and so it is the case in the field of Machine Consciousness. The sources of biological inspiration can be considered at different levels of description: physical, chemical, biochemical, biological, physiological, cognitive, etc. The research conducted in the present thesis is centered on the functional level of description typically associated with cognitive science. Therefore, the main tasks are to analyze, understand, model, and implement cognitive functions associated with consciousness. It is expected that the analysis of these cognitive functions, the design of an effective interaction between them, and their implementation in artificial agents will contribute to the scientific understanding of such an elusive phenomenon as consciousness.

1.1 Working hypotheses

Given the existing controversy in the scientific study of consciousness, where many fundamental points are still intensely debated, it is required to clearly establish which is the initial position adopted in the present doctoral thesis. Specifically, it is required to establish a set of working hypotheses in order to clearly delimit the research scope and unequivocally define the objectives of the thesis. Generally, any approach to the study of mind implies the advocacy to certain hypotheses, which after the required research can be finally validated or refuted. In the following the fundamental points of controversy are specified (they are discussed in detail in Chapter 2) and the corresponding working hypotheses adopted initially in this thesis are described.

The position expressed in the following working hypotheses may not correspond with the true nature of consciousness. However, as this doctoral thesis is part of the efforts to unveil such nature, obtained conclusions shall be used to validate or refute some of the proposed hypotheses.

1.1.1 Phenomenal consciousness can be studied in artificial systems

A distinction is frequently made between functional and phenomenal aspects in the context of consciousness studies (Block 1995). Access or functional consciousness (A-Consciousness) refers to the mechanism of the mind that makes possible the access to contents that are explicitly used in processes like reasoning, speech, and decision making. According to the global access hypothesis (Baars 2002), there exists a mechanism in the brain that permits the selection of a very limited number of mental contents to be sequentially processed – thus forming a conscious thread – while at the same time the vast majority of information processing takes place concurrently and unconsciously. Phenomenal consciousness (P-Consciousness) refers to the subjective experience *per se*, the set of sensations that human beings experiment when we are conscious of something. *Qualia* (plural of *quale*) is the name used to refer to the manifestation of such subjective conscious experiences.

According to Dennett (1988), the term qualia is *"an unfamiliar term for something that could not be more familiar to each of us: the ways things seems to us"*. The expression *"what is it like"* is commonly used to define qualia. This term was coined by Nagel, who used it to refer to phenomenal states and the characteristic subjective point of view of conscious beings (Nagel 1974). The redness that we consciously perceive when we look to something red, or the painfulness of the pain sensation are typical examples of *what is it like to see* red or *what is it like to feel* pain.

The specific meaning of the concept quale is still controversial. Qualia are usually assigned several specific properties, being the following generally accepted:

- Qualia are *ineffable*. In other words, they cannot be reported or acquired by any means different of direct subjective experience. The third-person observation of a quale is by definition not possible. In fact, although humans commonly report about their conscious experience, saying for instance "I see the color red", there

is no possibility of direct observation of the redness qualia that the subject is experimenting subjectively.

- From the former property derives that qualia are eminently *private*. That is to say, qualia are completely subjective, i.e. qualia from different persons cannot be directly compared. When confronted with the same red visual stimulus, two different persons might have very different subjective experiences; however, both would say "I see the color red" (Shoemaker 1982). Therefore, verbal report of qualia is useless in terms of accurate comparison of the subjective experience of individuals.
- Qualia have *structure*. There exist similarities and differences between sensations corresponding to different sensory modalities. For instance, visual experiences have some qualities in common that are not shared by auditory experiences (people do not have visual experiences of sound or hear colors, except in cases of synesthesia (Cytowic 2002, Ramachandran, Hubbard 2001)).

The concept of "artificial quale" is defined in the specific context of this thesis (see Chapter 6) characterized as the specification of the content of conscious experience simulated or modeled in an artificial system. Practically all computational models of consciousness are exclusively based on A-Consciousness. This is because functional aspects of the access mechanism are easily identified, while the functional role of P-Consciousness is not clear and widely debated (Dennett 1988, Jackson 1982, Gray 2003, O'Regan 2010).

Despite of conceptual and scientific difficulties associated with the study of phenomenal aspects (see Section 2.1), rigorous research in the domain of Machine Consciousness is expected to cover all dimensions of consciousness. In the present thesis, a Synthetic Phenomenology (SP) approach will be adopted (Chrisley 2009). This SP approach will allow for the exploration of the nature of first-person experience through the use of computational models. The functional point of view will be combined with the phenomenology perspective thanks to the specification of the contents associated with phenomenal states modeled or possessed by artificial systems.

In short, as for the distinction between A-Consciousness and P-Consciousness, the following working hypothesis has been adopted: *the Machine Consciousness research field does not only include purely functional cognitive aspects associated with consciousness, like attention, but also the phenomenal dimension. Taking an approach based on Synthetic Phenomenology will allow for the study of the content of conscious experience through the use of computational models.*

1.1.2 The functional role of consciousness: integration and adaptation

In relation with the former hypothesis, in which P-Consciousness is stressed as a possible subject of research, this hypothesis deals with the functional role of consciousness. This is a controversial point as some authors have argued that certain mental states are caused by physical brain mechanisms, but these states themselves have no causal effect in the world (epiphenomenalism) (Huxley 1898). According to this position, mental states would exist as mere lateral effects or derivate byproducts of the brain physical processes. These states, like for instance the sensation of pain, would be epiphenomena (they could not be the origin of any causal chain). In other words,

whether or not the organism experiments the sensation of pain, the behaviour would be exactly the same (in the case of not consciously feeling the pain there would be a representation of pain, but not conveying any associated phenomenal state). Epiphenomenalism argues that there are enough physical causes to explain behavior, and therefore there is no need to claim that conscious experience is the cause of anything, qualia are simply considered lateral effects.

The working hypothesis assumed in this thesis about the functionality of consciousness is radically opposed to the epiphenomenal vision. Apart from the functionality corresponding to cognitive skills like attention, anticipation, decision making, etc., the existence of conscious states is considered to provide a global function of cognitive integration, coherency, and unity (Tononi 2004, Seth et al. 2006).

As indicated by Haikonen (2009), humans perceive through qualia. If there are no qualia, there is no consciousness. Therefore, the generation of qualia cannot be ignored and should be regarded a key issue in the Machine Consciousness research. Additionally, it seems reasonable to suppose that consciousness has appeared thanks to evolution (Dennett 1991). The mechanisms that generate subjective experiences and the perception of the unity of the self – separated from the external world – have been selected by evolution as they are advantages for survival in complex environments.

In short, *possessing phenomenal states, and experiencing the world through qualia, is an evolutionary advantage that allows the subject to deal effectively with a great amount of information and generates effectively adaptive behaviours in unstructured environments*.

1.1.3 Consciousness is a process, not a property of the matter

There are essentially two groups of theories that try to explain P-Consciousness from very different perspectives. On one hand, theories based on the substrate try to explain consciousness in terms of the intrinsic properties of the physical medium in which the representations of the world are generated; on the other hand, theories based on information processing try to explain consciousness in terms of computational functions and the corresponding relational properties between represented contents (Atkinson, Thomas & Cleeremans 2000).

A classical example of a theory based in the substrate is the one proposed by Hameroff and Penrose (1996), known as the Orch-OR (Orchestrated Objective Reduction) theory. The Orch-OR model argues that phenomenal states are produced thanks to quantum mechanics effect that take place at a supramolecular level, in the microtubules of the neurons. According to these sorts of theories, producing consciousness in machines would require the same physic mechanisms, i.e. to use a substrate that supports the same mechanisms at a quantum level. Therefore, according to this particular theory, a quantum computer would be required in order to reproduce the phenomenon of orchestrated objective reduction, which would sustain the phenomenal states of the machine.

An example of theory based on information processing if the Global Workspace Theory (GWT) proposed by Baars (1988). In this case it is argued that consciousness is produced thanks to a specific way of computing. Therefore, consciousness production is independent of the substrate using to physically perform the computation.

Consequently, according to the GWT, any computer, independently of the underlying technology, would be able to produce consciousness as long as the representations of the world are manipulated using the adequate processes.

In the present thesis the following working hypothesis is supported: *the nature of consciousness is like that of a software process. It is immaterial but possible to be generated by any physical substrate able to perform the required computation. No specific substrate is required in order to build a mind that observes and subjectively experiences the world through qualia*.

1.1.4 Conscious processing is just the tip of the iceberg

Looking at the former hypotheses it can be inferred that consciousness appeared in nature as an effective form of adaptation to very complex and unstructured environments. Analogously, there seems to be evidence that consciousness in produced by great complexity structures in the nervous systems of mammals (Seth, Baars & Edelman 2005). In light of this, researchers tend to support the idea that an equivalent level of complexity is required in order to generate consciousness in artificial systems.

Former claim may be characterized according to two different criteria. On one hand, making the necessary distinction between conscious and unconscious processing; on the other hand, distinguishing between brute computational power and cognitive power. According to Moravec's paradox, conscious thought requires relatively little computation, while unconscious sensorimotor skills require great computational power: *"it is comparatively easy to make computers exhibit adult level performance on intelligence tests or playing checkers, and difficult or impossible to give them the skills of a one-year-old when it comes to perception and mobility"* (Moravec 1990).

A conceptual separation between conscious and unconscious processing can be drawn from the former paradox. However, according to modern cognitive theories, it is more accurate to consider consciousness as the tip of the iceberg of mental activity, where most of the processing takes place unconsciously. Therefore, the real problem consists on determining how conscious processing is produced out of a huge amount of unconscious mental activity. In other words, finding out how the information needs to be processed is more important than solving the problem of computing power requirements.

Taking into account the former considerations, the research direction on Machine Consciousness developed in this thesis in centered on the improvement of the design of cognitive architectures. It is assumed that *the most important lack of current artificial cognitive architectures is not their constrained computational power (which could be enhanced using parallel computing or distributed systems), but their limited cognitive computation capabilities. The root cause of this limitation lies in the very design of these systems*.

1.1.5 Consciousness can be studied scientifically

One of the main characteristics of consciousness, specifically in the domain of phenomenal consciousness (P-Consciousness), is that it is eminently private. The

painfulness of pain, the *what is it like* to be oneself, and every subjective experience in general, can only be experienced by the first-person observer. However, the scientific method calls for the observation of objective data, i.e. third-person observations.

Chalmers (2004) describes the science of consciousness as a process in which third-person observations – brain processes, behaviour, etc. – are associated with conscious experience first-person observations. The private nature of data from subjective experience is an obvious drawback when it comes to applying the scientific method. In most scientific areas, data are available intersubjectively; that is to say, data are available equally to a great number of observers. However, in the case of consciousness, subjective experiences are only accessible to the very subjects. For the rest of the observers, data about the experiences lived by the subject can only be obtained indirectly, using for instance the observation of behaviour or the inspection of brain processes at the neurophysiologic level.

Given that an ultimate theory of consciousness does not exist, obtaining objective observations from physical inspection is not strictly possible. Nevertheless, neural correlated of conscious states are being investigated (Rees, Kreiman & Koch 2002, Seth 2009). Usually, in humans accurate verbal report is an effective method of indirect observation (Seth, Baars & Edelman 2005).

In this thesis, the problem of first-person observation is addressed adopting the *heterophenomenology* approach proposed by Dennett (1988, 2003), which consists on an extension of objective science to include consciousness, but keeping the experimental method. Heterophenomenology (phenomenology of other different from oneself) proposes to use verbal report (and other sorts of report) as third-person observations that provide the researcher with valuable information about subject's beliefs regarding his/her own conscious experience.

In the case of artificial systems the first-person observation problem can also be addressed applying the concept of heterophenomenology. Although the machine under analysis is not endowed with an accurate verbal report mechanism, alternative ways of conscious experience specification could be considered. The field known as Synthetic Phenomenology provides a suitable framework for the study of subjective experience (Chrisley 2009). Synthetic Phenomenology considers the scientific validity of subjective experience specifications generated by a machine. What is actually analyzed is the value of the very specification produced by the machine, independently of whether or not the machine is considered to actually possess phenomenal states associated with the specified content. Additionally, this approach can be combined with third-person observations obtained from the inner inspection of the working artificial system.

In short, in this doctoral thesis the following hypothesis is supported: *Machine Consciousness can be studied scientifically thanks to the combination of: (1) purely objective observation of physical parameters of behaviour and inner working of the system, and (2) the heterophenomenological observation based on the specifications generated by the very artificial system regarding its own conscious experiences*.

1.1.6 The level of consciousness of an artificial system can be measured

Measuring consciousness is also a controversial topic. Although it is commonly accepted that consciousness appears as a continuum, there is no consensus about how to measure and characterize the level of consciousness of a system. The main reason for this lack of consensus is rooted in the fact that the formulation of an ultimate theory of consciousness is still a challenge for science. There exist scales that are commonly used for the clinical diagnosis of human consciousness (Jennett 2002, Giacino, Kalmar & Whyte 2004); however, these measures cannot be applied in the field of Machine Consciousness because they are based on physiological parameters.

The definition of similar mechanisms to those used in the clinical practice, but applied to the domain of artificial systems, would be of great help in order to effectively assess the progress in Machine Consciousness. There exist some proposals for measuring consciousness in machines (see Section 2.3); however, all of them have significant problems of application and scientific validity (Seth et al. 2006). Furthermore, none of the extant proposals allows for the calculation of a quantitative measure associated with the cognitive capabilities of the evaluated system.

Attending to the distinction between access and phenomenal consciousness, it seems logical to think that specific measures could be designed for each of these aspects. In fact, there exist proposals for the specific measurement of P-Consciousness, like Φ, which is based on the Information Integration Theory (Tononi 2004). However, while measures like Φ are exclusively focused on the hard problem, this thesis focuses on a pragmatic approach that requires the measurement of cognitive abilities associated with the so-called easy problems of consciousness (Chalmers 1995). In the realm of Machine Consciousness, and more specifically from an engineering perspective, it does not make sense to try to solve the hard problem of consciousness if the easy problems related with the cognitive abilities are not addressed in the first place. Additionally, having a mechanism for measuring the degree of resolution of the easy problems might shed some light about which is the most adequate characterization for the hard problem of consciousness.

In the framework of the present thesis, the hypothesis is supported that *it is possible to define a sequence of levels of development of consciousness based on the cognitive abilities demonstrated by an individual. These levels can be established as a reference to evaluate and compare different Machine Consciousness models and implementations.*

1.2 Objectives

The scientific study of consciousness, and more specifically the field of Machine Consciousness, is in an early stage of development, not having achieved yet a significant level of maturity that permits the existence of fundamental principles commonly agreed by the scientific community. Therefore, the main objectives of this thesis (described in detail in Chapter 3) are focused on the testing of the working hypotheses described above (Section 1.1).

In order to carry out a set of experiments designed to test the working hypothesis, some specific tools are required. Such tools have been developed as part of this doctoral

thesis research work. In particular, three interrelated research lines can be distinguished and the corresponding tools specified:

- **Design of an artificial cognitive architecture inspired by consciousness.** An artificial cognitive architecture based on a computational model of consciousness is to be designed and implemented. The architecture should be flexible enough to integrate several complementary mechanisms inspired by different theories of consciousness.
- **Design of a Synthetic Phenomenology model.** The former cognitive architecture is to be used for the design of a model of phenomenal consciousness. Such model and the corresponding implementation are expected to shed light on the problem of conscious experience generation in artificial systems.
- **Design of a scale for measuring the level of consciousness in artificial systems.** A theoretical framework and associated tools are to be defined and implemented in order to effectively analyze and evaluate the level of cognitive development of consciousness in artificial systems.

In short, although the major contribution of this thesis lies in the testing of the working hypothesis described above, a set of associated tools have been developed as platforms for experimentation. CERA-CRANIUM, an artificial cognitive architecture inspired by several theories of consciousness have been designed and implemented (see Chapter 4). *ConsScale*, a scale inspired by evolution of consciousness has been created in order to measure the cognitive development of both natural and artificial creatures (see Chapter 5). Finally, a Synthetic Phenomenology model has been defined in order to study the generation of artificial qualia using the CERA-CRANIUM architecture (see Chapter 6).

1.3 Structure of the thesis report

The structure of this report is described in this section in order to provide the reader with a clear outlook of the contents of this doctoral thesis.

Chapter 2 contains a comprehensive review of significant related work. This chapter, dedicated to the state of the art survey, begins with an introduction to the scientific study of consciousness and associated problems. Then, Machine Consciousness is introduced and current applications and models are discussed. Finally, this chapter includes a review of current techniques for measuring consciousness.

Chapter 3 provides a detailed description of the objectives of this doctoral thesis, specifying the deliverables, concrete direction of the research carried out, and expected results.

Chapter 4 is dedicated to the MC^3 computational model and derived CERA-CRANIUM cognitive architecture. MC^3 is a computational model of consciousness developed as part of this thesis work, and CERA-CRANIUM is a partial implementation of that model. This chapter includes the design principles and the implementation details of the cognitive architecture.

Chapter 5 is dedicated to *ConsScale*, the novel scale for measuring artificial consciousness that has been designed and developed as part of this thesis work. This

chapter includes the description of consciousness levels, associated tools, and evaluation mechanisms required for the practical application of the scale.

Chapter 6 explores the possibility of the specification of conscious experience contents. The proposed model for the generation of artificial qualia is described here, and the working hypotheses related with Synthetic Phenomenology are also discussed.

Chapter 7 describes the evaluation methods, tools, and experiments that have been carried out using the cognitive architecture, the Synthetic Phenomenology model, and the scale. This chapter includes results and discussion of experiments performed in the domains of mobile robotics and computer games.

Finally, Chapter 8 provides a summary of the main conclusions drawn during the development of this doctoral thesis. Additionally, main future work initiatives are discussed.

2 Estado del arte

El principal dominio científico-tecnológico que se aborda en la presente tesis es la Conciencia Artificial. Concretamente, se plantea el análisis, la evaluación y el diseño de modelos de Conciencia Artificial que potencialmente sean capaces de dotar a un agente autónomo de capacidades cognitivas superiores. Por lo tanto, en este capítulo se presentan los últimos y más importantes avances científicos relacionados con este campo de investigación.

Dado que la principal fuente de conocimiento sobre la conciencia proviene de su estudio en el reino animal y especialmente en el caso de los humanos, este capítulo comienza con un repaso de las principales líneas de investigación que se realizan actualmente con el objetivo de comprender la conciencia humana.

Una de las características principales del estudio de la conciencia es la confluencia de diversas disciplinas científicas, por lo que a menudo se hará referencia a trabajos procedentes de diversos campos que en mayor o menor medida están relacionados con la neurociencia, como por ejemplo, la psicología cognitiva, neurofisiología, filosofía de la mente, inteligencia artificial, etc.

Con el objetivo de ofrecer una panorámica completa del estado del arte en el campo de la Conciencia Artificial, se analizan en primer lugar las principales teorías que tratan de explicar cómo se produce la conciencia en los humanos (Apartado 2.1). Posteriormente, se repasan los principales avances conseguidos hasta la fecha en el ámbito de la Conciencia Artificial (Apartado 2.2). Finalmente, se estudian los principales métodos existentes actualmente para la medición del nivel de conciencia (Apartado 2.3).

2.1 El estudio científico de la conciencia

Desde los inicios de la ciencia el hombre se ha preguntado qué es y cómo funciona la conciencia. La principal respuesta que tenemos hoy en día a estas preguntas es que todavía no sabemos cuáles son exactamente los mecanismos biológicos que dan lugar a este fenómeno. Sin embargo, desde disciplinas como la psicología, la neurología o la filosofía se han planteado multitud de teorías e hipótesis acerca de la naturaleza de la dimensión consciente de los humanos y cuál es su origen y funcionamiento.

Actualmente es obvio pensar, que al igual que la inteligencia, la conciencia de la que están dotados los seres humanos tiene su origen en el cerebro. Diversas disciplinas científicas estudian el sistema nervioso central desde diferentes perspectivas con el

objetivo de entender cómo funciona realmente el cerebro humano. A este conjunto de disciplinas que estudian a diferentes niveles los procesos que tienen lugar en el cerebro humano se le denomina *neurociencia* (o *neurociencias*).

El estudio de la conciencia es uno de los mayores retos de la ciencia. A menudo ha sido un tema esquivo y ausente en los círculos puramente científicos. Sin embargo, con los últimos avances en neurociencia, en las últimas décadas el problema de la conciencia ha atraído el interés de científicos de diversas áreas. En cualquier caso, descifrar los misterios de la conciencia sigue siendo un auténtico desafío. Como dijo Leibniz hace tres siglos, *"si pudiéramos aumentar el tamaño del cerebro hasta hacerlo tan grande como una gran fábrica, podríamos entrar dentro y examinarlo. Pero aún así no encontraríamos la conciencia"* (Blackburn 2004). Ni siquiera podríamos apreciar cuál es el mecanismo que da origen a la voluntad humana. Simplemente observaríamos un incontable número de sucesos que tienen lugar en miles de millones de neuronas y las conexiones que las unen. Cada uno de estos sucesos seguiría inexorablemente las leyes físicas, no observaríamos ningún hecho particular que nos indicase la producción del fenómeno de la conciencia.

Desde un punto de vista puramente biológico, hoy en día la mayoría de científicos argumentan que la conciencia e inteligencia humanas han aparecido como una ventaja evolutiva, siendo a la vez el origen y el soporte del tejido social y cultural de las civilizaciones (Gavrilets, Vose 2006).

A menudo, la conciencia se define basándose en la relación existente en los humanos entre los siguientes procesos mentales: atención, razonamiento, reconocimiento y comportamiento (Grossberg 2003). Es decir, un ser consciente presenta la capacidad de atención hacia una cosa, y puede pensar acerca de ella, qué es, cómo es, por qué es así, etc., con el objetivo de reconocerla. Una vez que el objeto (o suceso) se identifica, el sujeto lo ha reconocido y entonces decide qué quiere hacer con él. Se dice que las representaciones mentales sobre las que operan estos mecanismos constituyen los contenidos conscientes de la mente humana. Como se verá en el siguiente apartado, el paradigma de la Conciencia Artificial se inspira en estos procesos observados en los humanos (y en otros mamíferos superiores) con el objetivo de conseguir sistemas artificiales que presenten capacidades y funcionalidades análogas a las naturales.

El diseño de un modelo de Conciencia Artificial implicaría un conocimiento extenso y preciso de las teorías más consagradas que explican el origen de la conciencia en los humanos. Sin embargo, como ya se ha mencionado anteriormente, no existe hoy en día un cuerpo de conocimiento establecido que explique definitivamente qué es y cómo surge la conciencia. En lugar de esa situación ideal, se dispone de un amplio conjunto de modelos y teorías sin confirmación o refutación absoluta. Es más, a menudo no existe una coherencia entre las hipótesis planteadas por diferentes autores. Incluso se mezclan diferentes niveles de complejidad y conocimiento de forma confusa, lo que dificulta en gran medida el entendimiento y la comparación de las hipótesis existentes.

En la presente tesis se ha pretendido identificar los mecanismos clave que se conocen en la actualidad sobre el funcionamiento de la conciencia en los humanos, con el objetivo de extrapolarlos a modelos computacionales que puedan ser de utilidad en el ámbito de la Conciencia Artificial. Es decir, que puedan servir tanto para ofrecer retroalimentación en el estudio de la propia conciencia humana como para contribuir al diseño de máquinas más eficientes.

2.1.1 Tipos de conciencia

Con el objetivo de establecer un contexto claro para los trabajos de investigación que se realizan en la presente tesis, evitando la composición de diferentes aspectos de la conciencia que a menudo se confunden y mezclan indiscriminadamente, se tomará como referencia la distinción establecida por Block acerca de las diferentes dimensiones de la conciencia (Block 1995). Block distingue entre:

- Conciencia Fenomenológica (*Conciencia P*).
- Conciencia de Acceso (*Conciencia A*).
- Conciencia de Monitorización (*Conciencia M*).
- Autoconciencia (*Conciencia S*).

La conciencia fenomenológica o Conciencia P se refiere al conjunto de experiencias subjetivas (*qualia*) que un ser siente por el hecho de ser consciente. La conciencia de acceso o Conciencia A es el mecanismo que posibilita el acceso a los contenidos de la mente para su uso en el razonamiento, el habla y los actos volitivos (la toma de decisiones consciente). La conciencia de monitorización o Conciencia M engloba los procesos de percepción interna o introspección. Por último, la autoconciencia o Conciencia S es la capacidad de auto-reconocerse y razonar acerca del yo reconocido.

Por lo tanto, se distinguen diversas formas de ver la conciencia. Por un lado un sujeto es consciente cuando presta atención a un objeto del exterior, conociéndolo y comprendiéndolo mientras éste es foco de su atención (Conciencia A); por otro lado, el sujeto puede percibir y sentir cualidades específicas – el cómo es sentir o percibir un determinado estímulo – por ejemplo, la rojez del rojo (Conciencia P). Adicionalmente, el sujeto también puede percibir hasta cierto punto su propio estado interior (Conciencia M), e incluso ser consciente de su propio *yo* (Conciencia S).

Si bien es cierto que el cuarteto propuesto por Block es muy útil para analizar tanto las teorías actuales sobre la conciencia como los posibles modelos de Conciencia Artificial propuestos, es también preceptivo conocer otras clasificaciones y definiciones que se utilizan a menudo en el estudio de la conciencia. Una distinción típica es la que se realiza entre la conciencia en relación a un sujeto (*conciencia de la criatura*) y en relación a un estado (*conciencia de un estado*). No es lo mismo decir que una persona (u otro organismo) es consciente que decir que uno de los estados mentales que posee esa criatura es consciente (Manson 2000). En general se dice que un sujeto es consciente porque es capaz de mantener estados mentales conscientes. En otras palabras, los seres conscientes tienen tanto estados mentales conscientes como estados mentales inconscientes.

Dentro de lo que se denomina conciencia de criatura también se suele distinguir entre *conciencia intransitiva* y *conciencia transitiva*. Decir que un organismo tiene conciencia intransitiva significa que el sujeto está en vigilia y alerta, en contraposición a estar dormido o en estado comatoso. Sin embargo, decir que un organismo es consciente de algo (conciencia transitiva) significa que está percibiendo ese algo de forma explícita y experimenta estados fenomenológicos asociados a los contenidos percibidos (qualia).

2.1.2 Principales áreas en el estudio científico de la conciencia

Uno de los problemas de base existentes en el estudio científico de la conciencia es la imposibilidad de aplicar el método científico tal y como lo conocemos en otros dominios, pues la experimentación sobre la conciencia es subjetiva y los resultados sólo se pueden obtener mediante la introspección o la comunicación verbal. Dennett acuñó el término *heterofenomenología* para referirse a una modalidad del método científico adaptada al estudio de la conciencia. El método heterofenomenológico se basa en combinar la información que proporciona el sujeto con las evidencias observables por terceras personas en cuanto a su estado mental (Dennett 1991). Este tipo de enfoques filosóficos, junto a los avances logrados en las últimas décadas en el terreno de la neurociencia, han dado lugar a lo que se puede denominar el método actual de estudio científico de la conciencia.

Desde que Santiago Ramón y Cajal proporcionara a la comunidad científica internacional las bases de la neurociencia moderna, se ha realizado una búsqueda incansable por parte de prestigiosos científicos de los llamados Correlatos Neuronales de la Conciencia o CNC (Crick 1994, Crick & Koch 2003, Edelman 1992). La identificación exacta de los procesos particulares que dan lugar a la conciencia y sus propiedades específicas implicaría la comprensión de los mecanismos subyacentes. Aunque existen diversos candidatos defendidos por numerosos científicos, parece claro que la búsqueda de estos correlatos neuronales no ha llegado a su fin. Análogamente, en el campo de la Conciencia Artificial, se podría intentar caracterizar la conciencia en función de sus correlatos computacionales (Cleeremans 2005). Algunos aspectos importantes en este ámbito serían la diferenciación entre procesos conscientes e inconscientes y sus características asociadas, ya sean relativas al tipo de conocimiento gestionado, el tipo de procesamiento realizado, etc.

Otro aspecto importante de la conciencia es el relativo a las emociones. Se sabe que los seres conscientes reaccionan ante cualquier estímulo de forma racional, pero también existe una respuesta emocional. El componente emocional no se puede, por tanto, obviar en el estudio de la conciencia, y por ende en los sistemas de Conciencia Artificial. Las emociones afectan a los mecanismos de atención, abstracción y de producción y percepción del lenguaje (Ciompi 2003). Además, las emociones proporcionan al sujeto una sucinta evaluación sobre la consecución de sus metas (Marina 2002, Damasio 1995). Por ejemplo, una persona siente alegría cuando las cosas salen de acuerdo a lo que ha planeado. Sin embargo, cuando aparecen problemas inesperados surge el enfado o la desesperación. Existen modelos psicológicos que relacionan las emociones con el comportamiento y algunos aspectos de la conciencia. Como se verá más adelante, estos modelos se pueden aplicar (evitando inicialmente la dimensión fenomenológica) en trabajos relacionados con la Conciencia Artificial.

Es importante destacar que la inspiración en las teorías de las emociones que se ha adoptado en la presente tesis doctoral se enmarca siempre en teorías más ámplias que tratan de explicar la conciencia. Es decir, no se pretende aplicar un modelo de emociones per se, sino que se consideran las emociones como un aspecto clave enmarcado en el ámbito de la conciencia.

Aunque hoy en día es comúnmente aceptado que el estudio de la conciencia debe ser muy interdisciplinar, típicamente la investigación se ha realizado de forma relativamente aislada desde diversos paradigmas o áreas de conocimiento. Esta es, probablemente, una de las principales causas por las que existen actualmente diversas

teorías parciales de la conciencia; que tratan de explicar, cada una desde su disciplina de conocimiento, ciertos aspectos relacionados con la conciencia. Con el objetivo de ofrecer una panorámica completa de las teorías más relevantes planteadas hasta la fecha, se describen a continuación los principales intentos de explicar la conciencia que se han realizado desde los siguientes campos de investigación:

- Neurobiología (Apartado 2.1.3).
- Psicología cognitiva y filosofía de la mente (Apartado 2.1.4).
- Física (Apartado 2.1.5).

Actualmente, la denominada *neurociencia cognitiva* se ha erigido como un campo eminentemente multidisciplinar que trata de explicar cómo aparece la cognición en el sistema nervioso combinando niveles de descripción tan variados como los correspondientes a la neurobiología y a la psicología cognitiva. Además, nuevos campos como la neurociencia computacional y la Conciencia Artificial también están comenzando a aportar nuevo conocimiento sobre la conciencia. En definitiva, se trata de poner en común los resultados obtenidos desde las diversas disciplinas con el objetivo de lograr nuevos avances en cuanto a la comprensión de la conciencia humana.

Las hipótesis de trabajo planteadas en esta tesis doctoral (Apartado 1.1) son un ejemplo ilustrativo de algunos de los temas controvertidos para los que la ciencia aún no tiene una respuesta firme. La investigación interdisciplinar realizada en esta tesis, y en general en el campo de la Conciencia Artificial, puede proporcionar nuevas visiones que permitan una mayor comprensión de un fenómeno tan esquivo para la ciencia como es la conciencia. Los modelos propuestos en la presente tesis doctoral se centran en el nivel de descripción proporcionado por la psicología cognitiva y la filosofía de la mente (Apartado 2.1.4). Aunque las teorías físicas y neurobiológicas también pueden proporcionar inspiración para la realización de modelos computacionales (tal y como se describe en el Apartado 2.2.2), el enfoque elegido en esta tesis está orientado a la construcción de arquitecturas cognitivas artificiales. Este tipo de arquitecturas permite abordar problemas más complejos y la realización de aplicaciones prácticas empleando diversidad de agentes artificiales.

2.1.3 Estudio neurobiológico de la conciencia

Para poder diseñar un modelo de conciencia aplicable a un sistema artificial se necesita un modelo que represente fielmente el mejor ejemplo de conciencia que se conoce: el ser humano. Cualquier modelo que explique en cierta medida el funcionamiento de la conciencia en los seres humanos debe basarse en los estudios científicos realizados al respecto. A continuación se repasan los fundamentos y las principales teorías que tratan de explicar la conciencia desde el punto de vista neurobiológico.

2.1.3.1 Fundamentos del estudio neurobiológico de la conciencia

Aunque históricamente el estudio científico de la conciencia humana estuvo de alguna forma relegado, desde que el premio Nobel Francis Crick (co-descubridor de la

estructura del ADN) proclamó que la conciencia debía ser abordada desde una perspectiva científica, y que tal perspectiva debía ser liderada por la neurociencia (Crick 1994), empezaron a salir a la luz numerosas investigaciones científicas orientadas a explicar cómo se produce la conciencia en el cerebro. Según Crick, *"nuestros gozos y nuestras penas, nuestros recuerdos y nuestras ambiciones, nuestro sentido de identidad personal y de libre albedrío, no son en realidad sino la conducta de vastos ensamblajes de neuronas y de sus moléculas asociadas"*.

Existen otras corrientes de pensamiento que no están de acuerdo con la aserción de Crick (enfoque materialista), y que indican que se necesita algo más que tejido nervioso para dar lugar a la conciencia (enfoque dualista). En este sentido, es interesante destacar que durante las últimas décadas los avances científicos en el campo de la neurología han propiciado la aparición de numerosas teorías acerca de la base neurobiológica de la conciencia (Atkinson, Thomas & Cleeremans 2000).

El estudio científico de la conciencia se realiza a distintos niveles. A nivel neurológico el objetivo es la identificación y comprensión de los procesos neuronales que producen los pensamientos conscientes (Crick & Koch 2003). A este nivel, la analogía correspondiente en el campo de la Conciencia Artificial sería la identificación de la arquitectura necesaria para producir conciencia y la determinación del tipo de procesamiento de la información necesario. En el plano de las ciencias cognitivas, el reto se enmarca en la identificación de las funcionalidades clave de la conciencia y la sinergia existente entre las mismas.

Dentro del ámbito de los estudios neurobiológicos, los trabajos de investigación suelen agruparse en los niveles molecular, neuronal y de redes o grupos de neuronas. A nivel molecular se llegan a estudiar incluso las interacciones cuánticas entre los elementos intracelulares como los microtúbulos (Hameroff, Penrose 1996). A nivel neuronal, toma especial relevancia el funcionamiento concreto de cada célula cerebral y los procesos de sinapsis. Por último, el estudio de organizaciones de neuronas se centra tanto en los pequeños grupos funcionales de neuronas, como en las columnas corticales o grandes asambleas de neuronas como las de percepción visual (Crick 1994).

Según Crick y Koch (2003), en un momento dado existen procesos cerebrales que se correlacionan con la conciencia, denominados Correlatos Neuronales de la Conciencia (CNC). Al mismo tiempo existen otros procesos que no tienen relación directa con la misma. Las preguntas que la neurociencia debe responder con respecto a este tema son las siguientes:

- ¿Cuál es la diferencia entre los procesos mentales que se correlacionan con la conciencia y los que no?
- ¿Las neuronas implicadas en los procesos conscientes son de algún tipo en particular?
- ¿Hay algo diferente en la forma en que se conectan?
- ¿Existe alguna diferencia en el modo de activación?

Cualquier avance científico que ayude a desvelar las respuestas a estas preguntas podría potencialmente ser aplicado al campo de la Conciencia Artificial. El problema de la búsqueda de los CNC es que la propia definición de conciencia es tenue e incluso cambiante dependiendo del ámbito de investigación. Se deben establecer una serie de criterios para estudiar la localización de la conciencia en el cerebro, de forma que se puedan obtener conclusiones factibles sobre las correlaciones correspondientes. Sin

embargo, la elección equivocada de estos criterios puede dar lugar a una conclusión errónea sobre las zonas del cerebro que producen la conciencia. Es más, la propia búsqueda de los CNC tal y como la planteaban Crick y Koch podría ser en sí un error conceptual. De hecho, otros investigadores, como O'Regan (2007), sugieren que la conciencia fenomenológica es un proceso análogo al de la vida (ver Apartado 2.1.4.4), no siendo posible definir éste como una propiedad estructural o funcional del cerebro, si no como un fenómeno que se manifiesta durante la interacción del sujeto con su entorno.

En el sistema nervioso central humano existen multitud de redes de neuronas que tienen una determinada función. Por ejemplo, los nervios ópticos se encargan de transmitir y procesar la información visual proveniente de los ojos. Teniendo en cuenta esta distribución de las funciones cerebrales, parece lógico pensar que existen en el cerebro determinadas estructuras en las cuales reside la conciencia. La búsqueda de estas zonas del cerebro ha sido el objetivo de muchos investigadores del campo de la neurociencia. El análisis de las áreas del cerebro encargadas de las distintas funciones mentales se realiza principalmente mediante las siguientes técnicas de diagnóstico por imagen (Rees, Kreiman & Koch 2002, Montz Andrée et al. 2002, Taylor 1999, Fenwick et al. 2003):

- FMRI (Resonancia Magnética Funcional por Imágenes),
- PET (Tomografía por Emisión de Positrones),
- MEG (Magnetoencefalografía) y
- MFT (Tomografía por Campo Magnético).

Estas herramientas de diagnóstico, combinadas con otras como el EEG (electroencefalograma), proporcionan a los neurólogos indicios sobre la actividad cerebral de los sujetos sometidos a observación. Mediante diferentes pruebas se sacan conclusiones sobre las áreas del cerebro implicadas en las distintas funciones mentales, incluido el fenómeno de la conciencia. Este análisis no es simple, ya que la experimentación está sometida a un fuerte componente subjetivo por parte del paciente observado. Además, tanto las funciones mentales como las correspondientes áreas del cerebro implicadas suelen superponerse, haciendo más complicada la interpretación de los datos obtenidos durante la experimentación.

En general, se suelen estudiar independientemente funciones cognitivas concretas asociadas con la conciencia. Por ejemplo, los mecanismos de atención también se han estudiado del mismo modo (Fenwick et al. 2003, Itti, Rees & Tsotsos 2005). Se han realizado multitud de estudios en los que se analiza la información de los escáneres MEG realizados a los pacientes, concluyendo que el cerebro en vez de registrar los estímulos ambientales, lo que hace es extraer su significado, enmarcado en el momento y la situación de la percepción. Lo que se observa en estos experimentos es que todas las áreas del cerebro están ampliamente interconectadas y la información fluye de una región a otra muy rápidamente.

Todos los autores de investigaciones orientadas a la búsqueda de los CNC están de acuerdo en que la conciencia en los humanos está relacionada con las zonas más modernas del cerebro (evolutivamente hablando): el cortex cerebral. Concretamente, la mayoría de los investigadores suelen apuntar al complejo talamocortical (Llinás et al. 1998, Edelman, Tononi 2000, Laureys, Owen & Schiff 2004). Se ha descubierto que la rica interconectividad de este complejo de conexiones reentrantes convierte al tálamo en una especie de centro de comunicaciones, que permite una conexión flexible entre las

diversas áreas del cortex. La actividad cerebral que tiene lugar durante la existencia de estados mentales conscientes parece corresponder con la función global que desempeña el complejo talamocortical.

Partiendo de esta base existen diferentes teorías y modelos que proponen una explicación, e incluso una localización concreta, para ciertos procesos cerebrales asociados con los estados conscientes. Se han buscado de forma exhaustiva las zonas del cerebro que tienen que estar presentes para que exista conciencia en el individuo, así como la actividad neuronal que se correlaciona con los procesos conscientes. Sin embargo, todavía hay mucha incertidumbre acerca de los criterios que se deben usar para determinar si una determinada zona del cerebro participa en la producción de la conciencia. De hecho, aún hoy en día se estima que más del 40% de los pacientes con trastornos de la conciencia son diagnosticados erróneamente (Schnakers et al. 2009). Recientemente se ha demostrado que pacientes que se creía estaban en estado vegetativo, en realidad sufren otro tipo de trastornos, como el estado de mínima conciencia o el síndrome de enclaustramiento (Monti et al. 2010). Desafortunadamente, y aunque se ha realizado numerosas investigaciones, no se conoce aún con precisión cuál es exactamente el mecanismo biológico que genera la conciencia (especialmente, su dimensión fenomenológica).

Aunque no hay un consenso sobre la forma concreta en la que la conciencia se genera en el cerebro, sí que existen ciertas propiedades, características y suposiciones comúnmente aceptadas por la comunidad científica dedicada a las neurociencias. En concreto, la literatura contiene alusiones recurrentes a procesos de sincronización globales, coherencia o coaliciones neuronales (ver Apartado 2.1.3.2). También se acepta generalmente la idea de que la conciencia ha aparecido en la tierra porque supone una ventaja evolutiva (ver Apartado 2.1.3.3).

En base a estos conocimientos sobre neurobiología se han propuesto diversas teorías o hipótesis que tratan de dar explicación a la producción de la conciencia en el cerebro. Las siguientes propuestas son de especial interés para la investigación en Conciencia Artificial y se describen brevemente a continuación:

- La sincronización y coherencia neuronal.
- Los enfoques evolucionistas.
- La hipótesis del Núcleo Dinámico (ND) y la Teoría de la Selección del Grupo de Neuronas (TSGN) de Edelman y Tononi (Edelman, Tononi 2000, Tononi, Edelman 1998, Edelman 1987).
- Teoría de la Integración de la Información (TII) de Tononi (2004).
- Teoría del Marcador Somático (TMS) de Damasio (Damasio, Everitt & Bishop 1996).
- Modelo CODAM de Taylor (Taylor 2003).

2.1.3.2 Sincronización global, unidad, coherencia y coalición

Una de las características principales de la experiencia consciente es que es unitaria. Sin embargo, gracias a los estudios neurobiológicos, se sabe que el cerebro no funciona de forma unitaria. Dado que el contenido de la conciencia es único e integrado

en cada momento, mientras que el procesamiento inconsciente de la información es masivamente paralelo, muchas de las teorías analizadas se basan de una u otra forma en los conceptos de sincronización, coherencia o coalición. En definitiva, cualquier teoría de la conciencia tiene que explicar cómo es posible que se genere un único hilo secuencial coherente a partir de mecanismos masivamente paralelos y aparentemente asíncronos. Es decir, una buena teoría de la conciencia tiene que dar solución al problema de la unidad de la experiencia consciente (*binding problem*).

En el cerebro las propiedades de los objetos que se perciben están separadas. Regiones neuronales específicas detectan el color, la forma, el movimiento, etc. Sin embargo, la percepción consciente de los objetos se manifiesta como una unidad. Los científicos por lo tanto concluyen que debe existir un mecanismo cerebral que permite unir e integrar todos los aspectos de la percepción para generar la experiencia consciente. Este mecanismo tiene que estar relacionado con alguna forma de sincronización y conexión de diversas regiones cerebrales.

A nivel neuronal, la sincronía se entiende como un número elevado de neuronas, correlacionadas entre sí, que se activan de forma similar en el mismo instante. Se ha especulado acerca de que la sincronía neuronal proporciona una especie de suma que puede dar lugar a un conjunto integrado con entidad única (Senkowski et al. 2007). En su búsqueda de los Correlatos Neuronales de la Conciencia, Crick y Koch llegaron a argumentar que la activación sincronizada a 40 Hz (ondas gamma) de coaliciones de neuronas era la base física de la conciencia (Crick, Koch 1990). Aunque más tarde se retractaron al comprobar que estas activaciones entre 35 y 75 Hz en el cortex cerebral no tenían que estar necesariamente relacionadas con los procesos conscientes (Crick, Koch 2003). Muchos investigadores de la conciencia han estudiado la sincronía gamma y su posible relación con la conciencia, sin embargo aún no está claro el papel que juegan en la producción de la conciencia las ondas gamma producidas en el neocortex (Vanderwolf 2000, Doesburg, Kitajo & Ward 2005).

Situada en una línea de investigación análoga, la teoría del cerebro oscilatorio argumenta que los procesos oscilatorios que tienen lugar en el cerebro, y que se manifiestan por medio de las oscilaciones en el EEG, son el puente que puede explicar la relación entre las funciones cognitivas y los estados conscientes del cerebro (Basar et al. 1999). Este tipo de teorías podría proporcionar a la neurofisiología la clave de la integración de la actividad de las áreas cerebrales con las correspondientes funciones mentales desde un punto de vista global.

También Damasio habla de coherencia en otro sentido similar (Damasio et al. 1990). La reverberación en zonas neuronales de convergencia sensorial integra la información de cada sentido. A su vez, toda la información procedente de cada sentido se integra en una única zona de convergencia multimodal que daría lugar a los contenidos conscientes. De forma similar Schacter et al. (1995) plantean la teoría de que múltiples módulos especializados mandan información a un único módulo consciente y analizan evidencias de la disociación de los diferentes tipos de conocimiento en el cerebro.

Tal y como se explica en el Capítulo 4, la arquitectura cognitiva propuesta en la presente tesis implementa mecanismos de sincronización global y coherencia. Aunque estos mecanismos no se logran mediante técnicas directamente inspiradas en los estudios neurobiológicos descritos anteriormente, la funcionalidad final obtenida es equivalente. Por ejemplo, desde el punto de vista de la integración de la información sensorial, la arquitectura propuesta presenta mecanismos de coalición que permiten

fusionar datos sensoriales mono-modales (ver Apartado 4.5) y multi-modales (ver Apartado 7.4.3.1).

2.1.3.3 La evolución de la conciencia

Algunos autores piensan que se puede aprender mucho sobre la conciencia humana estudiando como ésta ha aparecido en la evolución de la Tierra, es decir, analizando la existencia de conciencia en algunos mamíferos muy evolucionados como los primates o los delfines (Gallup 1977, Riba 1998). En los animales el estudio de la conciencia se realiza principalmente en base a su comportamiento (Seth, Baars & Edelman 2005), considerándose que el nivel de conciencia más avanzado se encuentra en los homínidos, siendo equivalente a la capacidad de conciencia desarrollada por un humano de 2 a 3 años de edad (Byrne, Whiten 1988, Savage-Rumbaugh, Mintz Fields & Taglialatela 2000). Aunque la conciencia suele asociarse exclusivamente con los mamíferos superiores, hay investigaciones que indican la existencia de comportamientos asociados a la conciencia en otros animales como cefalópodos y aves (Edelman, Baars & Seth 2005).

El hecho de que la conciencia haya aparecido en nuestro planeta como resultado de un proceso evolutivo de selección natural es una teoría mayoritariamente aceptada (en el ámbito científico). Las teorías evolucionistas tratan de explicar por qué la conciencia supone una ventaja evolutiva para los seres que la poseen. En este sentido, es útil aplicar de nuevo la distinción entre conciencia A, conciencia P, conciencia M y conciencia S.

Según Edelman (1989), la conciencia P apareció en la evolución cuando la capacidad de categorización perceptual se enlazó, mediante procesos de reentrada (ver Apartado 2.1.3.4), con la memoria, creando así el llamado “presente recordado”. La conciencia S y la capacidad de construir imágenes mentales pasadas y futuras aparecerían más tarde en la evolución, cuando los caminos neuronales de reentrada enlazasen las categorizaciones con las capacidades de lenguaje y memoria conceptual (Edelman 2003).

Según Damasio (1999), la mente humana existe hoy en día tal y como la conocemos porque hace mucho tiempo que nuestra especie ha aprendido a tener regulaciones innatas para la supervivencia, tales como el miedo al peligro y la apetencia por los alimentos. En resumen, las emociones conscientes son un arma para la supervivencia.

Una de las hipótesis defendida por algunos autores acerca del origen de la autoconciencia se basa en la capacidad de establecer una correspondencia entre la propia imagen corporal y la de otros individuos (Zlatev 2000). Esta correspondencia sería posible gracias a un “sistema espejo” parcialmente innato. Este sistema, combinado con la interacción social daría lugar a la intersubjetividad. El ciclo de desarrollo de esta conciencia de uno mismo, o autoconciencia, se basaría en los siguientes procesos:

- Incremento del reconocimiento objetivo de la propia imagen corporal.
- Incremento del control volitivo.
- Incremento del entendimiento de la intencionalidad de los otros.

- Incremento del entendimiento de la intencionalidad de uno mismo.

Según esta hipótesis, la autoconciencia y la empatía están co-determinadas. Esta capacidad de "ponerse en el lugar de uno mismo" y "ponerse en el lugar de otro" se denomina capacidad de *teoría de la mente* (Vygotsky 1980), y se considera una habilidad vital en entornos sociales complejos. Otros autores argumentan que la característica diferenciadora que hace posible la autoconciencia es la capacidad de lenguaje que tienen los humanos (Carruthers 2002). Desde luego, la capacidad de comunicación verbal en los humanos juega un papel clave en las relaciones sociales y es también un indicio claro de la existencia de conciencia (Seth, Baars & Edelman 2005). En sintonía con este razonamiento, algunos autores argumentan que como la conciencia está directamente relacionada con la aparición de individuos pertenecientes a una sociedad, la experimentación con entornos "sociales" multi-robot o robot-persona sería el marco adecuado para el desarrollo de modelos eficientes de Conciencia Artificial (Zlatev 2000, Gavrilets & Vose, 2006).

En relación con la capacidad de los humanos para la teoría de la mente se ha planteado la hipótesis del cerebro social o hipótesis de la "inteligencia maquiavélica" (Byrne, Whiten 1988), que pretende dar una explicación a la rápida evolución del cerebro humano. El cerebro humano ha evolucionado mucho más rápido que el de otros mamíferos. En sólo 25 millones de años han tenido lugar multitud de mutaciones en los genes humanos. La hipótesis de la inteligencia maquiavélica podría explicar este fenómeno y también el hecho de que los humanos tengan un cerebro tan grande y complejo. De acuerdo con esta teoría, la intensa competición social fue (y sigue siendo hoy en día) la razón principal por la cual el cerebro humano ha evolucionado hasta convertirse en un órgano extremadamente complejo y que consume el 20% de nuestra energía (Raichle 2006). La selección natural promocionó a aquellos individuos cuyas estrategias sociales les proporcionaban éxitos social y reproductivo. Sofisticadas estrategias "maquiavélicas", que implicaban comportamientos sociales como las mentiras, la astucia o la creación de grupos sociales fueron la forma de tener éxito en la emergente y compleja sociedad humana.

Gavrilets y Vose (2006) han proporcionado datos que apoyan esta hipótesis. Estos autores han desarrollado un modelo matemático que simula el desarrollo del cerebro humano de acuerdo a la teoría de la inteligencia maquiavélica. En su modelo, los genes controlan el cerebro que inventa y aprende estrategias sociales. Estas estrategias las usan los machos en su competición por aparearse. El modelo sugiere que la capacidad cerebral evoluciona más rápidamente que la capacidad de aprendizaje y que la ventaja de tener un gran cerebro deja de serlo tanto cuando los humanos estamos expuestos a una sociedad moderna.

Tomando como inspiración la evolución de la funcionalidad asociada a la conciencia que puede observarse empíricamente en los organismos biológicos, en esta tesis se ha propuesto una escala que permite medir el nivel de desarrollo de la conciencia también en sistemas artificiales (ver Capítulo 5).

2.1.3.4 La hipótesis del núcleo dinámico

La hipótesis del núcleo dinámico, propuesta por Edelman y Tononi (1998, 2000), se enmarca en el contexto planteado por la teoría de la selección del grupo de neuronas

(TSGN) (Edelman 1987). A la TSGN se le conoce también como Darwinismo neuronal, ya que se basa en un mecanismo de selección para explicar el desarrollo y la función del cerebro.

De acuerdo a la TSNG, en el cerebro se genera un gran número de caminos neuronales de entre los cuales un reducido grupo (de más valor para la supervivencia) tiene que ser seleccionado para generar comportamientos adaptativos. En otras palabras, el cerebro es un sistema selectivo. El proceso de selección se basa en el desarrollo y la experiencia. Esa es la razón por la cual los circuitos neuronales son muy diferentes si comparamos el cerebro de una persona con el de otra.

Durante el desarrollo del cerebro y el aprendizaje, se generan grupos de neuronas que tienden a emitir impulsos de forma sincronizada. Un gran número de estos grupos o circuitos se seleccionan de acuerdo a su valor como generadores de comportamientos útiles y adaptativos.

La TSNG explica la conciencia en términos de lo que se denomina el *Núcleo Dinámico*, que consiste en extensas interacciones reentrantes en el sistema talamocortical. Aquí el concepto de reentrada se refiere al proceso dinámico de interacciones rápidas y recíprocas entre mapas y núcleos neuronales. Edelman y Tononi argumentan que la conciencia se produce gracias a estos procesos de reentrada, los cuales serían capaces de proporcionar la capacidad discriminatoria de la experiencia consciente así como su integración. Según estos autores, una escena consciente particular se experimenta como un todo:

- *integrado* (una escena consciente es indivisible) y
- *diferenciado* (cada escena consciente es única).

La principal afirmación de la hipótesis del núcleo dinámico es que los qualia son estas discriminaciones (integradas y diferenciadas a la vez) realizadas por el cerebro. En pocas palabras, la TSGN defiende la caracterización de la conciencia como un proceso dinámico, donde los Correlatos Neuronales de la Conciencia no se pueden identificar en un punto específico del cerebro, sino que se asocian con la dinámica de reentrada existente en el complejo talamocortical. La hipótesis sugiere que las fronteras del núcleo dinámico cambian a lo largo del tiempo, habiendo grupos de neuronas que entran a formar parte del núcleo y otros grupos que salen de él. Estas transiciones ocurren debido a la influencia de señales externas e internas.

2.1.3.5 Teoría de la integración de la información

De acuerdo con la teoría de la integración de la información (TII), la conciencia se corresponde con la capacidad de integrar información que tiene un sistema (Tononi 2004). Esta teoría se basa en los conceptos de integración y diferenciación presentados previamente como parte de la TSGN. En este contexto la conciencia se caracteriza como un equilibrio entre la integración y la diferenciación. La diferenciación se refiere a la disponibilidad de un gran repertorio de posibles experiencias conscientes. Cada experiencia concreta se diferencia (o se discrimina) de las otras. La integración se refiere al carácter unitario de cada una de estas experiencias, es decir, los contenidos conscientes se experimentan como una unidad, incluso aunque estén compuestos por muchas dimensiones.

Tononi introdujo el valor Φ como una medida de la conciencia de un sistema. De hecho Tononi caracteriza la conciencia como la capacidad de integrar información, siendo Φ la medida de la cantidad de información que es capaz de integrar el sistema: *"Φ es la cantidad de información causalmente efectiva que se puede integrar a través del enlace de información más débil de un subconjunto de elementos"*. Al igual que la hipótesis del núcleo dinámico, la TII indica que el sistema talamocortical constituye el sustrato neuronal de los procesos que producen la conciencia P.

La principal implicación de la hipótesis propuesta por Tononi es que cualquier sistema físico capaz de integrar información tiene un nivel de consciencia tan alto como indique la medida Φ. Esto implicaría que las implementaciones artificiales capaces de integrar información y que tienen un valor de Φ alto, tienen experiencias conscientes (Koch, Tononi 2008).

2.1.3.6 Hipótesis del marcador somático

La hipótesis del marcador somático propuesta por Damasio proporciona una explicación de cómo se produce la modulación del comportamiento en base a procesos emocionales (Damasio, Everitt & Bishop 1996). Aunque esta hipótesis no trata directamente con el problema de la conciencia, sí lo hace con aspectos cognitivos muy relacionados, como la toma de decisiones y las emociones.

La toma de decisiones en los humanos normalmente involucra evaluaciones tanto racionales como emocionales. Ante un repertorio de posibles acciones, que en el mundo real suelen conllevar la existencia de incertidumbre y alternativas conflictivas, parece que el uso de las emociones para elegir la mejor opción es una ventaja evolutiva.

Los marcadores somáticos son asociaciones entre determinados estímulos y estados afectivos. Damasio sugiere que estos marcadores constituyen una asistencia valiosa en el proceso de toma de decisiones. Se argumenta que los marcadores somáticos se localizan en el cortex prefrontal ventromedial (parte de los lóbulos frontales del cerebro humano) y que son capaces de modular el procesamiento cognitivo. En los casos en que se necesita tomar una decisión complicada o bajo incertidumbre, los marcadores somáticos asociados con experiencias de castigo o de recompensa se combinan para producir un estado somático neto o resultante. Este estado general se usa para dirigir (o modular) la selección de la acción más apropiada. Esta modulación o influencia emocional puede tener lugar de forma inconsciente (a través del tronco cerebral y el estriado ventral) o conscientemente (reclutando áreas corticales). Supuestamente, los marcadores somáticos alejan el foco de atención de las opciones más desfavorables, simplificando así el proceso volitivo.

Según Damasio (1995), las emociones y la razón son muy dependientes entre sí. Damasio ilustra este punto con un análisis retrospectivo del famoso caso de Phineas P. Gage (1823-1860), cuyos lóbulos frontales se dañaron en un accidente (Damasio, Everitt & Bishop 1996). Gage sufrió una aparatosa lesión cerebral en el trabajo cuando una vara de hierro traspasó accidentalmente su cráneo (Figura 1). Se dice que el caso de Gage es la primera prueba clínica del papel del lóbulo frontal en la personalidad y la interacción social. De hecho, después de sufrir el accidente, los amigos de Gage comentaron que él ya no era la misma persona, se había convertido en un insociable.

Figura 1. Esquema del daño cerebral sufrido por Phineas P. Gage.

Desde el caso de Gage, Damasio y sus colegas han estudiado seis casos más de daños localizados en el cortex prefrontal ventromedial. Estos estudios han concluido que este tipo de daños incrementan los juicios morales utilitarios (Koenigs et al. 2007)[1].

Aunque esta teoría no proporciona una explicación concreta para la conciencia, supone la base para entender las emociones y la generación del comportamiento como procesos neuronales. Esta concepción de los fundamentos de los procesos mentales llevará a Damasio a formular una teoría más completa, que incluye una explicación tanto para la conciencia esencial (estado de alerta/vigilia) como para la autoconciencia (ver Apartado 2.1.4.3).

2.1.3.7 La atención y el Modelo CODAM

La atención es una cualidad muy relacionada con la conciencia, además, es una de las capacidades cognitivas que más se ha investigado en sistemas artificiales, obteniéndose resultados significativos en el campo de la visión artificial. El sistema VOCUS es un ejemplo destacado (Frintrop 2006). La atención es una característica deseable en un sistema autónomo y también se han estudiado en detalle sus bases neurobiológicas (Itti, Rees & Tsotsos 2005). Taylor describe como la atención en el cerebro humano involucra siempre a dos áreas diferentes: aquellas que se están controlando y las que están realizando el proceso de control requerido (Taylor 2003). El proceso de control en si mismo consiste en la amplificación de las entradas requeridas y la inhibición de las entradas "distractoras".

Se cree que este proceso se lleva a cabo mediante la retroalimentación competitiva existente entre el área controladora y el área controlada. La modulación resultante se logra por la multiplicación de la actividad neuronal en la zona controlada, o sesgando el

[1] En este contexto, el término utilitario viene del utilitarismo y se refiere a los juicios, o dilemas morales, en los que el conflicto se produce entre una agregación de elementos y un comportamiento emocionalmente inaceptable. Por ejemplo, tener que sacrificar la vida de una persona para salvar un número mayor de vidas

límite que dispara la respuesta. Al mismo tiempo es necesario disponer de áreas neuronales donde procesar las entradas a las que se atiende, y si es necesario, mantenerlas por un periodo de tiempo. Estas áreas se conocen como memorias intermedias (*buffers*). Para una atención efectiva también se requiere que se monitoricen las entradas para detectar posibles errores.

Dada esta descripción del mecanismo de atención, algunos autores sugieren que el marco de la ingeniería de control es adecuado para explicar e implementar la atención en sistemas artificiales. Este enfoque sería aplicable tanto a la atención sensitiva como para el aprendizaje y la atención en robots autónomos. De acuerdo con este planteamiento, Taylor ha propuesto el modelo CODAM (*COrollary Discharge of Attention Movement*) (Taylor 2003). Este modelo considera que la llamada "descarga corolario" (o copia de la señal de control que causa el movimiento del foco de atención) es crucial tanto para el mecanismo de atención como para la conciencia misma. En el ámbito de la teoría de control se sabe que el uso de una copia de la señal de control es un mecanismo que proporciona un incremento de la velocidad y la precisión de los sistemas. En el modelo CODAM se propone este mismo mecanismo como un nuevo componente que puede hacer más efectivos los modelos de atención. Según Taylor, este nuevo componente puede explicar la creación de un yo interior, caracterizado como dueño de un estímulo al que se está prestando atención. Este proceso de "atención provocando la conciencia" se desarrolla en el cerebro durante un periodo de tiempo que no es despreciable (del orden de cientos de milisegundos). Por lo tanto, la copia de la señal de control del movimiento del foco de la atención es clave en la protección y aceleración del proceso.

Adicionalmente, se han planteado otros modelos que también se centran exclusivamente en los mecanismos de atención, como el modelo unificado de la atención de Hunt (Hunt, Lansman 1986). En este caso se trata de una simulación donde el mecanismo de atención proporciona el medio para la selección y ejecución de producciones en tiempo real.

2.1.4 Teorías cognitivas de la conciencia

El problema de la comprensión de la conciencia se aborda también desde el punto de vista de los procesos mentales realizados por el sistema nervioso central. Es decir, desde el punto de vista cognitivo. Aunque este tipo de descripciones se puedan realizar de forma prácticamente independiente de los mecanismos neurobiológicos que operan en las redes de neuronas, es común que también se busquen las bases neurológicas que apoyan a las teorías cognitivas de la conciencia.

Las teorías cognitivas analizan la adquisición y el uso del conocimiento en la mente humana. En el caso específico de la conciencia, tratan de comprender, a nivel de procesos, como se gestiona el acceso al conocimiento y el control de un vasto conjunto de complejos procesos paralelos (inconscientes) desde un único hilo secuencial (consciente).

Las teorías sobre la conciencia consideran que en la mente existen dos tipos distintos de procesos: conscientes e inconscientes. Desde el punto de vista de los contenidos con los que operan estos procesos, esta dualidad se expresa en base a las diferentes formas de representar y procesar el conocimiento. Dependiendo de la naturaleza consciente o inconsciente de los procesos el conocimiento que utilizan puede

ser declarativo o procedimental, localizado o distribuido, procesado en serie o en paralelo.

Las hipótesis que consideran la separación entre consciencia e inconsciencia, tienen que plantear los criterios de separación entre ambos dominios, así como el funcionamiento característico de cada uno de ellos. La representación del conocimiento puede ser implícita o explícita. Los procesos conscientes usan información explícita directamente accesible, mientras que los procesos inconscientes manejan información implícita que no es accesible si no es a través de mecanismos interpretativos.

A continuación se describen las principales teorías de la conciencia que ofrece una explicación a nivel cognitivo[2]:

- Teoría del Espacio de Trabajo Global (ETG) de Baars (1988).
- Teoría de las Versiones Múltiples (VM) de Dennett (1991).
- Teoría de las Emociones, Sentimientos y Conciencia (ESC) de Damasio (1999).
- El Enfoque Sensoriomotor de O'Regan y Noë (2001, 2007).
- Teoría de los pensamientos de orden superior (POS) de Rosenthal (2005).

2.1.4.1 La teoría del espacio de trabajo global

Baars utiliza la metáfora de un teatro para dar forma a su teoría conocida como Espacio de Trabajo Global (*Global Workspace Theory*) o ETG (Baars 1988). Baars habla de un "teatro" en el que el foco de la atención se representa como el foco de luz sobre el escenario. El escenario completo se corresponde con la memoria de trabajo, que es el sistema de memoria que almacena los contenidos conscientes. La información obtenida bajo el foco de luz se distribuye de forma global a través del teatro a dos clases de procesadores inconscientes: los que forman la audiencia reciben información del foco de luz; mientras, entre bastidores, los sistemas inconscientes contextuales dan forma a los sucesos que ocurren en escena (ver Figura 2).

La metáfora del foco luminoso es también utilizada por Crick argumentando, acerca del procesamiento de la información visual, que fuera de la iluminación proporcionada por el foco – atención visual – la información se procesa menos, de forma diferente o ni siquiera se procesa (Crick 1994).

No hay que confundir esta metáfora del teatro que usa Baars con otra metáfora denominada "Teatro Cartesiano", que es en esencia opuesta a la defendida por Baars, ya que consiste en una postura dualista que atribuye la conciencia a un punto concreto del cerebro: la glándula Pineal. Descartes pensaba que en esta glándula se localizaba el enlace del cerebro con el alma (Finger 1995).

Volviendo a la metáfora del teatro propuesta por Baars, es importante resaltar que el "escenario" está compuesto por la memoria de trabajo. Donde los "actores" compiten por aparecer en el foco luminoso de la atención, en el cual aparecen como contenidos completamente conscientes. La selección de los contenidos que aparecen bajo el foco de

[2] Estas teorías han constituido la principal fuente de inspiración para el diseño del modelo MC^3 descrito en el Capítulo 4.

atención se realiza en gran medida entre bastidores. Son los procesadores inconscientes los que llevan a cabo esta selección en base al contexto y a conjuntos de creencias (a menudo inconscientes) que determinan los pensamientos conscientes (la actuación que tiene lugar en escena).

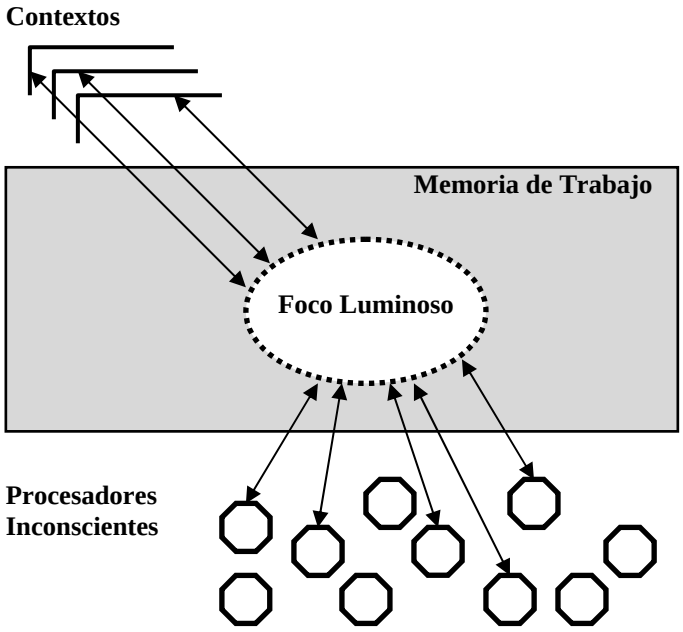


Figura 2. Esquema de la Teoría del Espacio de Trabajo Global.

Baars también indica que el foco luminoso de la conciencia es el instrumento que usa el "director" para tomar decisiones sobre los contenidos de la memoria de trabajo. Estas decisiones están guiadas por la persecución de metas. El director de la obra, también trabaja entre bastidores, lo que sugiere que en gran medida no tenemos acceso a las razones por las que hacemos las cosas. Este concepto encaja con el presentado por algunos autores (Rosenthal 2000a, Morin 2002), que afirman que el *yo* consciente *confabula* para deducir las razones por las que el sujeto lleva a cabo sus acciones.

Según Baars, la conciencia es un mecanismo efectivo para poder acceder de forma eficiente al vasto dominio de conocimiento y control que permanece inconsciente. La conciencia se usa para el aprendizaje rápido y el reconocimiento preciso. También activa un gran número de rutinas automáticas que constituyen acciones específicas, proporcionando coordinación y control. Las experiencias conscientes activan contextos inconscientes, que ayudan a interpretar sucesos conscientes futuros. En definitiva, la conciencia proporciona un marco para el acceso (y función de búsqueda global) a los vastos contenidos inconscientes de la mente.

Parece que las investigaciones realizadas con métodos de diagnóstico por imágenes indican que esta hipótesis podría ser cierta (Baars 2002, Baars, Ramsoy & Laureys 2003); en cualquier caso, se necesitan más análisis neurobiológicos para confirmar o desmentir con total seguridad las suposiciones de Baars.

Hay que destacar que la teoría del ETG proporciona una explicación plausible para la conciencia A, e incluso podría extenderse para dar explicación a la conciencia M y la

conciencia S. Sin embargo, es difícil encontrar una explicación válida para la generación de estados fenomenológicos atendiendo exclusivamente a la teoría de Baars. La presente tesis pretende avanzar en este aspecto proponiendo un modelo para la generación de qualia artificiales (ver Capítulo 6) y aplicándolo directamente a la arquitectura cognitiva propuesta (ver Capítulo 4), en la que una de las principales fuentes de inspiración para su diseño es la teoría del ETG.

2.1.4.2 La teoría de las versiones múltiples

El modelo de las versiones múltiples (*Multiple Draft Model*), descrito por Dennett (1991), emplea la metáfora de la revisión editorial para describir los procesos cognitivos de la conciencia. En este ámbito, las coaliciones de procesadores inconscientes son sometidas a un proceso de edición y revisión reiterativo (produciendo diversos borradores) hasta que se presentan oficialmente como contenidos conscientes de la mente.

Dennett argumenta que no existe ningún lugar en el cerebro donde se realice un control centralizado. Contrariamente, la actividad se desarrolla en subprocesos distribuidos, los cuales no tienen información sobre otros subprocesos que paralelamente se crean en la mente. Esto contrasta con la metáfora planteada por Baars, donde se supone que existe un lugar central de representación de los contenidos conscientes. Dennett también descarta explícitamente el dualismo, que llega a calificar como mero resultado de la ignorancia o incapacidad científica de otros autores.

Desde el punto de vista funcional, Dennett define al ser consciente como poseedor de un punto de vista. Como el cerebro tiene una limitación en cuanto a la cantidad de información que puede procesar en un determinado momento, una mente consciente es un observador que recoge un subconjunto parcial de toda la información que hay en el entorno. Esta información no se trata secuencialmente, si no que el cerebro dispone de numerosos módulos paralelos que procesan esta información. Dennett añade que además de procesarla, la información está constantemente siendo revisada y modificada para adaptarse a las nuevas informaciones que van entrando. Esto explica nuestra capacidad de ser conscientes simultáneamente de percepciones que tardan diferente tiempo en ser procesadas (por ejemplo, los estímulos visuales tardan más en ser procesados que los auditivos, sin embargo la experiencia consciente resultante no refleja esta asincronía).

De acuerdo con la visión de Dennett, lo que experimentamos de forma consciente es el resultado de multitud de procesos interpretativos. Muchas de las versiones que se van elaborando son descartadas y desaparecen a lo largo del tiempo. Cuando sólo queda una versión ganadora, ésta pasa a formar parte del flujo narrativo y se almacena en memoria. A su vez, estos contenidos de la memoria que afectan a la conciencia también están sujetos a revisión. Dado que en este modelo no existe un revisor central que de sentido a la narración que se está creando, no se puede especificar un momento concreto en el que un contenido que no es consciente pasa a serlo.

Dennett argumenta que la conciencia se ejecuta en una especie de máquina virtual modulada por el aprendizaje cultural y que se ejecuta sobre el hardware paralelo del cerebro (ver Figura 3). Esta máquina virtual instala en el cerebro un conjunto organizado de “hábitos de mente” que no son visibles a nivel neuroanatómico. El éxito

de la implantación de este tipo de maquinaria consciente se debe a un sinfín de micro-disposiciones existentes gracias a la plasticidad cerebral.

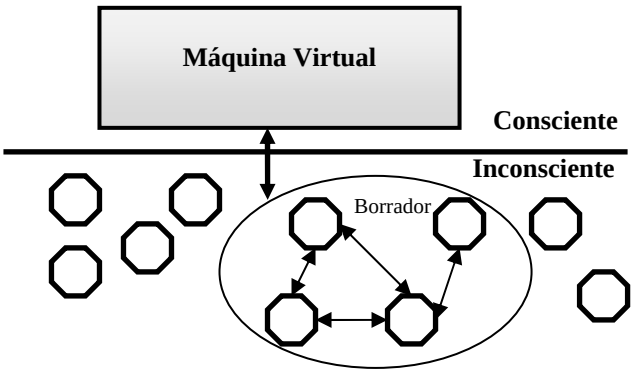


Figura 3. Esquema de la Teoría de las Versiones Múltiples.

En definitiva el modelo propuesto por Dennett implica que la versión consciente de nuestro pensamiento es la ganadora de un proceso continuo de revisión y cambio. Los factores que determinan qué versión es la "oficial" en cada momento pueden ser diversos. Por ejemplo, en momentos de ira, las coaliciones de procesadores inconscientes que presenten versiones más iracundas serán las ganadoras. En cualquier caso, el cerebro siempre busca una explicación que se adapte plausiblemente (según ciertos criterios) a las informaciones percibidas, pero no necesita necesariamente que la versión consciente se corresponda con la realidad. Esta posición concuerda con numerosos estudios realizados en el campo de la neurociencia, que demuestran que en casos en los que hay falta de información fidedigna el cerebro usa a menudo información inventada – falsa – para poder construir una escena consciente completa (Rubia 2000). El cerebro nos engaña para poder ofrecer una representación completa y coherente del mundo que nos rodea.

2.1.4.3 Teoría de emociones, sentimientos y conciencia

Actualmente los científicos aceptan mayoritariamente la idea general de que los sentimientos son el balance consciente de nuestra situación (Marina 2002). Según diferentes teorías psicológicas, los sentimientos son la forma en que los seres humanos son capaces de sintetizar su situación en el mundo dentro del ámbito limitado de la conciencia. Por lo tanto, la dimensión sentimental no puede ser obviada en el estudio de la conciencia, ya que ésta está directamente relacionada con la atención, el lenguaje y por supuesto el estado afectivo (Ciompi 2003).

El comportamiento y la percepción están profundamente influidos por el estado emocional del sujeto. Los psicólogos están de acuerdo en que existe un conjunto básico de emociones "trans-culturales" (Wierzbicka 1986). Sin embargo, la lista concreta depende de cada autor. La gran variedad de emociones identificadas es el resultado de modulaciones dependientes de la cultura o mezclas de las emociones básicas.

Un resumen extraordinariamente simplificado, sin entrar en el complejo mundo de los sentimientos, sería el siguiente: el sujeto siente alegría en caso de ver que sus objetivos se van cumpliendo de la forma prevista. En caso contrario, se siente frustrado. En los humanos los sentimientos influyen en la conducta, entre otras cosas, en base a un sistema de creencias. Por eso, bajo un estado de frustración unos individuos actúan desistiendo de sus objetivos originales por completo, mientras que otros optan por intentar diferentes alternativas.

Aunque Damasio se había centrado inicialmente en las bases neurológicas de las emociones (ver Apartado 2.1.3.6), la íntima relación existente entre emociones y conciencia le llevó a plantear también una teoría sobre la conciencia (Damasio 1999). Según Damasio, un sujeto alcanza la autoconciencia cuando su cerebro usa modelos de su cuerpo, del mundo exterior y de la interacción entre ellos. Cuando logra esto, el sujeto es autoconsciente, pero sólo instante a instante. Es decir, vive el momento. Si su cerebro también recuerda el pasado, entonces el sujeto también será consciente de sí mismo como una entidad que tiene una historia pasada.

Un sujeto que tiene emociones, que tiene una historia pasada y que es autoconsciente se convierte en un agente motivado que busca el placer, evitar el dolor y tiene miedo a la muerte. Se convierte en algo vivo.

Damasio cree que el mecanismo básico de autoconsciencia que poseen los humanos se produce de la siguiente forma (Doan 2009a): el cuerpo tiene multitud de sensores internos que captan información como la concentración de productos químicos existentes en la sangre, el ritmo cardiaco, etc. Este inmenso flujo de información sensorial es como una película (muy rica en contenidos multimedia) que se proyecta en el cerebro. Un ordenador con sensores internos también poseería esta capacidad de generar una película similar. Además, si tal ordenador estuviera equipado con el software adecuado, también podría formar un modelo del cuerpo actualizado a partir de la información de los sensores propioceptivos.

De acuerdo con la teoría de Damasio, la maquinaria biológica del cerebro también crea un modelo adicional. Un “modelo de segundo orden” que incluye al modelo del cuerpo mencionado anteriormente. En este modelo de segundo orden se representa el cuerpo, el mundo exterior tal y como lo perciben los sentidos exteroceptivos y también están representadas las interacciones entre este mundo exterior y el cuerpo.

En cada instante el cerebro tiene dos vistas simultáneas: una es la vista “desde el exterior” – como si fuera un observador, el cerebro ve (desde el modelo de segundo orden) el cuerpo que tiene una mano, y esa mano extendiéndose para tocar una taza de café. La otra vista es la del modelo del cuerpo, que también le dice al cerebro que sus ojos ven la taza y, ¡oh! Su piel nota que está caliente.

Por lo tanto, al cerebro se le presenta el siguiente hecho: instante tras instante, interactuando con un objeto tras otro, estas dos vistas siempre coinciden. Irremediablemente el cerebro detecta que ese cuerpo es su cuerpo: “la cosa que está tocando la taza soy yo mismo”. El cerebro se hace consciente de su ser – alcanza lo que Damasio llama la “conciencia primordial” (*core consciousness*). A partir de este punto, surge la conciencia humana. Según esta teoría, no hay razón aparente por la cual la conciencia artificial no pueda aparecer de la misma forma.

2.1.4.4 El enfoque sensoriomotor

Todas las teorías o hipótesis presentadas anteriormente asumen que la conciencia se genera en el cerebro. Sin embargo, el enfoque sensoriomotor de O'Regan y Noë va más allá: su tesis es que la conciencia P, los qualia, son en realidad una forma de explorar el entorno mediante procesos de transformación sensoriomotores (O'Regan, Noë 2001). Este punto de vista evoca una visión extendida de la mente, ciertamente anti-intuitivo, donde las experiencias conscientes son en realidad leyes sensoriomotoras que el cerebro aprende cuando se relaciona con su entorno.

O'Regan afirma que el enfoque sensoriomotor elimina el "problema duro" de la conciencia (Chalmers 1995), ofreciendo una explicación científica para los "misterios" de los qualia (O'Regan 2007). Según Chalmers, es imposible describir las sensaciones conscientes, es decir, son totalmente subjetivas. Este razonamiento ha llevado a calificar la búsqueda de una explicación para la conciencia P como el problema duro. Aunque otros autores, como Dennett (1991), ya han argumentado anteriormente que el problema duro no existe, no han conseguido dar una explicación convincente para la generación de los qualia. Es más, de acuerdo con el enfoque sensoriomotor, ni siquiera los neurobiólogos pueden ofrecer una explicación para la fenomenología.

Lo que un neurobiólogo puede explicar acerca de la rojez del rojo es que cuando una persona ve algo rojo en su cerebro hay un canal neuronal rojo-verde que se activa en la dirección del rojo. Sin embargo, eso no explica la experiencia asociada (la sensación de rojo). En otras palabras, no hay una forma obvia de salvar el hiato explicativo simplemente mirando resultados experimentales en el ámbito neurobiológico. Explicar el sentido de presencia o el hecho de "cómo es sentir algo" parece estar fuera del alcance de la fisiología.

O'Regan argumenta que podemos avanzar en la comprensión de la experiencia subjetiva considerando exactamente cuáles son los principales puntos que parecen difíciles de explicar. Chalmers sugiere que los principales misterios asociados a la capacidad de sentir conscientemente son:

- *La inefabilidad*. Imposibilidad de comunicar o informar acerca de los qualia de forma completa y detallada. Es decir, no se puede realizar una especificación completa para comunicar de forma precisa lo que sentimos subjetivamente.
- *La estructura de los qualia*. Los qualia tienen una estructura que permite comparar unos con otros. Tienen dimensiones como la modalidad sensorial: los qualia correspondientes a diferentes modalidades sensoriales son diferentes, sin embargo se pueden comparar en algunos aspectos. Por ejemplo, las sensaciones de un color suave y un tacto suave (ambas tienen una característica común, pero pertenecen a modalidades diferentes).
- *La sensación de presencia*. Los qualia tienen una sensación de presencia subjetiva o el "cómo es sentir algo" en primera persona. Las experiencias sensoriales tienen esta característica, mientras que otras actividades mentales como pensar o recordar no la tienen.

Estos hechos acerca de la capacidad de sentir conscientemente parecen muy difíciles de comprender en base a procesos físico-químicos del cerebro. Especialmente, la propiedad de presencia de los qualia es la que a menudo se considera algo especial y la que ha llevado a puntos de vista dualistas en el pasado.

Según O'Regan, estos misterios se pueden desvelar si en vez de pensar en el cerebro como en un generador de experiencias, se considera que el sentir es una forma de interacción con el mundo. Adoptando esta visión sensoriomotora, las cualidades experimentadas son equivalentes a leyes sensoriomotoras. La "rojez" no es algo generado por el cerebro, sino la forma en que las cosas rojas cambian la luz que índice sobre ellas. Sentimos la rojez cuando nos damos cuenta de que estamos interactuando de forma eficiente con un objeto que cambia la luz siguiendo un determinado patrón que conocemos como rojo.

De acuerdo con el enfoque sensoriomotor, los tres enigmas asociados a la conciencia P se pueden explicar de la siguiente forma:

- *Inefabilidad*. Los qualia son inefables porque no tenemos acceso cognitivo a todos los detalles de nuestras propias capacidades sensoriomotoras. Por ejemplo, cuando tensamos un músculo no tenemos acceso a la longitud de las fibras musculares.

- *Estructura de los qualia*. Las similitudes y diferencias entre sensaciones correspondientes a diferentes modalidades sensoriales se pueden explicar en base a las similitudes y diferencias que existen entre las leyes objetivas que determinan las interacciones sensoriomotoras involucradas. Por ejemplo, la cualidad asociada a la vista está constituida por todas las leyes que potencialmente se cumplen cuando se practica la capacidad de ver. Como confirmación de esa teoría, los experimentos de sustitución sensorial demuestran que por ejemplo la vibración en la piel se puede sentir como visión (Avillac et al. 2005), dado que las leyes relacionadas con la visión se cumplen igualmente cuando el observador se mueve utilizando un panel de vibración sobre el abdomen en vez de la vista.

- *Presencia*. Explicar el sentido de presencia o el "cómo es sentir algo" requiere explicar por qué las experiencias sensoriales tienen algo muy especial que otras actividades mentales no tienen. Para explicar esta diferencia en el marco de la teoría sensoriomotora basta con buscar diferencias objetivas entre las leyes sensoriomotoras que gobiernan los actos perceptuales y las leyes que gobiernan actividades mentales como pensar o recordar. Parece entonces que lo que caracteriza particularmente a las sensaciones perceptivas y lo que las distingue de las experiencias mentales es que las interacciones sensoriales tienen corporeidad (*bodiliness*) y son llamativas (*grabbiness*) (O'Regan 2007, 2010).

La corporeidad (*bodiliness*) radica en el hecho de que cuando el cuerpo se mueve, la entrada sensorial cambia. Las actividades mentales no tienen corporeidad. La capacidad de llamar la atención (*grabbiness*) consiste en el hecho de que los sistemas perceptuales están conectados de tal manera que interrumpen el proceso cognitivo cuando hay un cambio súbito en la entrada sensorial. Las actividades mentales tampoco poseen esta capacidad de atracción de la atención. Según la teoría sensoriomotora, la corporeidad y la capacidad de llamar la atención son hechos objetivos sobre las actividades sensitivas que parecen capturar intuitivamente el hecho de que somos sujetos de nuestras propias experiencias sensoriales – éstas se manifiestan en nosotros. Esto encaja con la noción de presencia o "cómo es sentir". Análogamente, el hecho de que las actividades mentales y las actividades asociadas con procesos autónomos del sistema nervioso no tengan corporeidad ni capacidad de atraer la atención brinda una explicación acerca de por qué sólo los estados sensitivos poseen "presencia" o un "cómo es sentir".

En resumen, en enfoque sensoriomotor proporciona una explicación para la inefabilidad, la estructura y el sentido de presencia de las sensaciones. También proporciona una forma de salvar el hiato explicativo en términos de las cualidades de la experiencia subjetiva. Este enfoque es también aplicable al modelo para la generación de qualia artificiales que se propone en la presente tesis (ver Capítulo 6).

2.1.4.5 La teoría de los pensamientos de orden superior

Las teorías llamadas de orden superior tratan de explicar la conciencia en base a la relación entre los estados conscientes y una representación de más alto nivel (de orden superior). Estas representaciones de más alto nivel o superiores pueden ser o bien una percepción de más alto nivel de ese estado, o bien un pensamiento o creencia de orden superior acerca del estado (Carruthers 2009). En general estas teorías tratan de explicar las propiedades distintivas de la conciencia y en especial la dimensión fenomenológica de la misma.

Las teorías de orden superior de la conciencia P afirman que un estado mental fenomenológicamente consciente es un estado mental que es el objeto de una representación de orden superior de un cierto tipo. Por ejemplo, la teoría de los pensamientos de orden superior (POS) establece que un estado consciente es aquel que va acompañado de un pensamiento – de orden superior – acerca de que ese estado es en el que se encuentra uno mismo (Rosenthal 2000b). La motivación principal de las teorías basadas en elementos de orden superior es que la mayoría de los estados mentales se pueden dar de forma consciente o inconsciente.

De acuerdo con Rosenthal (Rosenthal 2005, Rosenthal 2000b), un estado mental consciente M, que es mío, es un estado que en realidad está causando una creencia (generalmente no consciente) de que yo tengo M. Además, está causando esta creencia de forma que no se puede inferir su causa. La conciencia P se explica en base a que M debería tener un papel causal – y/o contenido de cierto tipo distintivo – que permita considerarlo como una experiencia; y que cuando M sea una experiencia (o una imagen mental o sensación) sea consciente en términos fenomenológicos cuando esté en el foco de atención.

2.1.5 Otras teorías de la conciencia

Las teorías expuestas en los apartados anteriores son las que cuentan con más popularidad en la comunidad científica y también las que en mayor medida suelen aplicarse como inspiración en el campo de la Conciencia Artificial (y de hecho son las que en mayor medida se han aplicado en la presente tesis). Sin embargo, existen más propuestas que pueden considerarse en algunos casos marginales o que simplemente no tienen, o no se ha encontrado aún, una aplicación práctica directa (como ocurre específicamente en el caso de las teorías basadas en la mecánica cuántica). Estas propuestas pueden dividirse en los siguientes grupos de teorías:

- Teorías basadas en la simulación interna y el auto-modelado.
- Teorías basadas en las propiedades físicas del sustrato neuronal.

2.1.5.1 Simulación interna y auto-modelado

Aunque los modelos conceptuales de alto nivel de descripción no especifican necesariamente los mecanismos neuronales (de bajo nivel) que dan lugar a la conciencia, pueden servir para proporcionar conocimiento útil acerca de los procesos que tienen lugar en el cerebro para que se produzca la conciencia.

La mayoría de los modelos de alto nivel propuestos se basan en la idea de que la conciencia surge de la simulación de las interacciones entre el organismo y el entorno que se realiza en el cerebro. Por lo tanto, estos modelos suelen hacer especial hincapié en el desarrollo del yo y la subjetividad.

Un ejemplo típico de teoría de la subjetividad es el modelo fenomenológico del yo propuesto por Metzinger (2003). Este filósofo propone una serie de modelos conceptuales que se centran en el papel que juega el *yo* (*self*) en la experiencia consciente para así explicar la subjetividad. Según Metzinger, los yos no existen, nunca han existido en el mundo tales cosas como los yos, lo que en realidad existe en el mundo son los "modelos fenomenológicos del yo" (MFY) o yos fenomenológicos. Un yo fenomenológico no es una cosa, sino un proceso continuo. Un MFY es el contenido de un "auto-modelo" que consiste en una representación dinámica del organismo y que es transparente; es decir, no puede ser reconocido como un modelo por el propio sistema que lo usa.

Según Metzinger, la existencia de MFY permite que se establezca una distinción entre las señales relacionadas con el entorno y las señales relacionadas con el organismo, las cuales a su vez permiten que el propio organismo modele la relación intencional entre el sujeto (MFY) y el mundo (Seth 2007). Metzinger sugiere que esta forma de crear modelos puede generar contenido fenomenológico, en cuyo caso el modelo se denomina "modelo fenomenológico de la relación de intencionalidad" (MFRI). Un componente subjetivo S (el MFY) se representa fenomenológicamente al ser dirigido hacia un objeto intencional O.

La teoría de Metzinger supone un análisis representacionalista y funcional de lo que es la experiencia consciente en primera persona. El objetivo último de esta teoría es explicar cómo emerge la "sensación de ser un sujeto" y cómo este proceso se puede analizar en base a niveles de descripción subpersonales. Esto enlaza con el enfoque heterofenomenológico propuesto por Dennett (1991), ya que también se pretende dar una explicación a la subjetividad desde la utilización de modelos conceptuales y herramientas que no se basan exclusivamente en la instrospección.

De acuerdo con Metzinger, los seres conscientes confunden constantemente el contenido de su MFY con su propio yo, es decir, piensan que son un yo real. Esto se debe a la naturaleza del proceso representacional que genera el auto-modelo. Este auto-modelo es en gran medida transparente, por lo que la información de que es un modelo no forma parte del contenido del propio modelo. Los seres conscientes ven a través de este modelo y por lo tanto tienen la impresión de estar en contacto directo con su propio cuerpo y el mundo que les rodea. Según Trehub, para que un sujeto pueda desarrollar un MFY es preciso que tenga una capacidad innata para crear representaciones egocéntricas (Trehub 2007). De acuerdo con este mismo autor, esta capacidad para

poder tener experiencias con una perspectiva egocéntrica se encuentra en las estructuras "retinocéntricas" del cerebro.

Aunque esta idea es anti-intuitiva, Metzinger trata de demostrar que es una visión más acertada de la realidad y aplica su teoría para explicar casos de trastornos neurofenomenológicos como los trastornos de identidad, síndrome del miembro fantasma o las experiencias extra-corporales (Metzinger 2003).

El proceso representacional también puede verse desde el punto de vista de la simulación interna. Concretamente, Revonsuo propone utilizar la metáfora de la realidad virtual (Revonsuo 2005). Según este enfoque, la conciencia emerge como una inmersión total en un mundo simulado que representa las interacciones entre el organismo y el entorno. Esta misma idea del modelado o simulación interna es defendida por Hesslow (2002), quien plantea las siguientes suposiciones (en base a pruebas neurológicas):

- *Simulación de las acciones*. Podemos activar las estructuras motoras del cerebro de una forma parecida a la ejecución de una acción normal, pero sin llegar a efectuar ningún movimiento en la realidad.
- *Simulación de la percepción*. Imaginar percibir algo es esencialmente lo mismo que percibirlo de verdad, la única diferencia esencial es que no hay un estímulo exterior.
- *Anticipación.* Existen mecanismos asociativos que permiten que la actividad perceptual o motora provoque la activación de las áreas sensoriales del cerebro. De hecho, una acción simulada puede provocar actividad perceptual que se parece a la que se generaría si la acción se hubiera realizado realmente.

Teniendo en cuenta estos puntos, Hesslow argumenta que la teoría de la simulación interna puede explicar las relaciones existentes entre las funciones sensoriales, motoras y cognitivas y la aparición de un mundo interior fenomenológico. En línea con estas investigaciones es interesante destacar que recientemente se ha demostrado que incluso pacientes con trastornos de la conciencia – aparentemente inconscientes e inmóviles – pueden simular internamente la realización de diversas actividades (Monti et al. 2010).

2.1.5.2 Propiedades físicas del sustrato neuronal

También existen teorías que tratan de explicar el origen del aspecto fenomenológico de la conciencia en base a diversas propiedades físicas del sustrato neuronal que la produce. De entre las teorías que siguen esta línea de investigación destacan las basadas en la mecánica cuántica. La teoría más popular desarrollada en este sentido es la defendida por Hameroff y Penrose (1996), que se basa en fenómenos cuánticos observados en los microtúbulos que hay en el cerebro.

La búsqueda de explicaciones basadas en la mecánica cuántica se produjo, entre otros motivos, por la aparente incapacidad de la neurociencia convencional para dar una explicación al problema de la conciencia. La propuesta original de Hameroff y Penrose para dar una explicación a la conciencia se basaba en la idea de que algunos aspectos de la física cuántica, como por ejemplo la coherencia cuántica, son esenciales para la conciencia. De hecho, estos autores proponen un mecanismo concreto como responsable de la generación de la conciencia: una nueva forma de reducción de la onda cuántica

denominada OR (*objective reduction*) (Penrose 1994), que tiene lugar en algunas estructuras intracelulares de las neuronas como los microtúbulos del citoesqueleto. Esta afirmación dio lugar a la teoría conocida como Orch-OR (Reducción del Objetivo Orquestado) (Hameroff, Penrose 1996).

Las características particulares de los microtúbulos los hacen especialmente adecuados para producir efectos cuánticos. Según la teoría Orch-OR, las tubulinas (proteínas que forman los microtúbulos) están asociadas a sucesos cuánticos internos e interactúan (computan) con otras tubulinas. La superposición coherente macroscópica de los estados de las tubulinas emparejadas cuánticamente tiene lugar a lo largo de áreas cerebrales de tamaño significativo, proporcionando la unidad global que requiere la conciencia (ver Apartado 2.1.3.2).

Según Hameroff y Penrose, la emergencia de la coherencia cuántica en los microtúbulos corresponde al procesamiento pre-consciente (por un periodo de hasta 500 milisegundos), hasta que la diferencia masa-energía entre los diferentes estados de las tubulinas alcanza un umbral relativo a la gravedad cuántica. De acuerdo con la teoría OR de Penrose, cada uno de los estados superpuestos tiene sus propias geometrías espaciotemporales. Cuando el grado de diferencia de masa-energía coherente lleva a una separación suficiente de la geometría espaciotemporal, el sistema tiene que decidir entre reducirse o colapsar a un estado simple. De esta forma, se produce una superposición temporal de geometrías ligeramente diferentes hasta que una reducción cuántica clásica tiene lugar abruptamente. A diferencia de la reducción subjetiva aleatoria (SR o R) de la teoría cuántica estándar, que es causada por la observación, la teoría Orch-OR sugiere que la reducción que tiene lugar en los microtúbulos es un auto-colapso, que provoca patrones concretos de estados que regulan las actividades neuronales (incluyendo las funciones sinápticas). Las probabilidades de que se produzcan estados post-reducción en las tubulinas están determinadas por factores como el acoplamiento de otras proteínas asociadas a los microtúbulos, que actúan como nodos que modulan u "orquestan" las oscilaciones cuánticas.

Más recientemente, Hameroff (2009) ha propuesto una nueva metáfora, denominada el "*piloto consciente*", basada en la teoría Orch-OR para explicar la conciencia. De acuerdo con este modelo, las funciones cognitivas del cerebro, incluido el procesamiento sensoriomotor, se entienden como "neurocomputación" que tiene lugar en las redes sinápticas. Cuando la neurocomputación cognitiva se realiza de forma consciente está acompañada de una sincronía gamma. Esta sincronía observada en el EEG se produce en gran medida gracias a grupos de neuronas que están enlazadas entre ellas por uniones entre dendritas, formando una red dendrítica temporal (*dendritic web*) que opera en paralelo al flujo axonal-dendrítico. Como las uniones dendríticas se abren y cierran, la red dendrítica que produce la sincronía gamma puede cambiar su topología y moverse como una especie de "envoltura espaciotemporal" que crea una integración y que se correlaciona con la conciencia. El "piloto consciente" es una descripción metafórica de la red dendrítica móvil sincronizada en la banda gamma como vehículo para una agente (o piloto) consciente que experimenta y asume el control de lo que de otra forma sería neurocomputación inconsciente.

2.2 Conciencia Artificial

El campo de la Conciencia Artificial, conocido en inglés por la denominación "*Machine Consciousness*"[3], vive actualmente un momento de creciente interés propiciado, en gran medida, por los avances en los conocimientos sobre la conciencia humana descritos en el apartado anterior. En cualquier caso, se trata de una disciplina muy joven, con marcado carácter multidisciplinar y que debido a su juventud cuenta con escasas aportaciones científicas o ingenieriles de gran trascendencia. No obstante, el potencial teórico de las líneas de investigación incluidas bajo el paradigma de la Conciencia Artificial es muy grande, incluyendo retos de tal envergadura como la contribución al entendimiento de la mente humana y la construcción de robots conscientes.

2.2.1 Introducción

Hasta fechas recientes, los avances científicos logrados en el estudio de la conciencia habían tenido una influencia modesta en los sistemas artificiales de inspiración biológica. En el marco de la Inteligencia Artificial (IA), los trabajos explícitamente orientados a la aplicación de modelos de conciencia eran relativamente escasos (Aleksander 2005). Uno de los principales motivos de la poca influencia en la IA de los avances científicos sobre la conciencia es la propia naturaleza de algunos de estos postulados. Los modelos basados en la mecánica cuántica (Apartado 2.1.5.2) o en los efectos relativísticos (Rakovic 1997) son difícilmente aplicables en entornos informáticos convencionales. Otro motivo, es el desconocimiento generalizado en el ámbito ingenieril de las teorías y modelos desarrollados en otras disciplinas directamente relacionadas con las ciencias cognitivas y las neurociencias. Si bien es cierto que la idea de crear máquinas conscientes es tan antigua como la humanidad, el campo científico de la Conciencia Artificial, enmarcado en la IA, prácticamente acaba de comenzar su andadura: durante la última década se viene celebrando con periodicidad anual una reunión internacional de expertos sobre la materia. En vista de las líneas de investigación presentadas en estas reuniones, en 2009 se creó la primera revista internacional dedicada exclusivamente a la Conciencia Artificial[4] (Chella 2009).

Las principales áreas que cubre el campo de la Conciencia Artificial son (Gamez 2008):

- La construcción de máquinas que desarrollen comportamientos asociados con la conciencia.
- La construcción de máquinas con capacidades cognitivas asociadas a la conciencia.

[3] La denominación "*Artificial Consciousness*" también se usa en la literatura de lengua inglesa para referirse a este campo de investigación, sin embargo parece que la expresión "*Machine Consciousness*" se ha convertido finalmente en la más popular.

[4] *International Journal of Machine Consciousness*.

- La construcción de máquinas basadas en arquitecturas que se cree que pueden producir (o al menos son necesarias para la producción de) conciencia.
- La construcción de máquinas fenomenológicamente conscientes.

El término máquina se usa aquí para referirse a cualquier implementación artificial y no está limitado a las máquinas físicas, sino que se incluyen en esta categoría los agentes software desarrollados bajo los mismos principios.

De las cuatro áreas enumeradas anteriormente, es la dimensión fenomenológica de la conciencia la más incierta en los estudios científicos y, por lo tanto, la que también presenta más dificultades a la hora de diseñar modelos computacionales. Por el contrario, se considera que es más viable la aplicación de modelos funcionales de la conciencia basados en aspectos cognitivos.

De hecho, los aspectos de acceso de la conciencia son muy interesantes en cuanto a su posible aplicación en sistemas artificiales. La conciencia puede ser considerada como una especie de pasarela que brinda un acceso resumido a una gran cantidad de contenidos de la mente. De acuerdo con los estudios neurológicos, en cada momento hay un gran número de procesos neuronales inconscientes ejecutándose en paralelo; sin embargo, sólo ciertos contenidos se muestran a la conciencia en un instante dado. Este "cuello de botella" en el acceso a los contenidos de la mente es útil para gestionar, por ejemplo, una percepción visual efectiva (Sperling 1960).

Aunque el aspecto de la evaluación del nivel de conciencia no suele considerarse como una de las áreas principales de este campo (Aleksander 2005, Gamez 2008), es sin duda un punto muy importante que requiere de mayor atención.

El trabajo en las áreas de investigación mencionadas anteriormente se realiza como medio para alcanzar los siguientes objetivos:

- Alcanzar un mejor entendimiento de lo que es la conciencia (confirmar o refutar, entre otras, las hipótesis planteadas en el apartado 1.1).
- Comprender cómo se genera la conciencia en los seres humanos.
- Descubrir y caracterizar el tipo y nivel de conciencia existente en otras criaturas (animales y máquinas).
- Construir máquinas o robots capaces de realizar tareas complejas, que involucren, por ejemplo, una interacción natural con los humanos.

Actualmente, el campo de la Conciencia Artificial está centrado principalmente en el diseño e implementación de modelos de conciencia, la mayoría de ellos basados en avances provenientes de las neurociencias. Este tipo de trabajos abarca desde modelos puramente funcionales, donde lo importante para la atribución de conciencia es el comportamiento resultante, hasta modelos basados en información detallada sobre la anatomía y dinámica del cerebro humano.

Desde el punto de vista de la funcionalidad, el objetivo último de las implementaciones basadas en Conciencia Artificial es resolver un determinado problema de tal forma que el usuario humano no pueda distinguir si ese problema lo ha resuelto otro humano o una máquina (Aleksander 2005). Se trata en definitiva de construir máquinas capaces de pasar versiones adaptadas de la famosa prueba de Turing (Turing 1950, Harnad 1994).

Desde el punto de vista fenomenológico, el diseño y desarrollo de modelos de la fenomenología – área normalmente denominada "*fenomenología sintética*" – trata de ahondar en la comprensión de los qualia y cómo se generan las sensaciones de la experiencia subjetiva (Chrisley 2009). La fenomenología sintética tiene como principal objetivo la caracterización de los estados fenomenológicos poseídos o modelados en agentes artificiales.

Desde el punto de vista de la filosofía de la mente, Searle indica que el estudio de los sistemas cognitivos es el estudio de la conciencia, de igual forma que el estudio de la biología es el estudio de la vida (Searle 1992). Por conciencia Searle no se refiere exclusivamente a la subjetividad, sino que hace referencia a toda la extensión y riqueza de la conciencia, dejando patente la categórica relación entre todos los procesos mentales y la conciencia. Por procesos mentales se refiere a la percepción, el aprendizaje, la inferencia, toma de decisiones, resolución de problemas, emociones, etc. Incluso yendo más allá, y pensando en la posibilidad de diseñar un modelo de la mente, Searle recuerda que no hay que olvidar que todas las características que los filósofos han definido como exclusivas de la mente son también dependientes de la conciencia: subjetividad, intencionalidad, racionalidad, libre albedrío y causalidad mental.

En resumen, según Aleksander (2005) y Sanz et al. (2007), el estudio de la conciencia mediante el diseño de máquinas puede proporcionar dos resultados principales:

- Un lenguaje común en términos computacionales que permita expresar con exactitud el concepto de conciencia.
- La definición de un amplio conjunto de métodos computacionales para la construcción de una nueva generación de máquinas con capacidades avanzadas gracias a la flexibilidad de los modelos de la conciencia.

Como se ha comentado en el capítulo de introducción, su carácter marcadamente multidisciplinar es una de las características principales de la Conciencia Artificial, pero también fuente de confusión en cuanto a la propia definición de este campo de investigación. La Conciencia Artificial toma como fuente de inspiración conocimiento proveniente de la filosofía, la psicología, la inteligencia artificial, las neurociencias, etc., pero sus objetivos siguen siendo los descritos anteriormente y no los establecidos para cualquiera de estas otras disciplinas con las que se relaciona. Es preciso tener claro este punto para no confundir el rumbo que deben tomar las investigaciones en el campo de la Conciencia Artificial.

En general, un programa de investigación en Conciencia Artificial debe distinguir entre dos modelos de conciencia diferenciados: un modelo de la conciencia humana, que reflejará en la medida de lo posible el conocimiento existente sobre la "conciencia natural", y otro modelo de conciencia computacional (posiblemente más limitado) que se basará en el anterior, pero deberá estar orientado a un enfoque pragmático que permita su implementación. La experimentación con el sistema artificial puede proporcionar retroalimentación útil para la redefinición del modelo de conciencia utilizado. Asimismo, el diseño de experimentos también puede estar sujeto a una redefinición guiada por los resultados preliminares obtenidos (ver Figura 4).

Una de las principales hipótesis de trabajo que se mantienen en el campo de la Conciencia Artificial consiste en la idea de que si se puede obtener un modelo computacional lo suficientemente preciso de los mecanismos que dan lugar a la

conciencia en los humanos, se podría aplicar este modelo para construir máquinas conscientes (Hernández, López & Sanz 2009).

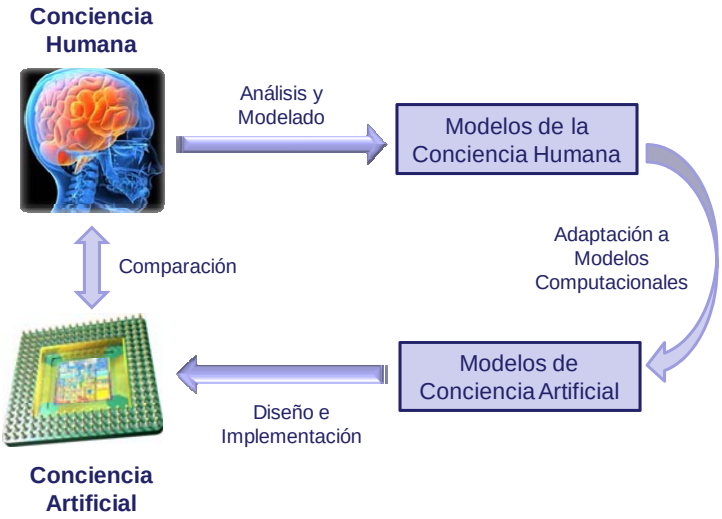


Figura 4. Conciencia Artificial.

En ámbitos como la visión artificial o el reconocimiento del habla se han seguido estrategias análogas. Sin embargo, en algunas ocasiones se obtienen mejores resultados con modelos que no imitan en absoluto los procesos biológicos. Por el contrario, en el caso de la Conciencia Artificial, la finalidad de conseguir modelos que se aproximen en la mayor medida posible a la realidad no es sólo el rendimiento, sino que también se pretende comprender mejor cómo se produce la conciencia en el cerebro humano.

Como se ha mostrado en el Apartado 2.1, los modelos de la conciencia se pueden describir a diferentes niveles de descripción: físico, fisiológico, funcional o cognitivo y mental. De forma análoga, los principales proyectos de Conciencia Artificial realizados hasta la fecha se pueden clasificar atendiendo a su complejidad y el nivel de descripción en el que se basan.

En los siguientes apartados se presenta un breve repaso de los proyectos de Conciencia Artificial más relevantes. En primer lugar se analizan las propuestas más simples desde el punto de vista arquitectural: los sistemas basados primordialmente en redes de neuronas artificiales (Apartado 2.2.2). Posteriormente, se analizan los sistemas que combinan diferentes técnicas de IA como base de un modelo computacional de la conciencia (Apartado 2.2.3). Finalmente, se presenta un estudio de las principales arquitecturas cognitivas basadas en modelos de la conciencia (Apartado 2.2.4).

2.2.2 Modelos basados en redes de neuronas artificiales

Típicamente las redes de neuronas artificiales (RNA) se han empleado para resolver diferentes problemas: reconocimiento de patrones, predicción de secuencias temporales, etc., entre los que no se encontraba usualmente la aplicación de modelos de Conciencia Artificial. Sin embargo, algunas RNA están inspiradas en mecanismos observados en el sistema nervioso humano. Inspiradas en los resultados de la neurociencia moderna, que

indica que la conciencia se produce en el cerebro, se han desarrollado diversas líneas de investigación encaminadas a la creación de modelos computacionales de la conciencia basados en RNA.

Del estudio detallado de los mecanismos presentes en el sistema nervioso central humano, se desprende que los modelos comunes de RNA no reflejan muchas de las características presentes en los sistemas biológicos (Holmes 2002). Algunas de estas características podrían ser esenciales para la creación de modelos de conciencia plausibles. Por ejemplo, en el tejido nervioso existen interconexiones por medio de ondas cerebrales (oscilaciones rítmicas) en diferentes bandas de frecuencias (ver Apartado 2.1.3.2). Aunque estas ondas cerebrales se correlacionan fisiológicamente y psicológicamente con funciones cerebrales, normalmente no son tenidas en cuenta en los modelos de RNA convencionales.

Otro problema patente de las RNA es su tamaño en comparación con el del cerebro humano (Buttazzo 2001): 10^{12} neuronas con 10^{15} sinapsis es un tamaño que está todavía lejos de poder ser replicado en un único ordenador (se necesitaría un mínimo de un petabyte de memoria principal). En este sentido, nuevas disciplinas como la ingeniería neuronal, la nanotecnología y la "neuroarquitectrónica" pueden ser opciones que impulsen el desarrollo de nuevos dispositivos de tipo neuronal (Lyshevsky 2002). Actualmente, proyectos como *Blue Brain* (Markram 2006), tratan de aplicar los principios de ingeniería inversa al estudio del cerebro de los mamíferos, usando técnicas de simulación precisa para comprender los mecanismos de la función cerebral. Sin embargo, este tipo de iniciativas se centran en el nivel neurobiológico, y aunque pueden contribuir significativamente a la comprensión del funcionamiento del cerebro a nivel neuronal, no es inmediato prever que puedan tener un impacto similar en cuanto a la comprensión de los procesos cognitivos.

Típicamente, tanto la estructura como los algoritmos de aprendizaje de la RNA simplifican en gran medida el funcionamiento real de la redes de neuronas de un cerebro biológico: por un lado, la activación de una neurona no es una suma aritmética de los efectos sinápticos, sino que es un proceso complejo, en el que intervienen diferentes periodos de tiempo correspondientes a los potenciales de acción de cada sinapsis; por otro lado, el sistema nervioso de un mamífero se compone de muchas clases diferentes de neuronas y de células gliales. La neuroglia (o conjunto de células gliales) no sólo proporciona apoyo, sino que también procesa señales y tiene sinapsis como las neuronas. Además, los potenciales de acción no determinan las señales nerviosas por sí solos. Otros potenciales y la señalización hormonal también juegan un papel importante, de igual forma que lo hacen los diferentes neurotransmisores.

Aunque el nivel de complejidad descrito anteriormente no está presente en las RNA actuales, esto no significa necesariamente que no se puedan construir modelos computacionales (limitados) de la conciencia, o al menos de determinados aspectos de la misma. De hecho, teniendo presentes estas limitaciones, algunos modelos basados en RNA pretenden reproducir mecanismos cognitivos asociados a la conciencia como la capacidad de atención. Por ejemplo, usando algoritmos de tipo *winner-takes-all* ("el ganador se lleva todo"), la activación de una neurona determinada establece el foco de atención. Es decir, la neurona ganadora establece el contenido que se procesará de forma explícita o "consciente". Este tipo de mecanismos son equivalentes a los procesos que realizan las neuronas de representación del contenido según algunos modelos de las estructuras corticales humanas (Towaga, Otsuka 1998).

Otros modelos basados en RNA definen mucho más nítidamente la diferencia entre contenido consciente e inconsciente (Kozma 1997). La idea básica de estos modelos es que la representación del conocimiento en una RNA se debe realizar a dos niveles distintos: el consciente y el inconsciente. Esta teoría se conoce como el paradigma local-global, refiriéndose esta contraposición a la forma de describir la conciencia en términos de los procesos de aprendizaje en RNA. La implementación de este enfoque se lleva a cabo mediante un algoritmo de aprendizaje estructurado, que se aplica en redes auto-organizadas con aprendizaje supervisado. La red de neuronas se divide en dos estructuras relacionadas, una se ocupa del conocimiento "consciente" y la otra del "inconsciente". Las dimensiones del espacio de problema han de reducirse en la parte consciente mediante una simplificación basada en la extracción de reglas, mientras que la parte subconsciente de la red mantiene todas las dimensiones del problema.

También se han utilizado otro tipo de modelos de conciencia más simplificados con el objetivo concreto de mejorar las capacidades de reconocimiento de patrones de una RNA (Kim 1997). Estos modelos proponen un "almacén central" o memoria jerárquica donde se guardan las "imágenes mentales" obtenidas mediante un algoritmo de reconocimiento. Este algoritmo se encarga de comparar los nuevos ejemplos presentados a la red con las imágenes mentales existentes, y si la comparación resulta positiva el ejemplar es reconocido, en caso contrario se crea una nueva imagen mental.

La mecánica cuántica es considerada por algunos autores como la clave en el estudio de la conciencia (ver Apartado 2.1.5.2). La aplicación de los principios de la mecánica cuántica en las RNA ha dado lugar a una disciplina conocida como Redes de Neuronas Cuánticas (RNC), que incluso cuenta con propuestas de implementación hardware (Behrman, Steck & Skinner 1999). Las RNC se conciben a menudo como una mejora de las RNA clásicas en la resolución de problemas relacionados con el reconocimiento de patrones (Zhou 2003). Desde el punto de vista computacional, existen dos motivos principales por los cuales se desea aplicar la computación cuántica a las RNA: para compensar el límite de desarrollo del hardware tradicional basado en el silicio (que supuestamente alcanzará su límite alrededor del año 2030 (Buttazzo 2001)) y para producir una capacidad computacional no disponible usando los enfoques neuronales convencionales.

La pregunta clave sobre este asunto es la siguiente: ¿es capaz la computación cuántica (y más concretamente las RNC) de proporcionar soluciones a los problemas que hasta ahora las RNA no han podido solucionar? Entre estos problemas se encuentra el diseño e implementación de modelos de conciencia. Parece que no hay una opinión unánime a este respecto. Si bien es cierto que se ha demostrado que la computación cuántica es superior a la computación clásica para determinados problemas, al menos en lo que respecta a la reducción de la complejidad algorítmica (Ventura 2001), no está claro que los procesos cuánticos sean necesarios para la aparición de la conciencia (Tegmark 2000). Incluso se ha especulado recientemente con la posibilidad de que el cerebro pueda realizar procesos "quasi-cuánticos", es decir, realizar computación basada en los principios de la mecánica cuántica (superposición, emparejamiento y colapso), pero utilizando mecanismos convencionales, que no requieren condiciones físicas exóticas (Haikonen 2010).

Aunque los modelos de la conciencia basados en RNA son apropiados para la investigación de los procesos que supuestamente dan lugar a la conciencia a nivel neurobiológico, normalmente no permiten un estudio arquitectónico y funcional de nivel superior. Dado que la presente tesis se centra en el diseño de arquitecturas

cognitivas artificiales, no se ha abordado específicamente el diseño de RNA, aunque se consideran componentes potenciales dentro de una arquitectura cognitiva.

2.2.3 Sistemas híbridos

Los modelos de Conciencia Artificial híbridos se caracterizan porque están basados en más de un único enfoque o técnica computacional. En realidad, la mayoría de las arquitecturas cognitivas más desarrolladas emplean diversas técnicas simultáneamente (o aplican la misma técnica, pero dos o más formas diferentes). Una estrategia básica para la organización de diferentes mecanismos en un mismo sistema es la estructura en capas o niveles. Por ejemplo, en la arquitectura CLARION los procesos cognitivos se sitúan en dos niveles, cada uno gobernado por mecanismos diferentes (Sun 1997). CLARION está construido en base a dos componentes principales: uno para el procesamiento inconsciente o implícito y otro para el procesamiento consciente o explícito. Ambos están implementados con redes de neuronas. Sin embargo, unas redes se ocupan de contenidos procedimentales (inconscientes), mientras otras se encargan del conocimiento declarativo (consciente).

Sun hace hincapié en las diferencias existentes entre el aprendizaje implícito y explícito y el papel que juega la conciencia en la combinación de ambos. Según Sun, cada tipo de conocimiento requiere un proceso de aprendizaje específico. Los humanos pueden aprender conocimiento procedimental a través del método de ensayo y error (sin tener conocimiento previo). El conocimiento declarativo, que reside en un nivel superior, se puede adquirir a través de la experiencia continuada en el mundo. Es importante que el conocimiento declarativo se aprenda a través del razonamiento acerca de las aptitudes (conocimiento procedimental) de más bajo nivel. Dada la naturaleza implícita del conocimiento procedimental, el aprendizaje de este conocimiento se suele realizar de forma inconsciente. El conocimiento procedimental representado de este modo no es interpretable. En CLARION cada nivel codifica un conjunto completo de conocimiento para su procesamiento. Estos dos conjuntos de conocimiento se solapan en gran medida, por lo que los resultados de ambos han de combinarse. Según Sun se produce una sinergia entre el procesamiento implícito (“inconsciente”) y el procesamiento explícito (“consciente”).

Otra combinación típica en el ámbito de la IA consiste en el uso conjunto de RNA y técnicas de computación evolutiva (Weiß 1994a). Estos modelos se basan en los principios de mutación y selección que la naturaleza ejerce sobre la evolución de las redes de neuronas biológicas. En relación a estos mecanismos se han propuesto varias teorías que combinan RNA con computación evolutiva, como la teoría de selección de grupos de neuronas (ver Apartado 2.1.3.4) o la teoría de los circuitos de aprendizaje evolutivos (Weiß 1994b). De acuerdo con estas teorías, la evolución ha sido la clave de la aparición de los procesos mentales en los humanos. Por lo tanto, la aplicación de modelos híbridos que combinan RNA con técnicas de computación evolutiva podría posibilitar la investigación sobre la reproducción de este tipo de procesos mentales en máquinas.

Las redes de neuronas biológicas tienen una estructura directamente relacionada y adaptada a la función que desempeñan. Se sabe muy poco sobre cómo se puede diseñar una determinada estructura de red para que realice una función dada. De hecho, el diseño de red sigue siendo un asunto crítico y de interés en el ámbito de las RNA. Dado

que el espacio de búsqueda sobre todas las posibles estructuras de red es infinitamente grande y poco manejable, parece indicado el uso de algoritmos evolutivos para el diseño de RNA. En cualquier caso, la aplicación de este tipo de enfoques en Conciencia Artificial es muy compleja, pues no está claro cómo ha de definirse la función de adecuación (*fitness function*). Es decir, se necesitan establecer a priori los procesos que supuestamente pueden dar lugar a la conciencia.

Otras propuestas de sistemas híbridos combinan RNA con procesamiento simbólico. Por ejemplo, en un robot mascota que debe mostrar comportamientos afectivos y aparentemente conscientes se combinan RNA con Sistemas Basados en Reglas (SBR) (Kubota, Kojima & Fukuda 2001). Mientras que las RNA proporcionan una inferencia basada en el procesamiento numérico, las reglas borrosas proporcionan una inferencia lógica basada en el procesamiento simbólico. Un modelo basado en la conciencia se encarga de seleccionar y combinar la salida de ambos subsistemas. En definitiva, el SBR constituye un módulo de toma de decisiones, mientras que la RNA puede formar parte del módulo de percepción y emociones. El modelo de conciencia combina estos módulos para generar un comportamiento equivalente al de un ser consciente. El modelo propuesto por Kubota distingue entre la toma de decisiones conscientes e inconscientes. Sólo cuando el sistema alcanza un determinado nivel de conciencia puede dejar de realizar una acción automática (inconsciente) y cambiar de comportamiento.

En los SBR se ha propuesto la introducción de operadores lógicos específicos relacionados con ciertos aspectos de la conciencia (Pietarinen 2002). Utilizando la lógica modal se exploran las posibles conexiones entre la lógica y la neurociencia. La particularidad de este enfoque es que se añade la lógica epistémica como otra disciplina adicional desde la que se puede estudiar la conciencia. Los defensores de este enfoque argumentan que la lógica puede utilizarse para representar fenómenos neurológicos, utilizándose el concepto de conciencia para distinguir entre el procesamiento de información implícito y explícito. En esta "lógica de la conciencia" se dice que un agente es consciente de una preposición *p* en aquellas situaciones en las que se mantiene un valor de verdad definido para la preposición. Es decir, un agente no puede ser consciente de preposiciones para las que el valor de verdad es indefinido. La conciencia funcionaría aquí como un operador que aplicado al conocimiento implícito lo convirtiese en conocimiento explícito.

Un punto interesante en relación a los sistemas híbridos basados en la conciencia es que podrían servir de base para resolver la rivalidad existente entre el enfoque simbólico y el subsimbólico (Wei, Wu & Chen 2000). Los modelos de conciencia podrían constituir un nexo que diera lugar a un enfoque que combine sólidamente estos dos paradigmas. En la presente tesis, se pretende abordar este problema mediante el uso de arquitecturas cognitivas artificiales.

2.2.4 Arquitecturas cognitivas artificiales

Un punto importante que es preciso destacar en cuanto a las implementaciones inspiradas en teorías cognitivas es que éstas a menudo proporcionan simplemente una metáfora que ayuda a entender de forma intuitiva el funcionamiento de la conciencia, pero no dan suficientes detalles como para realizar directamente un diseño computacional. Por lo tanto, uno de los principales retos de los modelos cognitivos de

Conciencia Artificial es diseñar los mecanismos concretos que den lugar a la funcionalidad deseada. Además, debido a que la descripción proporcionada por las teorías cognitivas es a menudo muy abstracta, es normal que existan diferentes implementaciones inspiradas en una misma teoría. La arquitectura propuesta en el Capítulo 4 es un ejemplo de arquitectura inspirada en diversas teorías cognitivas de la conciencia.

El nivel de complejidad también es variado en cuanto a las diferentes implementaciones existentes. Además, al igual que ocurre en el estudio de la conciencia en los humanos, conviene distinguir entre los diferentes aspectos o dimensiones de la misma que se pretenden reproducir. Los sistemas artificiales que tratan de emular las capacidades cognitivas de los humanos se inspiran en teorías cognitivas como las descritas en el Apartado 2.1.4. La realización de este tipo de sistemas se basa en arquitecturas software capaces de proporcionar una implementación de los modelos cognitivos asociados. A estas arquitecturas se les denomina arquitecturas cognitivas artificiales.

En el ámbito de la IA se han desarrollado, con cierto éxito, muchas arquitecturas cognitivas. Algunos ejemplos relevantes son ACT (Anderson 1993), ACT-R (Anderson, Matessa & Lebiere 1997), SOAR (Laird, Newell & Rosenbloom 1987), ART (Grossberg 1987), CLARION (Sun 2006), IDA (Franklin, Kelemen & McCauley 1998), LIDA (Ramamurthy et al. 2006, Franklin et al. 2007b), CogPrime (Goertzel 2009), EPIC (Kieras, Meyer 1997), Arquitectura Cognitiva de Haikonen (Haikonen 2007b) e Icarus (Langley, Choi 2006). Asimismo, el paradigma de los agentes BDI (*Belief-Desire-Intention*) constituye en sí mismo un enfoque cognitivo (Rao, Georgeff 1991, Rao, Georgeff 1995), cuyas implementaciones derivadas pueden considerarse arquitecturas cognitivas.

Cada una de las arquitecturas cognitivas existentes cubre ciertos aspectos específicos de la cognición humana. Asimismo, cada implementación se basa en técnicas concretas, normalmente provenientes de la IA, como las RNA o los sistemas basados en reglas. Sólo algunas de estas arquitecturas están inspiradas específicamente en modelos de la conciencia humana. En este análisis se incluyen sólo aquellas arquitecturas cuya principal fuente de inspiración es la conciencia y los procesos cognitivos asociados.

2.2.4.1 Arquitecturas basadas en un ETG

En el enfoque propuesto por Franklin en la arquitectura IDA (Franklin, Kelemen & McCauley 1998), se llama agente cognitivo con "conciencia funcional" a un agente software que presenta las siguientes características: percepción, memoria a corto y largo plazo, atención, planificación, razonamiento, resolución de problemas, aprendizaje, emociones, estados de ánimo, actitudes, etc. Se espera que los agentes que presenten estas características sean más flexibles, adaptativos y en definitiva más humanos gracias a su capacidad de aprender y tratar con situaciones desconocidas. IDA se basa en la teoría del Espacio de Trabajo Global o ETG (Baars 1988) (ver Apartado 2.1.4.1) y es una de las implementaciones de Conciencia Artificial más conocida.

IDA es un sistema de distribución de mensajes desarrollado para la marina de los Estados Unidos de América. IDA se diseñó para asignar tareas a los marineros al final

de su turno de trabajo. Para desarrollar esta misión, IDA tenía que poder desarrollar una conversación en lenguaje natural (usando el correo electrónico), acceder a diversas bases de datos, ajustarse a las políticas de la marina y también comprobar los requisitos de cada trabajo, el coste asociado y las preferencias del marinero. Todas estas funciones las desempeñaban un gran número de *codelets*[5] – pequeños programas especializados – gestionados mediante una arquitectura basada en un ETG.

En IDA el espacio de trabajo global, y por ende el modelo de conciencia, estaba basados en los siguientes componentes:

- Gestor de coaliciones entre los *codelets*. Sistema que permite la formación de coaliciones o asociaciones entre diversos *codelets* para que la información conjunta que proporcionan pueda competir para aparecer bajo el foco de atención.
- Controlador del foco de luz (foco de la atención). Módulo que establece el contenido específico que se considera "consciente" y por lo tanto se difundirá a toda la "audiencia" del ETG.
- Gestor de difusión de mensajes. Módulo que permite la distribución global (a todos los *codelets*) de los contenidos que son seleccionados bajo el foco de atención.
- *Codelets* de atención. Programas que comprueban constantemente ciertos criterios en la información sensorial de entrada en busca de condiciones que puedan requerir una intervención consciente.

Los *codelets* de atención de IDA son programas especializados que monitorizan la ocurrencia de un determinado suceso que puede requerir una intervención consciente. Cuando tal suceso tiene lugar, estos *codelets* forman coalición con otros *codelets* que contienen información sobre la situación. La coalición compite entonces para situarse bajo el foco de atención (*spotlight*). Si la coalición gana, sus contenidos se difunden llegando al resto de *codelets*. Alguno de los *codelets* que recibe la información puede decidir que ha encontrado una acción que resuelve la situación y por lo tanto competir para que esa acción se ejecute. En IDA la selección de comportamientos se hace en base a una serie de "impulsos" (*drives*) definidos. Aplicando este mecanismo se da un nivel de activación mayor a aquellos comportamientos que satisfacen los impulsos.

El modelo IDA también contempla un mecanismo de deliberación que explora diferentes escenarios y elige el mejor. La arquitectura cognitiva de IDA contempla además el concepto de emociones. Por ejemplo, la frustración aparece cuando el sistema no es capaz de comprender un mensaje de entrada enviado por un marinero.

Actualmente, el modelo computacional de conciencia definido en la arquitectura de IDA está siendo ampliado en un nuevo sistema denominado LIDA (*Learning IDA*). Este nuevo modelo computacional incluye nuevos mecanismos de aprendizaje y se pretende aplicar en nuevos entornos como robots autónomos (Franklin et al. 2007b). El modelo computacional propuesto en LIDA todavía no ha sido completamente implementado (Franklin, comunicación personal), si bien los autores han especulado acerca de su capacidad para desarrollar conciencia fenomenológica (Ramamurthy 2008).

[5] En español el término *codelet* se podría traducir como "programita" o "pequeña aplicación software".

Existen otras arquitecturas basadas también en la teoría del ETG. Moura y Bonzon (2004) se basan en el modelo de Baars para construir un sistema de agentes inteligentes que presenten algunas de las capacidades cognitivas asociadas con la conciencia. También Dehaene et al. (2003) han creado un sistema simulado para estudiar cómo el espacio de trabajo global y los procesadores especializados interactúan durante la tarea *Stroop*. La tarea Stroop consiste en presentar al sujeto una serie de tarjetas para que este diga tanto el nombre del color que está escrito en la tarjeta como el color de la tinta con la que está escrito el nombre del color. Esta tarea es más complicada de realizar cuando el color de la tinta no coincide con el nombre del color escrito, por ejemplo la palabra "rojo" escrita con tinta azul. El modelo propuesto por Dehaene et al. está compuesto por redes de neuronas que forman un ETG y sistemas adicionales de monitorización y recompensa que modulan la actividad del ETG. Las simulaciones realizadas con este modelo sugieren que las tareas fáciles se pueden llevar a cabo directamente con la ayuda de los procesadores especializados, sin mucha activación del ETG. Sin embargo, las tareas más complejas, como decir el color de la tinta cuando este no coincide con el nombre escrito, sólo se pueden realizar gracias a la activación del ETG y usando los sistemas de monitorización y recompensa para corregir los errores.

En trabajos más recientes (Dehaene, Sergent & Changeux 2003), los mismos investigadores han simulado el efecto conocido como "parpadeo atencional" (*attentional blink*). El parpadeo atencional tiene lugar en los humanos cuando dos imágenes parecidas se presentan rápidamente una después de la otra (típicamente entre 100 y 500 milisegundos después). En estos casos, suele ocurrir que el sujeto no detecta un cambio significativo realizado en la segunda imagen (Raymond, Shapiro & Arnell 1992). Usando el sistema simulado, Dehaene et al. (Dehaene, Sergent & Changeux 2003) han explicado por qué se produce este efecto en el cerebro humano. Cuando se presenta la primera imagen a un sujeto, la imagen accede al ETG gracias a las activaciones de largo alcance que tienen lugar entre diferentes áreas neuronales. Cuando el cerebro está en este estado, es mucho más difícil que la segunda imagen pueda difundir globalmente para acceder al ETG. Aunque las áreas locales del cerebro continúen procesando la información sensorial de la segunda imagen, esta información no es consciente porque no puede llegar a las áreas del cerebro responsables de la memoria y la comunicación.

Shanahan también ha propuesto recientemente una arquitectura cognitiva basada en la teoría del ETG (Shanahan 2006). Esta arquitectura se basa en módulos inspirados en estructuras funcionales del cerebro. Un bucle sensoriomotor de primer orden se encarga de dar respuestas motoras inmediatas, mientras que un bucle de segundo orden modula el comportamiento del bucle de orden inferior de acuerdo a la relevancia de las acciones. El bucle de primer orden se cierra por medio de la interacción con el mundo, mientras que el bucle de segundo orden se cierra internamente a través de un área de asociación. Este sistema de asociación simula el estímulo sensorial que seguiría a una acción motora de forma análoga a como lo hace la imaginación. Esta función de simulación se realiza usando una arquitectura basada en un ETG en el que las áreas de asociación reciben información de un módulo inspirado en el ganglio basal humano. En las áreas de asociación se produce una competición para seleccionar que información se difunde globalmente. Esta arquitectura proporciona un mecanismo para calcular cadenas de asociaciones y por lo tanto explorar las posibles consecuencias de las acciones antes de realizarlas.

2.2.4.2 Arquitectura cognitiva de Haikonen

Con un planteamiento bastante diferenciado de los basados en un ETG, Haikonen ha propuesto una arquitectura hardware basada en principios cognitivos (Haikonen 2007b). El modelo de Haikonen constituye una de las propuestas más completas para solucionar el problema de la Conciencia Artificial. Sin embargo, aunque Åberg y Rantala han desarrollado algunos microchips electrónicos basados en la arquitectura de Haikonen (Aberg, Rantala 2008), todavía no existe una implementación completa (Haikonen, comunicación personal). Haikonen argumenta que si se llegara a construir una máquina siguiendo los principios de esta arquitectura cognitiva, ésta podría tener conciencia. En la máquina teóricamente consciente descrita por Haikonen la percepción se realiza en base a "representaciones de señales distribuidas".

Doan describe la máquina de Haikonen de la siguiente forma (Doan 2009b): si las cámaras de la máquina de Haikonen están enfocadas hacia una pelota amarilla, el patrón de píxeles de las cámaras se convierte en la entrada de un circuito preprocesador que produce un vector de, por ejemplo, 10.000 señales. Se necesita un cable, o cualquier otro conductor, para transporta cada una de estas señales. Uno de los cables podría ser la salida del atributo "redondez" de la circuitería del preprocesador visual. En este caso, la señal indica "encendido" (*on*). Otro cable de entre estos 10.000, el correspondiente al atributo "cuadradez" estaría "apagado" (*off*) – teniendo un voltaje nulo, por ejemplo. Otro grupo de cables de entre este conjunto de 10.000 se correspondería con el análisis del espectro de frecuencias. El cable correspondiente a la frecuencia que los humanos reconocemos como "amarillo" estaría encendido, mientras que el resto de los cables de este grupo – "rojo", "azul", etc. – estarían apagados. También existirían muchos otros grupos de cables describiendo atributos como tamaño, borde, brillo, etc.

La máquina no representa internamente la pelota como un gráfico redondo, tampoco como un conjunto de números que indican diámetro, color, etc. si no a través de este vector de señales. Haikonen llama a esto una representación basada en señales distribuidas. Si a la máquina se le enseñan diferentes pelotas de diferentes tamaños, colores, etc. y cada vez que esto sucede su micrófono "oye" un patrón de sonido que los humanos entendemos como la palabra "pelota", la máquina asocia el patrón de sonido al patrón visual. El proceso de percepción de esta máquina se basa en la construcción de este tipo de asociaciones. Para la máquina asociativa de Haikonen, el aprendizaje, la fijación en memoria y el acto de recordar son simplemente algunos aspectos diferentes de los muchos que tiene la única cosa que la máquina sabe hacer: establecer asociaciones.

Los vectores de señales producidos a partir de estímulos sensoriales, como la visión de la pelota o la pronunciación de la palabra "pelota", se difunden a un gran número de "grupos de neuronas", algunos de los cuales almacenarán el patrón de señal usando bien el *método de circulación* o bien el *método sináptico* (Haikonen 2007b). Según la descripción de Haikonen, esta arquitectura cognitiva es capaz de reproducir procesos como la producción del habla e incluso habla interior. Además, una máquina implementada siguiendo estos principios sería capaz de percibir su propio cuerpo y percibir su estado interno. Por ejemplo, cuando sus manos tocan su cuerpo se producen dos conjuntos de señales que permiten a la máquina deducir que lo que está tocando es su propia piel, en vez de algo externo. De igual modo, en un estadio temprano del desarrollo cognitivo de la máquina de Haikonen, cuando sus manos se mueven frente a sus cámaras, descubriría – al igual que pasa en los bebés cuando descubren sus manos y

sus pies – que el objeto percibido no pertenece al mundo exterior, sino que es una parte de su propio cuerpo.

La arquitectura de Haikonen también incluye un mecanismo para la imaginación. La máquina de Haikonen puede imaginar cosas que nunca ha visto. Por ejemplo, habiendo aprendido lo que es una pelota amarilla y habiéndola visto rodar cuesta abajo, la próxima vez que vea (o que piense en) una pelota amarilla, puede imaginar una pelota de color diferente o una pelota rodando cuesta arriba. Este proceso de imaginación se produce de la siguiente forma: la representación de señales distribuidas "pelota amarilla" evoca otra representación de señales distribuidas "pelota", pero la señal "amarilla" en el vector de señales se apaga y otra señal se enciende. Otra posible opción es que la secuencia "pelota rodando" se evoca desde memoria y entonces la máquina invierte la secuencia temporal.

Los seres conscientes experimentan un flujo de imágenes mentales. Haikonen argumenta que su arquitectura también es capaz de generar este tipo de procesos. Por ejemplo, cuando aparece el estímulo visual de la pelota amarilla, en la máquina emerge un patrón de señales entre las que se encuentran atributos activos como "redondez" y "amarillo". La activación de este patrón evoca otros patrones conocidos como el patrón de sonido de la palabra "pelota", el conocimiento de la "pelota rodando", la memoria visual de una "pelota azul" que ha sido vista con anterioridad, etc. A su vez, todas estas representaciones pueden evocar otras relacionadas: "rueda más deprisa bajando una cuestas más empinada", la imaginación de "la pelota rodando cuesta arriba", la imaginación de "varias pelotas rodando juntas", etc.

Las emociones también se contemplan en el modelo planteado por Haikonen. En un momento dado, están activas un gran número de representaciones de señales distribuidas debido a información proveniente de sensores, evocada o imaginada. La máquina deberá prestar más atención a la posibilidad de que la pelota sea lanzada hacia ella misma, en vez de imaginar pelotas púrpuras inexistentes que ruedan cuesta arriba. Pero, ¿cómo sabe eso la máquina? La respuesta está en las emociones: Haikonen establece las bases para las emociones como una combinación de reglas globales y reglas a nivel neuronal.

Las reglas a nivel neuronal son denominadas *coincidencia*, *incongruencia* y *novedad*. Si la máquina está observando una pelota amarilla y no pasa nada más, el grupo de neuronas que procesa este vector de señales produce una señal de salida *coincidencia* (además del propio vector de señales de salida). Esto sucede porque el vector de señales que alimenta las entradas asociativas representa la pelota amarilla vista una fracción de segundo antes; por lo tanto, la entrada asociativa coincide con el vector de entrada principal. Si la pelota empieza a alejarse, el nuevo vector de entrada es diferente del patrón de la entrada asociativa (la pelota se ve más pequeña). En este caso se genera una señal *incongruencia*, por lo menos durante unos instantes, atrayendo la atención de la máquina. Si después la máquina mira una caja cuadrada que no ha visto nunca, entonces la entrada asociativa no contiene nada con lo que comparar y por lo tanto se produce una señal *novedad* en la salida del grupo de neuronas.

Aparte de las señales *coincidencia*, *incongruencia* y *novedad*, la máquina también tiene reglas para construir su capacidad de saber qué es el dolor y qué es el placer, además de saber qué es bueno y qué es malo. Algunas reglas son de la forma: "el dolor es malo", "daños en los sensores de la piel es dolor" y "evitar cosas malas". Otro conjunto de reglas, que se evocan cuando la batería se está cargando son: "un aumento en el nivel de carga de la batería es placer", el placer es bueno" y "si siento placer,

seguir haciéndolo". En la naturaleza se supone que este tipo de reglas aparecen gracias a la evolución, donde individuos y especies con esas reglas han prosperado y sus descendientes las han heredado. Haikonen no deja claro si este tipo de reglas se pueden aprender por evolución o deben ser programadas por el constructor de la máquina. La combinación de estas señales y reglas dan lugar a las emociones de la máquina y sus comportamientos asociados, por ejemplo:

- Miedo: "malo"+ "dolor" = retirarse.
- Curiosidad: "bueno" + "novedad" = acercarse.
- Deseo: "bueno" + "placer" = acercarse.

La máquina de Haikonen tiene un mecanismo de atención asociado a las emociones. Los patrones de memoria que representan un episodio pasado infeliz, como por ejemplo "una pelota lanzada rasga la piel", no se almacenan de forma neutral. Lo que se almacena no es sólo la representación interna de la pelota, la memoria visual de la pelota volando hacia la máquina y los datos de los sensores de la piel. También se almacena el hecho de que el reporte de daños en la piel era *dolor*, y que esto era *malo*, además de la reacción de retirada de la máquina en ese momento.

Todas estas representaciones se agrupan. Las representaciones del *dolor* y lo *malo* dan a los patrones de memoria un alto valor con significado emocional. Cuando los patrones de memoria anteriores son evocados por los patrones de la pelota amarilla, el significado emocional asociado lleva al cerebro de la máquina a prestarles un alto grado de atención. Esto significa que estos patrones estén activos más tiempo, los voltajes de los vectores de señales permanecen altos, mientras que otros vectores de señales (como la imaginación de la "pelota morada rodando cuesta arriba") se diluyen.

La arquitectura de Haikonen también proporciona un mecanismo de introspección. Normalmente, las entradas del proceso de percepción provienen de sensores que miden información externa o interna. Pero también se pueden conectar a la entrada de los procesos de percepción las salidas de los propios procesos de percepción o los patrones construidos por procesos de imaginación. El resultado es que los grupos de neuronas de percepción de la máquina no sólo ayudan a percibir el mundo exterior sino que también perciben el hardware interno del cuerpo, ayudando a iniciar el flujo de imágenes mentales como se ha descrito anteriormente. Estos circuitos además permiten percibir el propio flujo. La introspección en este contexto consiste en la máquina percibiendo y examinando su propio flujo mental.

2.2.4.3 Otras arquitecturas cognitivas basadas en la conciencia

Aparte de las arquitecturas descritas anteriormente, existen otros modelos menos conocidos que también exploran diferentes aspectos de la conciencia y tratan de aplicarlos en sistemas artificiales.

Sugiyama propone un modelo cognitivo centrado explícitamente en la conciencia (Sugiyama 2000). Concretamente, sugiere el uso de una operación primitiva denominada "reflexión", que permitiría reconocer y comparar de forma diferenciada las percepciones exteriores e interiores del propio sujeto. Sugiyama plantea las siguientes conjeturas:

- Cuando se dispone de un mecanismo para comprender la diferencia entre dentro (la propia entidad) y fuera (el mundo exterior) se es capaz de distinguir un objeto entre varios mediante el reconocimiento. A este fenómeno se le llama conciencia si se produce de forma intencionada por parte de la entidad.
- Para reconocer un objeto se necesita tener algún tipo de imágenes internas de objetos. Comparando el objeto con estas imágenes, se dice que se entiende o reconoce un objeto cuando el agente puede encontrar una imagen igual o similar internamente.
- En una primera fase se desarrolla la conciencia mayoritariamente de forma interna, posteriormente se desarrolla una conciencia del propio ser y la relación existente entre el interior del sujeto y el mundo exterior.

Teniendo en cuenta estas conjeturas se puede construir una teoría de la conciencia basada en imágenes mentales de objetos, que se "reflejan" en el interior del sujeto, creándose las imágenes mentales internas que el agente tiene acerca del mundo exterior. La limitación de dicha teoría se sitúa en la propia primitiva "reflexión", cuyo funcionamiento no queda claro.

Otro autores, como Kitamura et al. (1995), siendo más pragmáticos a la hora de considerar la conciencia, siempre teniendo en mente la aplicación en robots autónomos, se preguntan si es posible el aprendizaje sin ningún tipo de algoritmo de aprendizaje específico. Sus experimentos demuestran que cierto grado de aprendizaje se puede realizar mediante una arquitectura que relaciona directamente la conciencia con el comportamiento. Pero también queda de manifiesto en sus experimentos que se necesitan algoritmos de aprendizaje para llegar a emular los niveles de aprendizaje de los que son capaces los mamíferos superiores.

Centrándose en la relación existente entre conciencia y comportamiento, Brown propone un mecanismo de células de memoria que dotaría a un robot de la ilusión interna de un entorno (Brown 1997). La idea es añadir al componente típico sensor/actuador un componente adicional sensor/sensor, que produce condiciones creadas internamente en el robot. En este enfoque, el proceso de aprendizaje es diferente, porque en vez de generar comportamientos únicamente a partir de las percepciones del entorno, se generan a partir de las percepciones internas también. Es decir, la unidad sensor/sensor genera sensaciones internas que se pueden usar para determinar el comportamiento; por ejemplo, cuando no es posible obtener información del entorno.

Samsonovich y De Jong (2004) proponen una arquitectura cognitiva basada en el concepto de "yo" (*self*). Para definir un sistema artificial cognitivo que tenga un yo, los investigadores proponen una representación formal basada en el concepto de esquema[6]. Los esquemas son una especie de modelos o plantillas que procesan elementos de información como primitivas motoras, datos sensoriales o conocimiento semántico. De acuerdo con estos investigadores, para ser autoconsciente se necesita tener una imagen del yo consistente con una serie de "auto-axiomas", atribuir a este yo las experiencias en primera persona y tener la creencia de que este yo es responsable de las acciones cognitivas y motoras del agente. Los auto-axiomas son creencias del agente sobre su

[6] Aquí el término *esquema* se utiliza refiriéndose al concepto "*schema*" introducido por Immanuel Kant en su obra La Crítica de la Razón Pura.

propio yo, como por ejemplo: “este agente es indivisible y mantiene su unidad a lo largo del tiempo, en toda circunstancia” (Samsonovich, Jong 2005).

Además de los conceptos vistos anteriormente, los modelos de Conciencia Artificial normalmente tienen en cuenta las emociones como parte del diseño. A menudo se considera que en un agente situado, como puede ser el caso de un robot autónomo, la capacidad de realizar evaluaciones de tipo emocional puede ser crucial (Manzotti, Metta & Sandini 1998). Aunque los enfoques clásicos de la IA típicamente han ignorado la dimensión emocional, ésta es cada vez más importante en las arquitecturas cognitivas artificiales.

Durante las últimas décadas, el estudio neurobiológico de las emociones ha proporcionado un nuevo entendimiento de los mecanismos subyacentes (ver Apartado 2.1.4.3). Estos conocimientos científicos han inspirado una serie de modelos cognitivos que pretenden aplicar los principios funcionales de las emociones en implementaciones computacionales. Se sabe que el comportamiento humano está fuertemente condicionado por el estado emocional del individuo. Por lo tanto, cualquier arquitectura cognitiva que pretenda modelar con precisión la generación de comportamiento de tipo humano debe incorporar principios funcionales basados en las emociones. En este sentido, como indican Franklin et al. (1998), las emociones proporcionan una valoración de lo bien que se están cumpliendo los objetivos del sistema.

2.2.5 Aplicación en robótica cognitiva

El estudio de la robótica puede abordarse desde diferentes perspectivas o niveles de abstracción: a nivel físico, hardware o mecánico los robots están compuestos de mecanismos diseñados para moverse, incluyendo motores, engranajes, correas, ordenadores de control, sensores y electrónica de comunicaciones (Siegwart, Nourbakhsh 2004). Desde un punto de vista más elaborado, que podríamos denominar nivel de dispositivo, los detalles del hardware se abstraen y un robot se puede considerar como un conjunto de servicios software para motores, odómetros o controladores. Una serie de bibliotecas software darían acceso a todas estas abstracciones. Desde un punto de vista aún más abstracto, a nivel computacional, podríamos considerar un robot autónomo como un conjunto de módulos software enmarcados en una arquitectura que típicamente estaría compuesta de un subsistema de control de la movilidad, un subsistema de control de los sensores y un subsistema de interpretación de la información de los sensores. Por último, el más alto nivel de abstracción es el referido por el término *robótica cognitiva*, donde el robot se modela como un conjunto integrado de funcionalidades cognitivas basadas en la percepción, el razonamiento y el comportamiento.

La habilidad de moverse intencionadamente es inherente a la mayoría de animales. La máxima expresión de esta capacidad se presenta en los seres humanos, caracterizados por un comportamiento inteligente. La creación de máquinas capaces de moverse y operar de forma inteligente en su entorno es uno de los retos principales de la robótica cognitiva. Independientemente de los métodos físicos usados por un robot para desenvolverse en su entorno, el verdadero reto se encuentra en los principios computacionales que se apliquen para gobernar su comportamiento.

La investigación en robótica tradicionalmente se ha centrado en las tareas de control y el procesamiento de las lecturas de los sensores, la planificación de caminos y

el diseño de manipuladores. Por el contrario, la investigación en robótica cognitiva se centra en dotar a los robots y agentes software de funciones cognitivas superiores que les permitan razonar, actuar y percibir de forma robusta en entornos desconocidos y cambiantes (Beetz et al. 2006).

Este tipo de robots deben, por ejemplo, ser capaces de razonar acerca de metas, acciones y recursos (lineales o no lineales, discretos y/o continuos, renovables o imprescindibles); tienen que saber cuándo percibir y qué buscar, inferir los estados cognitivos de otros agentes, ser capaces de ejecutar tareas colaborativas, etc. En resumen, la robótica cognitiva se refiere al diseño integrado de mecanismos de razonamiento, percepción y acción.

La Conciencia Artificial tiene un papel relevante en el campo de la robótica cognitiva. Los robots con capacidades cognitivas superiores deberían mantener la atención, planificar sus acciones, anticiparse y sobre todo ser capaces de razonar sobre otros agentes y sobre ellos mismos. De hecho, no se puede concebir un robot capaz de lograr objetivos complejos en un entorno social humano si éste no implementa al menos alguna forma simple de Conciencia Artificial.

Algunos de los procesos mentales que suelen atribuirse a los seres conscientes son los siguientes: deseo de vivir, reconocimiento de los otros, reconocimiento de uno mismo, emociones, sentido de la duración del tiempo, voluntad, deseo, reconocimiento de los sentidos corporales, creación de expresiones abstractas, capacidad de atención y conocimiento de la existencia de la muerte. Algunas de estas características podrían ser útiles en robots destinados a realizar tareas complejas que normalmente realizan humanos.

Desde un punto de vista evolucionista, se considera la hipótesis de la evolución de la conciencia en los humanos como una ventaja evolutiva (ver Apartado 2.1.3.3). Es decir, la aparición de una conciencia más avanzada en los homínidos modernos habría supuesto un valor añadido (un conjunto de funcionalidades) que favorece la supervivencia de los individuos. Baars sugiere un conjunto detallado de funciones cognitivas en las que la conciencia juega un papel clave (Baars 1997): adaptación, reclutamiento de procesadores especializados, toma de decisiones, corrección de errores, control de las acciones, aprendizaje, planificación, auto-monitorización y optimización. Por lo tanto, el campo de la robótica cognitiva está íntimamente relacionado con la Conciencia Artificial. Normalmente la implementación de funcionalidades asociadas a la conciencia está enmarcada en el ámbito de una arquitectura cognitiva. La conciencia *per se* no tiene sentido a no ser que se integre en el sistema sensoriomotor de un sujeto capaz de desarrollar procesos de percepción y acción.

El objetivo último del desarrollo de arquitecturas cognitivas para robótica es el de construir máquinas que sean capaces de *"saber lo que hacen"*, siendo así más robustas, flexibles y adaptables en el desempeño de sus funciones. Los robots sociales son un ejemplo significativo del tipo de aplicaciones que una máquina con capacidades cognitivas podría realizar. La interacción con humanos es una tarea de extremada complejidad para la que se requieren capacidades cognitivas avanzadas. Se espera que los robots cognitivos del futuro sean capaces de interactuar con los humanos, desenvolviéndose y aprendiendo de forma satisfactoria en entornos cambiantes (Fong, Nourbakhsh & Dautenhahn 2003). En este contexto la aplicación de modelos de Conciencia Artificial podría proporcionar el avance necesario para diseñar robots de este tipo.

Los avances logrados en el campo de la robótica han propiciado la aparición de multitud de dispositivos mecánicos autónomos o semiautónomos, desde robots dedicados a la fabricación hasta asistentes domésticos automáticos. Estos últimos, que actualmente son muy limitados en cuanto a sus prestaciones, incluyen robots de limpieza como la aspiradora Roomba (Forlizzi, DiSalvo 2006). Las limitaciones actuales de este tipo de robots ponen de manifiesto que el amplio abanico de aplicaciones que se conciben para estas máquinas no puede ser alcanzado sin un sistema de control cognitivo. Parece claro que las técnicas clásicas de diseño de software y los enfoques tradicionales de la IA no son suficientes para lidiar con la gran complejidad involucrada en los procesos de percepción y acción en entornos desestructurados. Uno de los principales objetivos de las arquitecturas cognitivas es mejorar el funcionamiento y la sociabilidad de los robots de servicio.

La investigación en ciencias cognitivas y neurociencias es una valiosa fuente de inspiración en el diseño de sistemas artificiales de tipo cognitivo (ver Apartado 2.1). Normalmente, el trabajo en robótica cognitiva se basa en los principios y descubrimientos provenientes de la psicología cognitiva y la neurobiología. Aspectos como la memoria y sus mecanismos neuronales subyacentes se pueden tratar de imitar en los cerebros artificiales de los robots. Sin embargo, algunos de los algoritmos clásicos usados en robótica no suelen tener en cuenta estos conceptos cognitivos (Thrun 2000). Una posible razón de esto es que en entornos controlados los algoritmos clásicos funcionan mejor o suficientemente bien. La aplicación de modelos cognitivos en robótica es un campo relativamente joven, y por lo tanto, queda mucho por investigar en este sentido. Como se ha explicado en el Apartado 2.2.4, existen diferentes enfoques para la aplicación de los modelos cognitivos humanos en sistemas artificiales. Por ejemplo, el modelado a nivel de sistemas considera las áreas funcionales del cerebro y su interacción, como por ejemplo en la arquitectura Ikaros (Balkenius, Morén 2003).

Considerando un robot como un sistema cognitivo se pueden diferenciar los siguientes procesos que tienen lugar en el sistema: percepción, razonamiento y comportamiento. En general el robot es capaz de captar información del medio a través de sus sensores, durante el proceso de percepción. Posteriormente, esta información es interpretada y se transforma en conocimiento útil para el robot. El proceso de razonamiento se refiere a la manipulación automática del conocimiento adquirido para generar nuevo conocimiento. De esta forma, el robot puede responder de forma inteligente a los estímulos del medio a la vez que persigue las metas establecidas en la misión encomendada. Finalmente, el comportamiento que muestra el robot se genera en base a una serie de acciones que realiza de acuerdo a los resultados del proceso de razonamiento.

El proceso de percepción es un requisito imprescindible para implementar incluso los comportamientos más simples de forma autónoma. Este proceso toma como entrada los datos del medio obtenidos por los sensores del robot. Existen multitud de tecnologías orientadas a los mecanismos de medida o telemetría (en el capítulo 7.3.2 se describen los sensores utilizados en la presente tesis doctoral). Los sensores y los algoritmos de percepción son imprescindibles para que, por ejemplo, un robot pueda saber dónde está y pueda razonar acerca de cómo llegar a un determinado lugar.

El control autónomo de un robot móvil no es una mera colección de algoritmos de planificación de caminos, representación de conocimiento, percepción computacional y razonamiento automático. En realidad, un robot autónomo supone una plataforma integradora donde los conceptos teóricos y los algoritmos se prueban físicamente en el

mundo real, enfrentándose a una enorme complejidad. De hecho, las aspiraciones iniciales en el campo de la robótica, representadas a menudo en la ciencia ficción por robots conscientes antropomórficos, han resultado ser inalcanzables para la tecnología actual, rebajándose hasta el nivel de inteligencia mostrada por los insectos (Brooks 1986). Las principales dificultades para emular las capacidades humanas no se encuentran en los componentes físicos (sensores, motores y materiales), sino en el software que gobierna el robot. No se sabe cómo programar un robot para que sea tan inteligente como una persona. De hecho, aún no se sabe cómo programar un robot para que sea al menos tan inteligente y hábil como un pequeño roedor. La inspiración en los procesos cognitivos observados en los humanos, diseñando arquitecturas cognitivas de control, parece ser un camino prometedor para avanzar en este ámbito (McFarland & Bösser 1993). Sin embargo, no hay que olvidar que para determinadas tareas o circunstancias, los mecanismos pre-programados pueden ser superiores a los mecanismos cognitivos.

Aunque existe un gran parque mundial de robots industriales, la mayoría de los robots móviles autónomos que actualmente existen son experimentales. Sólo un pequeño número de robots móviles autónomos se están empezando a usar en tareas domésticas (Forlizzi & DiSalvo 2006). En cualquier caso, se prevé que el número mundial de robots de servicio supere los 11 millones de unidades en 2010 (unos 5 millones de unidades de robots domésticos y más de 6.5 millones de unidades de robots de entretenimiento)[7]. Normalmente, el entorno de operación para los robots de servicio está caracterizado por su complejidad y variabilidad (en contraposición a los típicos entornos controlados de los robots industriales). Adicionalmente, los robots experimentales suelen estar concebidos para desarrollar su misión en ausencia de operadores humanos (debido a lo inhóspito del entorno o el coste y riesgo de enviar humanos a ese lugar. Algunos ejemplos de este tipo de entornos son la exploración espacial, la minería, los dominios microscópicos, la exploración de fondos abisales, vigilancia y seguridad, etc. En definitiva, se trata de entornos muy exigentes, donde se espera que el robot tenga un alto grado de autonomía y flexibilidad. Es en este tipo de ámbitos donde se requieren las capacidades cognitivas superiores asociadas con la conciencia.

Algunos ejemplos de robots experimentales que se beneficiarían de capacidades cognitivas superiores son:

- Robonaut, un robot diseñado por la NASA para participar en misiones espaciales como ayudante de los astronautas humanos (ver Figura 5).
- BEAR, un robot diseñado para las fuerzas armadas de EEUU con el objetivo de rescatar automáticamente soldados caídos en batalla (ver Figura 6).

Aunque existen multitud de robots basados en arquitecturas cognitivas y que incluso imitan la forma humana (antropomorfos), son pocos los proyectos que han abordado directamente el problema de la Conciencia Artificial. Sin embargo, algunos robots ya clásicos en la literatura científica, como el robot humanoide Cog (Brooks et al. 1998) (ver Figura 7) y otros más modernos, como iCub (Tsagarakis et al. 2007) (ver Figura 8), se han usado como objeto de debate a la hora de plantear la posibilidad de reproducir la conciencia en máquinas (Dennett 1997a).

[7] Fuente: International Federation of Robotics 2009.

Figura 5. Robot de exploración espacial *Robonaut*.

Figura 6. Robot de rescate de soldados BEAR (*Battlefield Extraction-Assist Robot*).

Alguno de los comportamientos que Cog era capaz de desarrollar, como los relacionados con la atención, su sistema emocional y la teoría de la mente están directamente relacionados con la conciencia. Sin embargo, la incapacidad de este robot para integrar diversos comportamientos independientes hace patente la falta de un modelo inspirado en la conciencia (entendida ésta como una función integradora de las capacidades cognitivas).

El robot iCub, desarrollado como parte del proyecto RobotCub, tiene como objetivos el desarrollo de un robot cognitivo avanzado y el avance científico en la comprensión de la cognición humana (Tsagarakis et al. 2007). Se pretende que el robot alcance las capacidades cognitivas de un humano de dos años y medio de edad, por lo tanto, no se puede ignorar el desarrollo de la conciencia en este entorno.

En general, cabe destacar que, dado el fracaso en la construcción de robots cognitivos avanzados en el pasado, la tendencia actual en robótica cognitiva se centra en el diseño de robots que pasen por una fase de desarrollo cognitivo al igual que lo hacen los humanos durante su niñez (*robótica epigenética* o de desarrollo). Aunque la conciencia no es uno de los temas predominantes que se discuten en la literatura actual sobre robótica avanzada, sí que las capacidades cognitivas que se pretenden desarrollar están cada vez más relacionadas con la misma: atención, teoría de la mente, empatía, emociones, etc. Los proyectos que se presentan a continuación tienen en común que sus objetivos principales giran en torno al campo de la Conciencia Artificial.

Figura 7. Robot Cog**.**

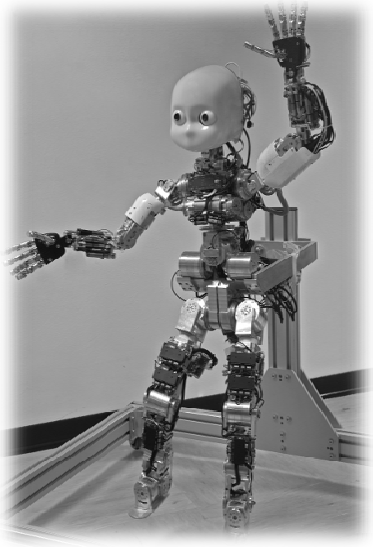


Figura 8. Robot iCub.

CRONOS es uno de los proyectos más importantes que se han llevado a cabo centrados exclusivamente en el problema de la Conciencia Artificial (Holland 2007). CRONOS es también el nombre del robot construido como herramienta central de este proyecto (ver Figura 9). El hardware de este robot está diseñado imitando lo más fielmente posible la estructura muscular y esquelética de un ser humano adulto (Holland, Knight 2006). Como parte del mismo proyecto, también se desarrollaron:

- SIMNOS: una simulación por ordenador de la física del robot en su entorno (ver Figura 10).
- Un sistema de visión inspirado en principios biológicos.
- SpikeStream: un simulador de redes de neuronas de impulsos (*spiking neural networks*).

Usando estas herramientas, el proyecto CRONOS se centró en los aspectos cognitivos, arquitecturales y fenomenológicos de la Conciencia Artificial.

Figura 9. Robot antropomórfico CRONOS.

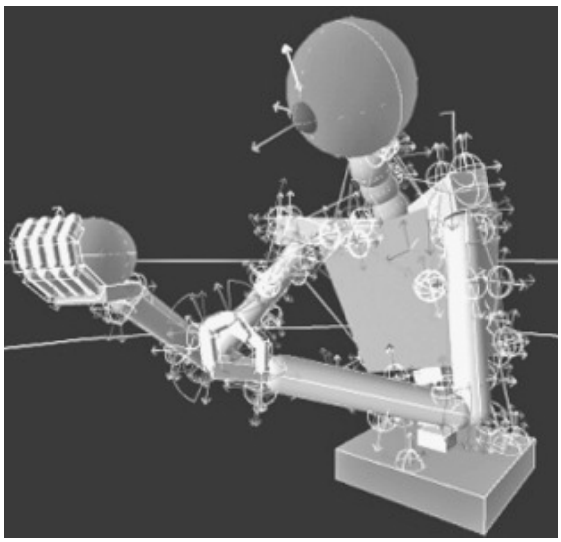


Figura 10. Robot simulado SIMNOS.

De acuerdo con Holland, CRONOS podría desarrollar conciencia gracias a los modelos internos que genera el robot. Estos modelos juegan un papel importante en los estados cognitivos conscientes y podrían ser la causa de, o al menos estar correlacionados con, la producción de conciencia en los humanos (Holland, Goodman 2003). Holland sugiere que los modelos internos que incluyen el propio cuerpo del agente y su relación con el entorno podrían dar lugar a una especie de auto-modelo como el definido por Metzinger (2003) (ver Apartado 2.1.5.1).

Para demostrar esta teoría acerca del modelado interno se utiliza SIMNOS como modelo interno de CRONOS (ver Figura 9 y Figura 10). Utilizando técnicas de localización automática y creación de mapas (SLAM), basadas en la información visual del "ojo" de CRONOS, se obtiene información sobre los movimientos del robot en relación a su entorno. Esta información se usa para actualizar continuamente el modelo

interno SIMNOS. Una vez que este mecanismo está en funcionamiento, también se puede usar para realizar simulaciones de las acciones potenciales del robot, pero sin que las acciones se realicen realmente en el hardware de CRONOS. Es decir, simular (o "imaginar") las consecuencias de posibles acciones del robot.

Adicionalmente, usando SpikeStream, el simulador de redes de neuronas de impulsos (Gamez 2009), Gamez ha desarrollado un sistema de control de inspiración biológica para la cámara de CRONOS (y también para el "ojo" simulado de SIMNOS). Cuando la red de neuronas está conectada al robot, genera movimientos oculares espontáneamente hacia diferentes partes del campo visual y aprende la asociación entre la posición del ojo y determinados estímulos visuales usando la plasticidad de los impulsos neuronales. Esta red de neuronas tiene un sistema emocional que cambia al modo "imaginación" cuando se encuentra con un objeto "negativo". Esto provoca la inhibición de la entrada sensorial y la salida motora mientras la red explora patrones sensoriomotores en busca de un estímulo positivo para su sistema emocional. Cuando esto ocurre, cesa la inhibición y la cámara se mueve directamente hacia una posición en la que tenga el objeto seleccionado en el centro del campo de visión.

Aparte del uso de complejos robots antropomórficos, también se han empleado robots mucho más simples, como Khepera (ver Figura 11), para probar modelos relacionados con la conciencia. Gamez apunta que aunque son muy simples, este tipo de robots pueden moverse autónomamente por su entorno; y por consiguiente pueden construir representaciones susceptibles de ser analizadas en términos de modelos internos o imaginación (Gamez 2008).

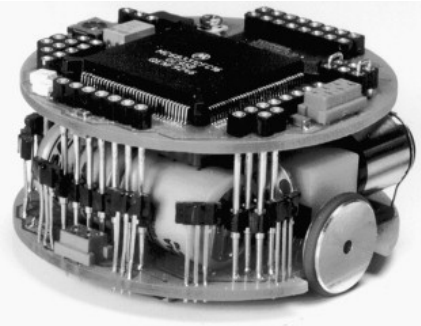


Figura 11. Robot Khepera.

Holland y Goodman han usado el robot Khepera para construir modelos a partir de los datos sensoriomotores (Holland, Goodman 2003). Usando el método de cuantificación de vectores por asignación adaptativa de recursos (CVAAR) (Linaker 2000), que se basa en el hecho de que la información sensoriomotora del robot se mantiene relativamente estable a lo largo del tiempo, se definen "conceptos". En este contexto, el término *concepto* se refiere a la combinación relativamente estable de entrada sensorial y salida motora que se establece mediante el algoritmo CVAAR. Estos conceptos se pueden usar para almacenar largas secuencias de experiencias en poco espacio de memoria, etiquetando los conceptos y almacenando sólo el número de veces que se repiten.

En los experimentos de Holland y Goodman se programó un robot simulado Khepera para que recorriera el espacio a lo largo de las paredes de su entorno a la vez que sorteaba posibles obstáculos. Al mismo tiempo se aplicaba el algoritmo CVAAR

para calcular conceptos. Cada uno de los conceptos obtenidos se correspondía con características del entorno que provocaban lecturas positivas en los sensores de proximidad y la respuesta motora del robot hacia ese estímulo. Usando esta representación se construía un mapa del entorno del robot, usando el proceso que Linåker llama *inversión* (Linaker 2000). "Invirtiendo" los conceptos del robot de esta forma, se producía una representación gráfica del modelo interno del robot y se analizaba como ésta podría usarse para controlar el robot. Holland y Goodman descubrieron que el modelo interno construido a partir de conceptos se podría usar para controlar de forma efectiva el robot, incluso procesando datos nuevos o incompletos y detectando anomalías. La conclusión de los investigadores fue que los modelos internos se pueden desarrollar y estudiar incluso en sistemas simples. Además, puede que jueguen un papel importante en la producción del comportamiento de un organismo. Lo que no está tan claro es que los modelos internos, tal y como se describen en este trabajo, sean un requisito para la producción de una conciencia artificial.

El robot Khepera también se ha usado para investigar la imaginación en sistemas artificiales. Ziemke et al. (2005) realizaron diversos experimentos controlando un robot Khepera con una red de neuronas basada en un módulo sensoriomotor y un módulo de predicción. El módulo sensoriomotor mapeaba la entrada sensorial con la salida motora con el objetivo de evitar obstáculos y seguir una trayectoria recta. El módulo de predicción se encargaba de prever cuál sería la entrada sensorial en el siguiente instante. Los pesos de las redes de neuronas de ambos módulos se ajustaban usando un algoritmo evolutivo. Cuando el robot recibía una entrada sensorial real, el módulo sensoriomotor generaba la señal de control correspondiente. Sin embargo, cuando se "le tapaban los ojos al robot", de forma que no recibía información alguna desde los sensores, la señal de control se generaba alimentando el módulo sensoriomotor con la salida del módulo de predicción. Se comprobó que el comportamiento generado en base a las entradas sensoriales "imaginadas" era muy similar al comportamiento generado usando los datos reales de los sensores. También se demostró que estas capacidades cognitivas asociadas con la conciencia (el modelado interno y la predicción), pueden suponer una mejora considerable en el rendimiento de un robot.

Basándose en el trabajo descrito anteriormente, Stening et al. sustituyeron las redes de neuronas usadas por Ziemke en el robot Khepera por el algoritmo CVAAR (Stening, Jacobsson & Ziemke 2005), usándolo para identificar combinaciones de entrada sensorial y salida motora que fueran relativamente invariantes a lo largo del tiempo. Los conceptos generados aplicando esta técnica se usaron como entrada de una red de neuronas entrenada para predecir cuándo tendría lugar el siguiente concepto. En los experimentos descritos por Stening et al. (2005), el robot se controlaba inicialmente con una red de neuronas entrenada previamente para producir un comportamiento simple de seguimiento. Al mismo tiempo el algoritmo CVAAR extraía las características básicas de la interacción sensoriomotora del robot. Después, se utilizaban las predicciones del siguiente concepto generadas por la otra red de neuronas como entrada de la red de neuronas de control. De esta forma el robot podía simular internamente una secuencia de conceptos sin necesidad de ningún movimiento real. Las representaciones gráficas obtenidas a partir de las representaciones internas del robot se podrían considerar en el marco de la fenomenología sintética (si éstas se considerasen como una especificación del contenido consciente del robot).

Chella et al. (2008) también están trabajando en una arquitectura basada en la conciencia para un robot móvil llamado Cicerobot (ver Figura 12). Este robot móvil está diseñado para desempeñar la tarea de guía en el museo arqueológico de Agrigento en

Italia. De manera semejante a CRONOS, la arquitectura cognitiva de Cicerobot está basada en una simulación interna en tres dimensiones que se actualiza al mismo tiempo que el robot navega por su entorno.

Figura 12. Robot Cicerobot.

Cuando el robot se mueve, envía una copia de los comandos motores al módulo de simulación. Éste calcula constantemente las expectativas acerca de la siguiente posición del robot y la imagen que se espera obtener de la cámara instalada en el robot. Una vez que el movimiento se ha ejecutado, se compara la imagen esperada con la imagen real y se usan las discordancias encontradas para actualizar el modelo tridimensional (de forma que refleje más fielmente la situación actual). Cicerobot usa su simulación interna para planificar las acciones y explorar diferentes escenarios de forma equivalente a como lo hace la imaginación humana (Chella & Macaluso 2009). Los mismos investigadores involucrados en el diseño de Cicerobot especulan acerca de la presencia de estados fenomenológicos en el robot (Chella & Gaglio 2009).

Adicionalmente al uso de robots físicos y simulados, en el campo de la Conciencia Artificial se han usado también agentes virtuales o agentes simulados. Un ejemplo clásico es *CyberChild* (Cotterill 2003), un bebé simulado basado en un sistema neurológico artificial (ver Figura 13). El control de la simulación está basado en diversas redes de neuronas inspiradas en áreas específicas del cerebro, como la amígdala, el hipocampo, etc. CyberChild se basa en un sistema básico de músculos que controlan aspectos como la producción de la voz y el movimiento de los miembros. Adicionalmente la simulación también incorpora algunos mecanismos asociados con la alimentación y el metabolismo (estómago, vejiga, nivel de glucosa en sangre, gasto de energía, etc.). Cuando el bebé toma la “leche virtual”, ésta se metaboliza, convirtiéndose en orina simulada, que se acumula en la vejiga e incrementa el nivel de incomodidad del bebé. El bebé simulado tiene impulsos que le llevan a alimentarse y evitar la incomodidad, pero debe recurrir a la comunicación con un adulto (operador) para poder satisfacer sus necesidades. El objetivo de este proyecto es usar la simulación para identificar correlatos neuronales de la conciencia.

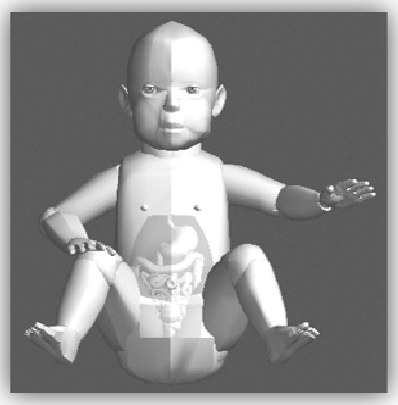


Figura 13. Bebé simulado CiberChild.

Goertzel (2008) también ha empleado agentes simulados, usando la plataforma *Second Life*[8], para experimentar con el desarrollo cognitivo en agentes artificiales. Goertzel relaciona el concepto de Inteligencia Artificial General (IAG) con el concepto de conciencia. Se define IAG como la capacidad de alcanzar metas complejas en entornos complejos, usando recursos computacionales limitados. Para diseñar sistemas capaces de cumplir estos objetivos Gloertzel propone el uso de una arquitectura cognitiva comercial denominada *Novamente Cognition Engine* (NCE), que también tiene una variante de código abierto denominada *OpenCog Prime* (OCP) (Goertzel 2009). La arquitectura OCP es similar a LIDA (ver Apartado 2.2.4.1) y es capaz de simular internamente "películas" (secuencias de imágenes mentales) basadas en contenidos de la memoria o contenidos inventados. Usando esta arquitectura se han realizado experimentos con mascotas virtuales a las que se les puede enseñar a realizar comportamientos simples.

2.2.6 Fenomenología Sintética y Qualia Artificiales

Tal como se ha descrito en el Apartado 2.2.1 y así como también apunta Chrisley (2009), no toda la investigación en Conciencia Artificial se centra exclusivamente en la generación de estados fenomenológicos en sistemas artificiales. También se pueden diseñar sistemas que no tengan en sí mismos estados fenomenológicos, sino que persigan exclusivamente el objetivo de simular o modelar los organismos que sí que los poseen. En este contexto, Chrisley usa la expresión *Fenomenología Sintética*[9] (FS) para referirse al campo de investigación que trata de simular o modelar estados fenomenológicos en una máquina (real o simulada). Es decir, Chrisley excluye de este campo los intentos de instanciar de forma efectiva estados fenomenológicos en un artefacto. Sin embargo, otros autores usan el término para referirse también a la identificación e instanciación de estados fenomenológicos en implementaciones de Conciencia Artificial (Gamez 2008).

[8] http://secondlife.com/

[9] El término "Fenomenología Sintética" fue usado por primera vez en el título de una charla dada por Scott Jordan en 1998 en el Instituto Max Planck de Múnich. Más tarde se empezó a usar en el ámbito de la investigación en Conciencia Artificial.

En general, podría decirse que un programa amplio de investigación en FS debe abordar tanto el problema de la especificación del contenido de los estados fenomenológicos, como la detección de la existencia de los mismos. Ya que normalmente el contenido de la experiencia consciente no puede ser especificado de forma precisa usando herramientas lingüísticas, la FS también debe proporcionar herramientas para la especificación de este tipo de contenidos. Chrisley aboga por el uso de representaciones gráficas (*depiction*) para especificar el contenido de la experiencia visual; y usa una arquitectura basada en expectativas para demostrar su uso práctico (Chrisley 2009).

Los trabajos realizados aplicando el algoritmo CVAAR a la navegación del robot Khepera (ver Apartado 2.2.5) también constituyen un ejemplo de investigación en FS. La representación gráfica de los conceptos generados en el robot al navegar podría considerarse una especificación funcional de lo que sería la experiencia subjetiva del robot (Holland, Goodman 2003, Stening, Jacobsson & Ziemke 2005). Aunque los autores no se manifiestan en cuanto a la existencia de estados fenomenológicos asociados, sí se demuestra que las representaciones generadas son útiles para generar comportamientos consistentes con los que podrían ser generados en situaciones análogas por seres conscientes.

Un enfoque típico para abordar el problema de la FS es considerar una teoría de la conciencia dada y comprobar si el sistema analizado cumple con los requisitos que establece la teoría para la producción de estados fenomenológicos. El problema es que no existe una teoría comúnmente aceptada. En cualquier caso, algunos autores han aportado sus propuestas en este sentido. Por ejemplo, Aleksander y Morton establecieron ciertos criterios que una arquitectura debe cumplir para considerar que pudiera generar estados fenomenológicos. Concretamente, estos autores afirman que "para que un sistema *S* produzca estados fenomenológicos, debe tener un mecanismo capaz de representar como se ve el mundo y el sistema *S* que está dentro de él desde el punto de vista del mismo *S*" (Aleksander, Morton 2006).

En contraposición a la argumentación anterior, hay autores que defienden la idea de que no es posible identificar los criterios que establecen si un sistema es capaz de tener estados fenomenológicos (Moor 1988, Prinz 2003). En base a este tipo de posturas se podría plantear la investigación en Conciencia Artificial con un enfoque puramente funcional, sin considerar la posibilidad de estados fenomenológicos en máquinas. Sin embargo, según Block (Block 1980), si una teoría funcional no contempla los qualia, no es una teoría de la mente válida. Es decir, si una teoría de la mente no es capaz de explicar la experiencia cualitativa de los seres conscientes, no sirve para explicar la mente humana.

La postura de Chrisley presentada anteriormente soslaya este problema centrándose en la posibilidad de modelar y simular la experiencia consciente (Chrisley 2009). Concretamente, se propone la representación de "contenido no conceptual". Es decir, la especificación del contenido de la experiencia se basa en una representación de los propios estados internos por los que la máquina transita (Chrisley 1995). Este punto de vista permite al menos trabajar en la definición de modelos que sean capaces de describir los qualia que podrían generarse en las máquinas sin tener que enfrentarse directamente al "problema duro" (Chalmers 1995).

Siguiendo una estrategia similar a la de Chrisley, Gamez ha propuesto un enfoque distinto para tratar el tema de la FS. Este autor propone una forma de dividir los estados internos de un sistema en una serie de representaciones estructuradas que están

relacionadas con estímulos concretos del entorno. Estas estructuras proporcionan una descripción de los estados por lo que transita el sistema sin necesidad de recurrir al lenguaje (Gamez 2006).

Un aspecto interesante relacionado con la FS (y que se desarrolla en detalle en el Capítulo 6) es el concepto de qualia artificiales. La identificación, caracterización, análisis y modelado de los qualia en el contexto de la Conciencia Artificial es un área emergente que potencialmente puede aportar nuevo conocimiento científico sobre la conciencia. Los trabajos existentes relacionados con esta línea de investigación son escasos y preliminares. A continuación se describen algunos ejemplos destacados en los que se aborda directamente el problema de los qualia artificiales.

El análisis de las correlaciones computacionales de los qualia artificiales realizado por Chella y Gaglio es uno de estos ejemplos significativos (Chella, Gaglio 2009). En este trabajo, que se basa en la arquitectura cognitiva desarrollada por los mismos investigadores (Chella, Frixione & Gaglio 2008), un proceso activo integra los flujos de información interna y externa para reconstruir una visión subjetiva de la escena percibida por el robot. Los autores argumentan que este proceso de integración entre la reconstrucción del modelo interno del mundo y la percepción propioceptiva genera los qualia artificiales del sistema.

La arquitectura cognitiva de Haikonen (ver Apartado 2.2.4.2) es otro ejemplo de modelo computacional que tiene en cuenta la generación de qualia artificiales (Haikonen 2009). Haikonen argumenta que la realización de su arquitectura estaría dotada de un mecanismo inherente de producción de qualia con significados basados en la situación del agente en el mundo que lo rodea (*grounded meaning*). Además, dicha máquina sería capaz de comunicar estos perceptos con significado a través de símbolos secundarios como las palabras de un lenguaje.

Aunque otras implementaciones de Conciencia Artificial no han sido diseñadas originalmente para tener conciencia fenomenológica, se puede explorar su capacidad para generar, o al menos especificar, estados fenomenológicos. Por ejemplo, en el caso de LIDA (ver Apartado 2.2.4.1), que se ha diseñado para ser un modelo de conciencia de acceso, ¿qué mecanismos sería preciso añadir para que LIDA fuera fenomenológicamente consciente? Tal y como apuntan sus creadores (Ramamurthy 2008), la implementación de un mecanismo para la generación de un mundo perceptual estable y coherente según las líneas establecidas por Merker (Merker 2005, Franklin 2005b) (es decir, la eliminación del movimiento aparente producido por el movimiento de los receptores sensoriales) podría contribuir al diseño de máquinas fenomenológicamente conscientes.

Otro trabajo significativo orientado hacia la definición de arquitecturas de procesamiento de la información equiparables a los humanos es H-CogAff (Sloman, Scheutz 2002), sistema en el que se añade una capa de meta-gestión basada en procesos reflexivos. Este caso particular de CogAff (Sloman 2001), pretende especificar una arquitectura mínima equiparable cognitivamente a un humano. En CogAff el problema de los qualia se aborda como un concepto basado en la arquitectura (Sloman, Chrisley 2003). Concretamente, se considera que las arquitecturas de máquina virtual son un dominio adecuado para la experimentación en Conciencia Artificial. El funcionalismo basado en máquinas virtuales proporciona una explicación para la fenomenología en la que la máquina virtual, aun no siendo física, es una máquina real que puede afectar y es afectada por el mundo físico que la rodea. En definitiva, Sloman argumenta que llevar el análisis de los qualia al nivel de descripción de las máquinas virtuales – el cual no es

definible en términos de las ciencias físicas – podría aclarar lo que realmente son los qualia (ver [Sloman, Chrisley 2003] para una argumentación detallada sobre este punto).

Algunos autores han abordado el problema de los qualia artificiales desde puntos de vista similares a los adoptados en la presente tesis (ver Capítulo 6), aunque centrándose más en aspectos neurológicos o matemáticos. Por ejemplo, Lehky y Sejnowski diseñaron una red de neuronas artificial capaz de mapear la variable de entrada longitud de onda en un espacio de color de tal forma que podía predecir la aparición del quale del color blanco – del cual se sabe que no está directamente relacionado con ninguna longitud de onda en particular (Lehky, Sejnowski 1999).

Los qualia también se han caracterizado desde el punto de vista de la Teoría de la Integración de la Información (ver Apartado 2.1.3.5). Balduzzi y Tononi han propuesto recientemente una representación matemática para caracterizar las relaciones existentes en el "*espacio de los qualia*" desde el punto de vista de la información que éstos integran (Balduzzi, Tononi 2009). El modelo de generación de qualia artificiales (ver Capítulo 6) propuesto en esta tesis se basa en la arquitectura cognitiva también realizada como parte de esta tesis (ver Capítulo 4), centrándose por lo tanto en la funcionalidad asociada a los qualia en términos de interacción sensoriomotora.

2.3 Medición del nivel de conciencia

Uno de los problemas clave que existe actualmente en el campo de la Conciencia Artificial es la necesidad de evaluar de forma precisa el nivel de conciencia que un agente artificial podría desarrollar. Una de las propiedades de la conciencia es que se trata de un fenómeno que puede darse en diferentes grados. En general, la conciencia implica diferentes estados, cualidades y niveles. Sin embargo, la tarea de evaluar el nivel de conciencia, o los estados conscientes que una entidad posee en un momento dado, presenta múltiples problemas y controversias. De hecho, las evaluaciones del nivel de conciencia en humanos y en otros animales sigue siendo un reto científico (Seth, Baars & Edelman 2005).

¿Por qué es tan complicada esta tarea de medición? La raíz del problema viene del hecho de que la conciencia es un fenómeno privado que se da en primera persona, mientras que el método científico está basado completamente en las observaciones objetivas en tercera persona. ¿Significa esto que la conciencia está fuera del alcance de la investigación científica? Dado que los análisis en primera persona no son suficientes por sí solos para ofrecer un valor científico convincente, se pueden adoptar puntos de vista alternativos basados en la observación en tercera persona (ver Apartado 1.1.5). Aunque estos enfoques en tercera persona no puedan ofrecer una inspección directa de la conciencia, tienen validez científica, al menos como herramientas para estimar la similitud entre los patrones de comportamiento producidos por los humanos y aquellos generados por implementaciones de Conciencia Artificial

Aunque instintivamente podría pensarse que existen únicamente dos estados posibles de conciencia: considerando que en un momento dado un sujeto está bien consciente o bien inconsciente, numerosas investigaciones sugieren que se puede diferenciar entre distintos estados de conciencia en los humanos (Seth et al. 2006). Estos cambios en el estado consciente se pueden comprobar observando tanto criterios neurofisiológicos, como el EEG (electroencefalograma), o parámetros de

comportamiento, como la diferencia entre sueño y vigilia o la comunicación verbal. La aplicación conjunta de estos métodos constituye un enfoque heterofenomenológico, pues se combina la observación en tercera persona con la comunicación de las observaciones en primera persona.

Otro problema clave en la aplicación de medidas de conciencia a las máquinas es que los métodos que se usan actualmente con humanos no son directamente aplicables en el dominio de la Conciencia Artificial. A continuación se describen las principales herramientas para la medición de conciencia en organismos biológicos (Apartado 2.3.1), para luego abordar el problema análogo de la medición de la conciencia en sistemas artificiales (Apartado 2.3.2).

2.3.1 Medición del nivel de conciencia en organismos biológicos

En un momento dado una persona es consciente de ciertas cosas, por ejemplo, probablemente el lector de esta tesis es ahora consciente de que está leyendo esta frase. Una persona puede detectar fácilmente que es consciente porque puede acceder (al menos en parte) a sus propios pensamientos y sentimientos. Pero, ¿cómo se puede saber si otro ser es consciente? Los investigadores que pretenden determinar el grado de conciencia de algunos animales utilizan diversas pruebas (Seth, Baars & Edelman 2005, Edelman, Baars & Seth 2005), como la resolución de algunos problemas que implican la necesidad de que el sujeto se reconozca a sí mismo en el mundo. Un ejemplo típico, denominado “la prueba del espejo”, consiste en emplear un espejo para comprobar si el animal en cuestión se reconoce a sí mismo (Gallup 1977).

Como en otros ámbitos del estudio de la conciencia, también con respecto a la medición conviene diferenciar entre conciencia funcional (conciencia A, S y M) y conciencia fenomenológica (conciencia P). La medición de la conciencia funcional abarca las dimensiones conocidas como conciencia A (conciencia de acceso), conciencia S (autoconciencia) y conciencia M (conciencia de monitorización).

La prueba del espejo es un ejemplo de experimento enmarcado en el dominio de la conciencia S, aunque también atañe claramente a la conciencia A. En realidad, se trata de una prueba de auto-reconocimiento en la que se determina si un sujeto es capaz de reconocer su propia imagen reflejada en un espejo (ver Figura 14). Hasta ahora sólo individuos pertenecientes a las siguientes especies han pasado la prueba del espejo: humanos (de más de 2 años de edad), grandes simios (bonobos, chimpancés, orangutanes y gorilas), monos rhesus, elefantes, delfines mular, ratas y pulpos (Gallup 1977, De Veer et al. 2003, Parker, Mitchell & Boccia 2006).

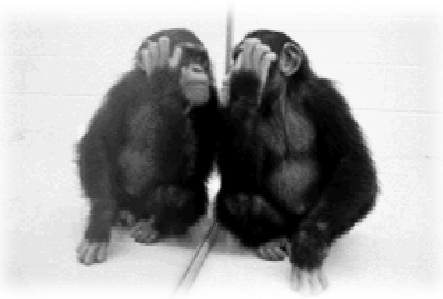


Figura 14. Chimpancé superando la prueba del espejo.

Es importante tener en cuenta que sólo un determinado número de individuos de estas especies han pasado la prueba, mientras que otros generalmente no lo consiguen. Obviamente la prueba tiene que adaptarse para cada especie, aunque típicamente consiste en hacer marcas no olorosas en la frente del animal mientras este está anestesiado. La prueba del espejo se considera por algunos autores como la mejor forma disponible de probar la autoconciencia de un organismo (Riba 1998).

En los humanos, es en el ámbito del diagnóstico clínico donde suelen aplicarse técnicas para la evaluación del nivel de conciencia de los pacientes. Normalmente, estas técnicas se basan en medidas neurofisiológicas y también en el comportamiento observado (Schnakers et al. 2009). En la actualidad, también se están empezando a usar herramientas de diagnóstico por imagen, especialmente en los casos de pacientes con trastornos de la conciencia (Laureys, Owen & Schiff 2004, Giacino 2004). En cualquier caso, las herramientas típicas que se usan en los servicios de urgencias son del tipo de la Escala de Coma de Glasgow o la más moderna Escala JFK de Recuperación del Coma en su versión revisada (Jennett 2002, Giacino, Kalmar & Whyte 2004). Otras técnicas de medida relacionadas, usadas en la investigación médica, incluyen marcadores neurofisiológicos como el índice biespectral (Rosow, Manberg 2001), los potenciales corticales evocados (Koivisto, Revonsuo 2003) o la sincronía neuronal (Vanderwolf 2000, Singer, Gray 1995).

También existen escalas psicológicas que se centran en aspectos específicos relacionados con la conciencia. Por ejemplo, la Escala de Autoconciencia Privada de Fenigstein (Fenigstein, Scheier & Buss 1975). Sin embargo, este tipo de escalas se centran en el diagnóstico de trastornos psicológicos. Es decir, sirven para describir rasgos de la personalidad o del comportamiento de la persona que están en mayor o menor medida relacionados con la conciencia.

Como se ha visto, el análisis del comportamiento es una de las metodologías comúnmente aplicadas para evaluar el nivel de conciencia de los seres humanos. También en base al comportamiento observado en diferentes seres, muchos autores argumentan que la evolución ha dado lugar a varios niveles de conciencia en las criaturas vivas (Dennett 1997b). Por ejemplo, Kitamura et al. (1995) presentan la siguiente lista ordenada según el supuesto nivel de conciencia del individuo:

1. Bacteria.
2. Insecto.
3. Pez.
4. Mamífero.
5. Mono.
6. Chimpancé.
7. Bebé humano.
8. Niño humano.
9. Adulto humano.

Aunque este tipo de clasificaciones o jerarquías evolutivas son claramente intuitivas, no existe una base científica sólida que demuestre mediante una medida objetiva los diferentes tipos de conciencia existentes en la naturaleza (Riba 1998). Actualmente se está investigando con diversas especies, observando tanto su comportamiento como el diseño de su sistema nervioso (Seth, Baars & Edelman 2005, Edelman, Baars & Seth 2005, Parker, Mitchell & Boccia 2006). El objetivo es establecer unos criterios más objetivos en relación a la cuestión de la conciencia animal.

Por ejemplo, Merker especula que la conciencia apareció en la evolución como un mecanismo eficiente para lidiar con el movimiento relativo de los órganos sensitivos (Merker 2005).

Otro enfoque similar, también basado eminentemente en el comportamiento observado y en la evolución, es el propuesto por el filósofo Tran Duc Thao (1971), que considera la filogenia del comportamiento animal y la ontogenia humana desde el punto de vista de la fenomenología y el materialismo dialéctico. En este sentido, se define conciencia como el conjunto de cosas que un animal puede sentir, recordar y pensar. Tran Duc Thao define el término *campo de conciencia* como el "lugar" donde aparece la conciencia en los animales, es decir, la función de la conciencia presente en cada especie. Kitamura describe este modelo, expresado en términos psicológicos, asociando comportamientos concretos a los niveles de filogenia animal y ontogenia humana establecidos por Tran Duc Thao (Kitamura, Otsuka & Nakao 1995). Se definen un total de 8 niveles asociados con diferentes fases del desarrollo de la conciencia en la evolución y en los humanos (ver Tabla 1 - adaptado de [Kitamura, Otsuka & Nakao 1995]). La idea básica de este modelo es que la conciencia aparece cuando una acción intencional de un nivel inmediatamente inferior es inhibida por causas internas o externas. En ese momento la conciencia provoca la elección de una acción intencionalmente. Si no hay nada que impida la ejecución de una acción, ni se prevé que lo vaya a haber, la conciencia desaparece y las acciones se realizan "automáticamente". Aunque este modelo proporciona una jerarquía de niveles, no establece ningún mecanismo concreto para determinar el nivel de conciencia de un individuo.

Este tipo de clasificaciones de los niveles de conciencia en base a la filogenia del reino animal o la ontogenia en el hombre, podrían servir como referencia en las implementaciones de Conciencia Artificial. Sin embargo, hay aspectos que estas escalas no cubren, como por ejemplo las diferentes dimensiones de la conciencia. La conciencia P está probablemente presente en todos los mamíferos y definitivamente en los primates. También es muy posible que los primates posean autoconciencia porque, como se ha mencionado anteriormente, son capaces de reconocerse a sí mismos en pruebas de reconocimiento con espejos.

En relación con el problema específico de la identificación de estados fenomenológicos (conciencia P), Metzinger propone un conjunto de restricciones que un contenido mental debe satisfacer para ser considerado como una representación fenomenológica (Metzinger 2003). Metzinger considera la conciencia P como un fenómeno gradual, siendo posible la existencia de diferentes niveles de conciencia asociados al grado de satisfacción de las restricciones consideradas.

También encaminadas a la medición de la conciencia P, existen algunas propuestas que se enmarcan dentro de las denominadas medidas de complejidad neuronal. Estas medidas suelen basarse en las relaciones causales entre los elementos que forman una red de procesamiento. La medida más popular basada en la causalidad es Φ (fi), propuesta en base a la TII (ver Apartado 2.1.3.5). De acuerdo con la TII, esta medida proporciona la "cantidad" de conciencia que puede generar un sistema (Tononi 2004). Φ se define como la cantidad de información causalmente efectiva que es posible integrar a través del enlace más débil (en el sentido de transmisión de información) de un sistema. El nivel de conciencia que se mide con Φ tiene un carácter de "disposición" o "potencialidad". De acuerdo con esta medida, son los valores de las variables que protagonizan el procesamiento de la información los que determinan el contenido de una imagen mental consciente.

Tabla 1. Niveles de conciencia en la filogenia animal y la ontogenia humana.

Nivel	Filogenia	Ontogenia Humana	Campo de conciencia	Categoría de comportamiento
8	Hombre.	2 años	Concepción.	Acción Lingüística.
7	Hombre.	2 años	Representación.	Fabricación de herramientas.
6	Simio.	18 meses	Imagen.	Uso de herramientas.
5	Mono.	1 año	Relación espacial/temporal de objetos.	Uso del medio.
4	Mamífero (cuadrúpedo).	9 meses	Emociones estables hacia los objetos.	Búsqueda, postura, posicionamiento de extremidades.
3	Pez.	5 meses	Ilusión temporal y emoción hacia el objeto presente.	Captura, acercamiento, ataque, evasión, escape.
2	Lombriz de tierra.	1 mes	Sensación de placer y dolor.	Orientación y posicionamiento de cuerpo y extremidades.
1	Anémona, esponja.	0	Sensación primitiva.	Desplazamiento reflejo. Alimentación.

Una de las características principales de Φ es que la medida se plantea como una condición suficiente para la aparición de conciencia P, de forma que cualquier sistema con un Φ suficientemente alto, sería consciente. Recientemente Tononi ha publicado una versión revisada de su teoría en la que Φ, además de medir la capacidad de integración, se convierte en una medida de la dinámica del sistema (Tononi 2008). En esta revisión, Φ mide la información generada cuando un sistema transita de un estado a otro. Concretamente, Φ mide la información que (durante la transición) un conjunto de elementos genera, adicionalmente a la suma de información ya generada por cada una de las partes. Es decir, la nueva Φ mide la capacidad de "emergencia" de nueva información. Una desventaja de esta medida es que su cálculo es prohibitivo en términos de requisitos computacionales.

2.3.2 Medición del nivel de conciencia artificial

Aunque, como se ha visto en el apartado anterior, existen multitud de trabajos dirigidos a la medición de la conciencia en humanos, los mismos índices y métodos no se pueden aplicar en el dominio de los agentes artificiales. El principal motivo es que el sustrato que potencialmente da lugar a la conciencia en ambos casos es distinto. Mientras que las medidas biológicas se centran en el sistema nervioso, las implementaciones artificiales están construidas con materiales y mecanismos radicalmente diferentes.

Como se ha visto en el Apartado 2.1.5.2, algunos autores defienden la necesidad de un sustrato específico para la producción de la conciencia. En el caso de estas teorías, la medición de la conciencia en sistemas que usan sustratos diferentes no tiene sentido. Sin

embargo, de acuerdo con el resto de teorías neurobiológicas y cognitivas, simplemente sería necesario establecer nuevos métodos de medición adaptados a las implementaciones de Conciencia Artificial.

Como se ha indicado en el apartado anterior, es común que el estudio de la conciencia en el mundo animal se realice por medio de la observación del comportamiento y el análisis del funcionamiento del cerebro. Sin embargo, cuando se traslada este problema a dispositivos artificiales nos encontramos con preguntas como estas: ¿cómo podemos establecer, por ejemplo, el nivel de conciencia de un software que implementa un modelo de redes de neuronas artificiales? ¿Qué propiedades evidencian la existencia de conciencia en un sujeto? ¿El auto-conocimiento? ¿El uso del lenguaje? ¿La resolución de determinados problemas?

Los procesos de identificación de comportamientos consistentes con la conciencia en organismos biológicos se pueden adaptar fácilmente al dominio de la Conciencia Artificial. Sin embargo, la medidas de tipo neurofisiológico no son válidas, a no ser que la implementación que se analice emule con alto grado de fidelidad los complejos procesos neurobiológicos y hormonales.

El famoso Test de Turing encaja en el marco de identificación de comportamientos consistentes con la conciencia (Turing 1950). Este test se basa en la siguiente suposición: si una máquina se comporta en todos los aspectos como inteligente, entonces es que es inteligente. Aunque habitualmente se define este test como una prueba de inteligencia, es inmediato extrapolar su aplicación al ámbito de la detección de la conciencia (si una máquina se comporta como un ser consciente, debe ser consciente). No obstante, no está claro como la aplicación del Test de Turing podría contribuir a una medición del nivel de conciencia, ya que en realidad esta prueba está orientada a determinar si el sistema artificial ha alcanzado ya un nivel equiparable al humano (no teniéndose en cuenta posibles niveles inferiores o superiores). Además la prueba original propuesta por Turing se basa exclusivamente en las capacidades lingüísticas, lo que podría considerarse una limitación si lo que se pretende es evaluar también la capacidad sensoriomotora de una máquina. En cualquier caso se podrían aplicar versiones más elaboradas del Test de Turing, como las propuestas por Hanard (Harnad 1994), que están basadas en el comportamiento y son mucho más exigentes. En cualquier caso, tanto el Test de Turing como las variaciones asociadas son demasiado genéricas como para establecer una medición del desarrollo o nivel específico de conciencia en un sistema artificial.

Otro problema que se puede plantear en relación al Test de Turing es la objeción de la habitación china propuesta por Searle (1980). El experimento mental de la habitación china consiste en una persona que se encuentra dentro de una habitación y recibe caracteres escritos en chino. Esta persona, aunque no sabe nada de chino, procesa los caracteres de acuerdo a un conjunto de reglas que tiene disponibles en la habitación (programa), y finalmente entrega unos resultados en base a esas reglas, pero sin haber comprendido nada de lo que se dice en los mensajes. El conjunto de reglas disponibles en la habitación podría ser tal que permitiese a la persona producir un comportamiento aparentemente consciente. Sin embargo, según Searle, la persona no sería consciente porque no tendría estados intencionales relacionados con los caracteres chinos.

Algunos autores argumentan que la objeción de la habitación china se puede superar si se aplican mecanismos de conexión entre conceptos y entorno (*symbol grounding*), de forma que los símbolos que maneja el procesador (persona en la habitación) tengan un significado (Harnad 1990). En este sentido, Haikonen también

propone un mecanismo para enlazar los conceptos o símbolos de alto nivel con las entradas sensoriales, confiriéndoles así un significado basado en la interacción del agente con su entorno (*grounded meaning*) (Haikonen 2007b). En resumen, se argumenta que si se pudiera añadir un nivel subsimbólico que proporcione este tipo de conexión con el entorno, la habitación china comprendería los símbolos que manipula.

Al considerarse el problema tan complejo que supone medir la conciencia en sistemas artificiales, y viendo que parece muy complicado proponer un método objetivo, algunos autores han optado por ofrecer predicciones relativamente plausibles basadas en diferentes suposiciones. Un ejemplo de este tipo de estrategias es la Escala de Probabilidad Ordinal (EPO) propuesta por Gamez (2005). Esta escala pretende proporcionar una medida de la probabilidad de que una máquina sea capaz de producir estados fenomenológicos. La escala EPO se basa en una medida de similitud entre el diseño de la máquina y el cerebro humano.

Uno de los trabajos más notables realizados en el ámbito de la Conciencia Artificial con el objetivo de establecer unos criterios claros para la detección de la conciencia es el conjunto de axiomas propuestos por Aleksander et al. (2003, 2007). Según Aleksander, se requiere un conjunto mínimo de axiomas para que un agente se considere consciente: sensación de lugar (*depiction*), imaginación, atención, planificación y emociones. Adicionalmente a estas capacidades, Haikonen también ha apuntado que la capacidad de comunicar estados mentales (a otros y a uno mismo) es también un requisito para tener conciencia. El habla interior con un significado basado en la interacción con el medio (*grounded meaning*) es una manifestación de la existencia de contenido mental; asimismo, la capacidad de comunicar este contenido es un indicador de la presencia de conciencia (Haikonen 2007b).

Como métodos más concretos, se han propuesto pruebas de conciencia específicas para máquinas, aunque normalmente estas pruebas se limitan a comprobar si el dispositivo mantiene una representación interna adaptada al entorno exterior (Brown 1997). En definitiva, se trata de comprobar si un robot tiene un modelo interno que le permite actuar incluso en el caso de ausencia temporal de datos provenientes de los sensores internos (Holland, Goodman 2003). Este tipo de experimentos son equivalentes a los desarrollados por Ziemke et al. con el robot Khepera (Ziemke, Jirenhed & Hesslow 2005) (ver Apartado 2.2.5). Desde un punto de vista funcional, este tipo de pruebas sólo demuestra que el robot es capaz de generar su comportamiento basándose en una simulación interna. Es decir, el robot presenta algunas funcionalidades relacionadas con los axiomas descritos por Aleksander – imaginación, atención, etc. (Aleksander, Dunmall 2003).

Análogamente a los experimentos que determinan la presencia de mecanismos de atención o imaginación en robots, también se pueden adaptar pruebas relacionadas con la autoconciencia. En este ámbito, la prueba del espejo es un ejemplo clásico que Gordon Gallup introdujo en los años 70 aplicándola a primates (Gallup 1977). Recientemente, se ha planteado la utilidad y la aplicabilidad de esta prueba en el dominio de la Conciencia Artificial y se han desarrollado varios experimentos con robots (Haikonen 2007a). La pregunta clave a este respecto es si se puede construir un robot que sea realmente capaz de pasar la prueba del espejo. Además, se tendría que determinar unívocamente si efectivamente el hecho de pasar esta prueba implicaría que el robot es verdaderamente autoconsciente.

Takeno et al. de la Universidad de Meiji en Japón creen que han superado la prueba del espejo con éxito al haber construido un sistema de reconocimiento de la propia

imagen reflejada en un espejo (Takeno, Inaba & Suzuki 2005). Estos autores definen cuatro pasos para sus experimentos, donde se han usado cuatro robots: el robot *yo* Rs, el robot *otro* Ro, el robot *controlado* Rc y el robot *automático* Ra. Los primeros dos robots están dotados con el sistema de reconocimiento de imágenes especulares. El tercer robot está controlado por el robot *yo*, mientras que el último se mueve automáticamente (ver Figura 15). Los cuatro experimentos planteados por Takeno se describen a continuación:

1. El *robot yo Rs* imita la acción de su propia imagen reflejada en un espejo.
2. El *robot yo Rs* imita una acción realizada intencionadamente por el *robot otro Ro* como un comportamiento de imitación.
3. El *robot controlado Rc* se controla completamente desde el *robot yo Rs* para imitar su comportamiento.
4. El *robot yo Rs* imita las acciones aleatorias del *robot automático Ra*.

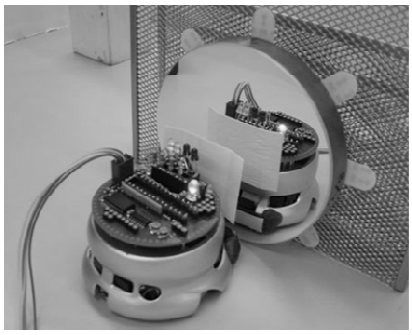


Figura 15. Robot Khepera frente a un espejo.

En estos experimentos descritos por Takeno el robot es capaz de reconocer su propia imagen reflejada en el espejo sin confundirla con la imagen de otro robot, que tiene el mismo aspecto físico. El sistema de reconocimiento de imagen especular está basado en una red de neuronas artificiales. El objetivo de este sistema es reconocer y diferenciar el comportamiento del propio robot del comportamiento de otro robot. Takeno también sugiere que la imitación es una prueba de conciencia ya que requiere el reconocimiento del comportamiento de otro sujeto y luego la aplicación de este comportamiento a uno mismo. Los resultados descritos por los autores indican que los robots pasan la prueba del espejo con una tasa del 70% de aciertos. En cualquier caso, estos experimentos no confirman exactamente que el robot diseñado sea autoconsciente. En realidad se podría decir que el robot presenta características de la conciencia S y conciencia A, puesto que es capaz de distinguir entre el comportamiento generado por sí mismo y el generado por un tercero.

El marco para la evaluación del nivel de conciencia propuesto en la presente tesis (ver Capítulo 5) pretende constituir un marco más completo que los enfoques expuestos anteriormente. Al tomar un enfoque basado en el desarrollo cognitivo el método de evaluación propuesto se puede aplicar prácticamente a cualquier tipo de sistema independientemente de su nivel de complejidad y de los mecanismos utilizados para su implementación.

2.4 Problemas éticos, sociales y legales

Aparte de los aspectos puramente científicos y tecnológicos asociados con la investigación en Conciencia Artificial, es preciso considerar también las posibles implicaciones que este campo podría tener en el ámbito social, ético y legal. De nuevo en este punto existe diversidad de opiniones: mientras que unos autores piensan que no es posible construir máquinas conscientes (Buttazzo 2001), otros argumentan que los robots conscientes existirán en un futuro próximo (Koch, Tononi 2008, Dennett 1997a). Incluso se especula con la posibilidad de que los robots conquisten el mundo y esclavicen a la humanidad (Vinge, Brin & Goertzel 2009). Otras visiones dan poco crédito a esa posibilidad, ya que contemplan la integración entre hombres y máquinas como el futuro más probable de los robots conscientes (Kurzweil 2000).

Algunos investigadores también apuntan a otros problemas éticos que podrían producirse a más corto plazo. Se especula con la posibilidad de que la experimentación en Conciencia Artificial pueda causar un sufrimiento considerable a las propias máquinas con las que se está experimentando. Metzinger llega incluso a comparar la investigación en Conciencia Artificial con el desarrollo de una raza de bebés discapacitados psíquicamente con los que se experimenta de forma indiscriminada, pudiendo causarles un sufrimiento cruel e inmoral (Metzinger 2003).

En el caso de que se consiguiera construir robots conscientes, otro punto de vital importancia sería su tratamiento legal. Actualmente, cuando un software falla, la responsabilidad recae en sus creadores, pero este punto no está tan claro en lo que se refiere a sistemas artificiales totalmente autónomos (Calverley 2005). Teniendo en cuenta la posible proliferación en el uso de robots de servicio en la sociedad de un futuro próximo, algunos gobiernos ya están empezando a preocuparse por la seguridad ciudadana. Países como Japón o Corea del Sur están trabajando en una legislación que contemple el uso masivo de robots domésticos autónomos. Recientemente, expertos en robótica japoneses han preparado un informe titulado "*Borrador de una Guía para Asegurar la Conducta Segura de la Próxima Generación de Robots*", con la intención de redactar una ley que proteja a la sociedad de forma efectiva (Lewis 2007).

2.5 Conclusiones

La comprensión de qué es y cómo se produce la conciencia es unos de los grandes retos de la ciencia moderna. La investigación en Conciencia Artificial pretende contribuir al conocimiento acerca de estas cuestiones y además pretende concebir una generación de máquinas mucho más autónomas, inteligentes y eficientes que las que se conocen en la actualidad.

Como se ha descrito en este capítulo, el estudio científico de la conciencia no está exento de controversia y diversidad de opiniones. No existe hoy en día una teoría definitiva que explique cómo se produce la conciencia. Incluso es muy complicado encontrar una definición de conciencia que satisfaga a la mayoría de los miembros de la comunidad científica. En este entorno, controvertido pero de creciente interés científico, nace la disciplina de la Conciencia Artificial, como prueba fehaciente del carácter multidisciplinar requerido en el ámbito de las ciencias cognitivas.

Todos los proyectos de Conciencia Artificial realizados hasta la fecha son relativamente pequeños y sus contribuciones científicas son modestas. Esta situación no es de extrañar si se tiene en consideración la poca madurez del campo y la pequeña envergadura de estos proyectos en términos de financiación y recursos disponibles. Aún así, se observa un claro avance que inspira la creación de nuevos proyectos e investigaciones relacionadas, como es el caso de la presente tesis doctoral.

Es significativo observar que algunas de las teorías propuestas para explicar la conciencia han tenido más impacto que otras a la hora de inspirar modelos computacionales. Este es el caso de la Teoría del Espacio de Trabajo Global. Sin embargo, es preciso tener en cuenta que muchas de las teorías expuestas en este capítulo son complementarias en algunos sentidos. Es decir, en la mayoría de los casos las diferentes teorías no ofrecen visiones contrarias acerca de la conciencia, sino que explican aspectos diferentes de la misma. En el trabajo desarrollado en esta tesis se ha evitado centrar los modelos computacionales exclusivamente en una teoría concreta, tomando ideas de diferentes enfoques, identificando las sinergias entre diversas hipótesis e intentando integrar todos estos conceptos de forma coherente.

Usando la diferenciación establecida entre conciencia A (de acceso), conciencia S (autoconciencia), conciencia M (de monitorización) y conciencia P (fenomenológica), se puede establecer una clasificación de las teorías de la conciencia descritas en base a los aspectos que abordan principalmente. La Tabla 2 resume las dimensiones de la conciencia en las que se centra cada una de las teorías analizadas. Como puede observarse, ninguna teoría es capaz de abarcar exhaustivamente todas las dimensiones de la conciencia. Aunque es cierto que todas estas dimensiones de la conciencia están relacionadas y no se pueden considerar de forma aislada más que de modo conceptual, esta clasificación pone de manifiesto la inexistencia de una "teoría unificada" de la conciencia.

Tabla 2. Principales aspectos que abordan las teorías de la conciencia analizadas.

Teoría	Conciencia A	Conciencia S	Conciencia M	Conciencia P
Hipótesis del núcleo dinámico	**X**			
Teoría de la Integración de la Información				**X**
Hipótesis del Marcador Somático			**X**	
Modelo CODAM	**X**		**X**	
Teoría del Espacio de Trabajo Global	**X**		**X**	
Teoría de las Versiones Múltiples	**X**			
Teoría de Emociones, Sentimientos y Conciencia		**X**	**X**	**X**
Enfoque Sensoriomotor			**X**	**X**
Teoría de los Pensamientos de Orden Superior			**X**	**X**
Modelo Fenomenológico del Yo		**X**		**X**
Orch-R – Reducción del Objetivo Orquestado				**X**

Un estado del arte en el campo de la Conciencia Artificial no puede concluir de otro modo más que afirmando que todavía es mucho más lo que queda por aprender que lo que se ha descubierto hasta la fecha. Si bien es cierto que los avances en el dominio de la neurobiología son muy prometedores en cuanto a la consecución de una mejor comprensión de la conciencia humana, también es cierto que queda mucho por hacer. Una de las características particulares de la investigación en Conciencia Artificial es que no se trata de un campo que se limite a tomar una inspiración biológica, sino que también puede proporcionar una retroalimentación útil a las neurociencias mediante el uso de modelos, simulaciones e implementaciones.

En cuanto al problema de la medición de la conciencia, queda patente la necesidad de modelos y técnicas más completos que permitan una evaluación lo más objetiva posible de la presencia de conciencia en sistemas artificiales. Incluso en el ámbito de la medicina, donde existen metodologías establecidas que permiten realizar una valoración del nivel de conciencia de los pacientes, queda mucho por hacer. Hay determinados casos, como los trastornos de la conciencia, donde no existe hoy en día un alto grado de confiabilidad en el diagnóstico.

Al igual que ocurre en el diagnóstico clínico, en el campo de la Conciencia Artificial, la determinación de niveles específicos es una propiedad deseable para una medida de conciencia. Tener una definición específica proporcionaría la posibilidad de realizar estudios comparativos entre diferentes modelos e implementaciones. Es más, el uso de una escala de medida establecida permitiría determinar el estado del arte en Conciencia Artificial con mayor precisión (ese es precisamente, uno de los objetivos principales de esta tesis). Además, en relación con los posibles problemas éticos, sociales y legales, la medición objetiva del nivel de conciencia puede proporcionar un marco valioso para el establecimiento de normas y buenas prácticas. Tanto en las tareas de investigación como en el posible despliegue y explotación de implementaciones de Conciencia Artificial estos aspectos deben ser tenidos en cuenta.

3 Alcance y Objetivos

Tal y como se ha explicado en el Capítulo 1, los objetivos de la presente tesis se caracterizan por la necesidad de confirmar o refutar una serie de hipótesis de trabajo planteadas acerca de la conciencia. Estas hipótesis (descritas en detalle en el Apartado 1.1) se basan en puntos de vista defendidos por corrientes filosóficas como el materialismo y el funcionalismo. En este contexto, esta tesis pretende contribuir al conocimiento científico de la conciencia mediante la aplicación de la tecnología informática y el planteamiento de nuevos modelos útiles para la experimentación en este campo. Paralelamente a la obtención de nuevo conocimiento científico sobre la conciencia, se explorará la posibilidad de la construcción de máquinas conscientes, evaluando las capacidades que algunos modelos computacionales pueden presentar en este sentido.

En esencia, se plantea la presente tesis como un esfuerzo dirigido hacia la obtención de pruebas objetivas, que usando modelos computacionales, respalden la concepción materialista-funcionalista de la mente. En otras palabras, se plantea la hipótesis general de que es posible construir máquinas con mentes conscientes sin necesidad de recurrir a un sustrato físico específico. En relación con esta visión, también se plantea el objetivo de determinar el tipo de conciencia que es posible desarrollar en una máquina. Adicionalemte, en esta tesis se argumenta que es factible al menos el desarrollo de capacidades asociadas con la conciencia A en sistemas artificiales. Sin embargo, el desarrollo de la conciencia S o la conciencia P en máquinas es un aspecto mucho más controvertido, aunque también se aborda como parte del trabajo realizado (ver Capítulo 6).

En relación al conocimiento científico sobre la conciencia y la posibilidad del desarrollo de la conciencia en sistemas artificiales, se perseguirán los siguientes objetivos:

- **Objetivo I. Demostrar que es posible estudiar la conciencia fenomenológica (conciencia P) mediante la aplicación de modelos computacionales.** Aunque la comprobación de la existencia de estados fenomenológicos en sistemas artificiales supone un gran problema, es posible al menos desarrollar modelos capaces de especificar el contenido de la experiencia consciente. La aplicación de los principios de la Fenomenología Sintética permitirá simular y modelar cuál sería el contenido de la experiencia de un sujeto consciente expuesto a los mismos estímulos que el modelo computacional.

- **Objetivo II. Demostrar, mediante el uso de modelos computacionales, que la conciencia tiene una función integradora y adaptativa.** La aplicación de un modelo computacional basado en la conciencia permitirá demostrar cómo diferentes capacidades cognitivas pueden ser integradas de forma efectiva. La generación artificial de qualia se considerará como el resultado de la sinergia emergente de la combinación efectiva de las funciones cognitivas.
- **Objetivo III. Demostrar que la conciencia se puede caracterizar mejor como un proceso dinámico y no como una propiedad de un sistema.** Usando los resultados obtenidos mediante la aplicación de un modelo computacional de la conciencia, se pretenderá demostrar que la naturaleza de la conciencia es comparable a la de un proceso software. Se intentará demostrar que no es necesario recurrir a un sustrato físico especial para desarrollar una mente que observa el mundo de forma subjetiva. Es decir, lo esencial para la generación de la experiencia subjetiva son los procesos computacionales asociados y no el sustrato material que da soporte a esa computación.
- **Objetivo IV. Demostrar que el "poder cognitivo" de un sistema artificial no se correlaciona con su potencia computacional.** Aunque intuitivamente se pueda pensar que es necesaria una gran potencia computacional para poder desarrollar capacidades cognitivas superiores en una máquina, se intentará demostrar que el problema radica principalmente en el diseño de la arquitectura de control. En otras palabras, se intentará demostrar que la investigación en Conciencia Artificial debe orientarse a la mejora del diseño de las arquitecturas cognitivas (y no en su capacidad bruta para el procesamiento de la información).
- **Objetivo V. Demostrar que es posible estudiar la Conciencia Artificial de forma estrictamente científica aplicando los principios de la heterofenomenología.** Frente al problema de la imposibilidad de estudiar de forma científica los estados puramente privados característicos de la subjetividad consciente, se tratará de abordar el problema del estudio científico de la Conciencia Artificial mediante un enfoque heterofenomenológico. Es decir, combinando la observación objetiva (comportamiento e inspección del funcionamiento interno) con la especificación del contenido consciente generado por el propio sistema que está siendo analizado.
- **Objetivo VI. Demostrar que es posible medir el nivel de conciencia de un sistema artificial, establecer un estado del arte en Conciencia Artificial y comparar objetivamente las diferentes implementaciones existentes.** Se pretende definir una serie de niveles de conciencia atendiendo a las capacidades cognitivas demostradas por el sistema que se esté evaluando. De este modo se podrían evaluar de forma ecuánime los modelos de Conciencia Artificial existentes e incluso establecer cuál es el nivel máximo alcanzado en la actualidad.

Para poder cumplir los objetivos mencionados anteriormente se requiere el desarrollo de herramientas que permitan la experimentación correspondiente. Para la consecución del Objetivo I se requiere diseñar e implementar un modelo computacional de la conciencia capaz de simular y especificar el contenido consciente que él mismo genera. Asimismo, es necesario disponer de un modelo de referencia que permita el estudio y el análisis de los posibles qualia artificiales. El Capitulo 6 aborda el problema de la definición, análisis, generación y especificación de la experiencia consciente en máquinas.

Para la consecución del Objetivo II es necesario contar con un modelo computacional capaz de integrar diferentes funciones cognitivas asociadas con la conciencia. Los objetivos III, IV y V también requieren un modelo computacional de la conciencia cuyo funcionamiento y rendimiento puedan ser analizados en detalle. En general, para poder realizar la experimentación correspondiente a los primeros cinco objetivos es necesario desarrollar un modelo computacional que sirva como banco de pruebas para la investigación en Conciencia Artificial y Fenomenología Sintética. Por lo tanto, como parte del trabajo de la presente tesis doctoral, se diseñará y desarrollará una arquitectura cognitiva artificial, inspirada en diversas teorías de la conciencia, y que tenga la flexibilidad requerida para servir a los propósitos descritos en los objetivos I, II, III, IV y V. Esta arquitectura cognitiva, denominada CERA-CRANIUM, se describe en detalle en el Capítulo 4.

Para la consecución del Objetivo VI (y en gran medida también para el Objetivo V) se necesita definir un marco de evaluación y caracterización del potencial cognitivo de los modelos y las implementaciones de Conciencia Artificial. En consecuencia, parte del trabajo de esta tesis doctoral se centrará en la definición de una escala que sirva para medir el desarrollo de las capacidades cognitivas asociadas con la conciencia. Este marco de evaluación, denominado *ConsScale*, se describe en detalle en el Capítulo 5.

En relación con los objetivos planteados, además de las herramientas necesarias, es importante determinar también el dominio concreto de aplicación y el diseño de los experimentos. En este sentido, la presente tesis se centrará en las posibles aplicaciones de la Conciencia Artificial en el dominio de la robótica cognitiva, planteándose el uso tanto de robots reales como robots o personajes simulados por ordenador en entornos virtuales.

Dado que los experimentos requeridos pueden alcanzar grados muy altos de complejidad, se tratará de especificar dominios de problema bien acotados, que aunque limitando las posibilidades de aplicación práctica a corto plazo, puedan ofrecer resultados significativos en cuanto a las hipótesis planteadas sobre la conciencia. Como norma general, y con el objetivo de mantener unos niveles de complejidad manejables, se diseñarán experimentos equivalentes a los realizados con humanos, pero adaptando el entorno para mitigar fuentes de ruido innecesarias (ver Capítulo 7). En cualquier caso, las herramientas desarrolladas durante la tesis deberán servir para la experimentación en diferentes entornos y aplicadas a diferentes dominios de problema.

Teniendo en cuenta los puntos anteriores, se han determinado los siguientes requisitos funcionales para el desarrollo de la arquitectura cognitiva CERA-CRANIUM:

- El modelo computacional de CERA-CRANIUM debe permitir la integración efectiva de diversas funciones cognitivas. Es decir, el modelo permitirá integrar en un mismo sistema mecanismos para funciones cognitivas como la atención, la gestión de memoria a corto plazo, el cambio de contexto o la comunicación del estado mental interno.
- Las implementaciones del modelo CERA-CRANIUM deben poder ser empleadas como sistema de control de un robot autónomo real (o en su defecto, de un robot simulado o de otro tipo de agente software).
- CERA-CRANIUM debe poder funcionar con agentes dotados de diferentes capacidades sensoriomotoras. Es decir, la arquitectura ha de manejar diferentes tipos de sensores y de actuadores (tanto reales como simulados).

- CERA-CRANIUM debe disponer de un mecanismo para especificar y comunicar de forma precisa la parte de su estado interno correspondiente al contenido de la experiencia consciente. En otras palabras, CERA-CRANIUM tiene que ser una plataforma válida para la experimentación en Fenomenología Sintética.

Para la definición del marco de evaluación *ConsScale* se tendrán en cuenta los siguientes requisitos:

- *ConsScale* debe definir una jerarquía de niveles de conciencia funcional que sirva para analizar, caracterizar y evaluar cualquier arquitectura cognitiva artificial, independientemente del dominio de aplicación para el que haya sido diseñada.
- La definición de *ConsScale* debe contemplar un método de evaluación preciso que permita su aplicación de forma sistemática.
- *ConsScale* debe proporcionar unas medidas cualitativas y cuantitativas precisas que permitan caracterizar la capacidad cognitiva global tanto de un modelo computacional teórico como de una implementación particular.

En definitiva, el trabajo realizado para la consecución de los objetivos I a VI debe conducir al establecimiento de nuevo conocimiento científico acerca de la veracidad de las hipótesis de trabajo planteadas:

- La conciencia fenomenológica también se puede estudiar mediante modelos computacionales.
- La generación artificial de qualia puede constituir una ventaja para el funcionamiento de los robots.
- Lo importante en la generación de la experiencia subjetiva son los procesos computacionales asociados, no el sustrato físico que da soporte a la computación.
- Es posible el estudio científico de la conciencia en máquinas aplicando un enfoque heterofenomenológico.
- La conciencia se manifiesta en diferentes grados y además es posible medir el nivel de conciencia funcional y la capacidad cognitiva asociada.

4 Arquitectura cognitiva CERA-CRANIUM

4.1 Introducción

Desde los inicios de la Inteligencia Artificial, los investigadores han especulado con la posibilidad de construir máquinas inteligentes que pudieran tener las mismas capacidades intelectuales que los humanos, o incluso superiores. Como se ha visto en el capítulo anterior, los científicos están de acuerdo en que una de las claves principales para construir este tipo de máquinas, equiparables a los humanos, reside en la conciencia y las capacidades cognitivas asociadas. En este ámbito, la investigación en Conciencia Artificial pretende dar respuesta a preguntas como las siguientes:

- ¿Pueden pensar los ordenadores o simplemente calculan?
- ¿Es la conciencia una cualidad que sólo puede estar presente en los humanos?
- ¿Depende la conciencia del material que compone el cerebro humano, o se puede generar en el hardware y/o software de un ordenador?

Tanto en el dominio de la ingeniería como en el de la biología, el reto al que se enfrenta la comunidad científica sigue siendo la unión entre lo mental y lo físico. Sigue existiendo un vacío en las teorías, de momento insalvable, que impide explicar de forma concluyente cual es la relación concreta que existe entre mente y cerebro (Levine 1983). Por consiguiente, tampoco se conoce la posible relación que podría existir entre la mente de un hipotético robot consciente y su hardware y software asociados. En la presente tesis se pretende arrojar luz sobre esta cuestión mediante el diseño, construcción y evaluación de un modelo computacional de la conciencia. Se espera que los resultados obtenidos durante la experimentación aporten nuevo conocimiento acerca de la conciencia y la posibilidad de su existencia en sistemas artificiales.

El marco computacional que se describe en este capítulo se basa en la hipótesis de que la conciencia ha aparecido y se ha desarrollado en la especie humana porque proporciona una serie de funcionalidades útiles para la supervivencia (ver Apartado 2.1.3.3). Mediante la implementación y la experimentación con el modelo propuesto se pretende confirmar cuáles son estas funcionalidades (o capacidades cognitivas) asociadas con la conciencia, cómo se pueden implementar de forma artificial y cómo se pueden articular las sinergias que potencialmente existen entre las mismas.

Una de las principales hipótesis de trabajo planteadas en esta tesis doctoral es que se pueden construir modelos de Conciencia Artificial basados en técnicas que no emulan directamente el funcionamiento a nivel neuronal del tejido nervioso. Es decir, se pueden confeccionar modelos que representen un nivel de descripción más alto, concretamente adoptando el punto de vista de la psicología cognitiva, donde se identifican funciones mentales y los componentes arquitectónicos asociados. Por lo tanto, el marco de trabajo propuesto en la presente tesis se basa en la experimentación con Modelos Cognitivos de Conciencia Artificial (MCCA). Específicamente, se ha diseñado un marco para la experimentación con MCCA que se basa en una arquitectura cognitiva genérica, que se puede integrar en entornos diferentes y se puede aplicar a dominios de problema diversos. La versatilidad del marco propuesto radica en la definición de un sistema de control cognitivo genérico para agentes abstractos. Es decir, un sistema que pueda servir como mecanismo de control para cualquier agente situado, pudiendo ser éste un robot real, un robot simulado o un agente software.

El marco propuesto está inspirado en las principales teorías cognitivas de la conciencia y proporciona mecanismos tanto para la adquisición de información sensorial como para la ejecución de acciones motoras. En general, las teorías cognitivas descritas en el Capítulo 2 (ver Apartado 2.1.4) se basan principalmente en metáforas que simplemente ayudan a comprender el funcionamiento de la mente humana de forma intuitiva. Si bien es cierto que una simple metáfora está muy lejos de constituir un cuerpo establecido de conocimiento científico, puede servir como herramienta útil para dirigir las líneas de investigación en direcciones concretas, que afirmen o desmientan las hipótesis planteadas. Sin embargo, desde el punto de vista del ingeniero, estas teorías, que típicamente provienen de la psicología y la filosofía, no proporcionan una explicación práctica acerca de cómo se podría producir la conciencia en una máquina. Es decir, estas teorías no se pueden traducir directamente en modelos computacionales, es necesario realizar una interpretación funcional específica para poder llegar al nivel de detalle requerido para realizar una implementación. Por otro lado, desde el punto de vista de la construcción de estos MCCA, el uso de estos esquemas simplificados de la conciencia tiene dos ventajas claras: la facilidad de implementación y la comprensión de los modelos; y por ende la posibilidad de experimentación con sistemas artificiales fácilmente observables, parametrizables y relativamente asequibles.

Para el diseño del marco de experimentación propuesto se han tomado como referencia las principales teorías cognitivas sobre la conciencia (ver Apartado 2.1.4). Específicamente, la Teoría del Espacio de Trabajo Global (Baars 1988) y el Modelo de las Versiones Múltiples (Dennett 1991) tienen aspectos en común que se han tomado como guía de diseño para una arquitectura capaz de integrar diferentes MCCA. Aunque los autores tienden a usar nombres y descripciones diferentes, las teorías cognitivas de la conciencia coinciden en la hipótesis de que la unidad del yo (o *self*) que se produce en los seres conscientes tiene su origen en mecanismos que no son en sí unitarios. Concretamente, estas teorías argumentan que los contenidos conscientes emergen como resultado de procesos de competición y colaboración entre procesadores especializados (Dennett 1991, Baars 1997, Minsky 1988, Hofstadter 1995, Shanon 2008). Las diferentes teorías ofrecen explicaciones o metáforas diferentes acerca de la forma específica en la que estos procesos de competición y colaboración podrían tener lugar. Sin embargo, todas las teorías coinciden en que su naturaleza es muy dinámica y presentan un alto grado de adaptación.

Dado que ninguna de estas teorías explica la forma detallada en la que se produce la conciencia, cada MCCA existente adopta un enfoque diferente en cuanto a la forma en

que los flujos de percepción y acción se construyen y se gestionan. Sin embargo, existe un denominador común en relación al mecanismo subyacente que se usa para realizar los procesos cognitivos de bajo nivel: *la colaboración y competición concurrente de multitud de procesadores especializados en un espacio de trabajo compartido*. Lo que difiere de un modelo a otro es la técnica específica que se emplea para orquestar los procesos de colaboración y competición. Asimismo, cada implementación particular generalmente está orientada a un entorno y a un dominio de problema específicos, haciendo muy complicada la tarea de comparar su rendimiento relativo.

El marco de experimentación descrito en este capítulo constituye un entorno en el que se proporcionan unos servicios comunes de ejecución concurrente de procesadores especializados, permitiendo que el MCCA concreto que se integre en la arquitectura cognitiva dirija de forma específica los procesos de competición y colaboración, dando lugar a flujos de percepción y acción particulares. Adicionalmente, tomando en consideración la necesidad de evaluar las propiedades de los MCCA bajo diversas condiciones de configuración y enfrentados a diversos dominios de problema, se ha implementado una arquitectura software genérica que también haga posible la aplicación práctica en entornos heterogéneos de los modelos de conciencia planteados. Utilizando esta arquitectura, un modelo de conciencia se puede integrar en dominios de aplicación dispares, como por ejemplo un sistema de control de un robot autónomo móvil o el control de un personaje sintético de un videojuego (Arrabales, Ledezma & Sanchis 2007, Arrabales, Ledezma & Sanchis 2009b). En general, para poder realizar experimentos de forma efectiva, no sólo se necesita implementar el propio modelo de conciencia, sino que también es necesario integrarlo en un sistema completo que afronte un determinado problema. En este sentido, la arquitectura software implementada también proporciona la posibilidad de enfrentar un mismo MCCA a diferentes problemas (ver Figura 16).

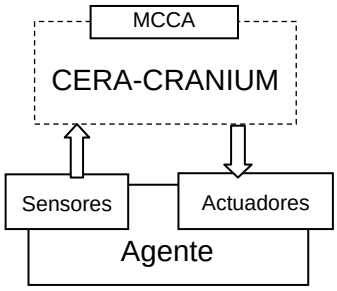


Figura 16. Esquema simplificado del marco de experimentación CERA-CRANIUM.

La plataforma de experimentación CERA-CRANIUM permite la evaluación y comparación de diferentes enfoques cognitivos (MCCA). Para hacer posible esto se ha implementado una arquitectura cognitiva genérica, pero configurable, basada en espacios de trabajo global que se distribuyen en capas. Esta plataforma proporciona las características funcionales principales de una arquitectura cognitiva de propósito general, que se puede usar como base para la realización de modelos computacionales de la conciencia de más alto nivel. En el diseño de CERA-CRANIUM se han tomado en cuenta los diferentes procesos de creación, asociación, combinación y competición de procesadores especializados que se describen en las teorías mencionadas,

implementando los mecanismos necesarios para regular estos procesos. Los componentes principales de la plataforma CERA-CRANIUM son:

- CERA (*Conscious and Emotional Reasoning Architecture*). Una arquitectura de control de inspiración cognitiva estructurada en capas.
- CRANIUM (*Cognitive Robotics Architecture Neurologically Inspired Underlying Manager*). Una herramienta para la creación y gestión de grandes cantidades de procesos paralelos en espacios de trabajo compartidos.

Como se ha descrito anteriormente, se ha adoptado una perspectiva cognitiva frente al problema de la Conciencia Artificial. Sin embargo, esto no significa que se ignore el problema de la fenomenología. Gracias a la posibilidad de implementar diversas variaciones de MCCA se podrán aplicar los principios de la Fenomenología Sintética para estudiar la capacidad que tienen los modelos cognitivos estudiados de especificar los contenidos de la experiencia consciente (ver Capítulo 6).

A continuación se describen los principales componentes software que se han diseñado y desarrollado con el objetivo de disponer de un marco para la experimentación con MCCA. Estos componentes son clave para la realización de los experimentos descritos en el Capítulo 7, cuyos resultados han proporcionado retroalimentación útil sobre los modelos de conciencia planteados.

4.2 CERA: Una arquitectura cognitiva inspirada en la conciencia

CERA es una arquitectura cognitiva para agentes autónomos, está estructurada en capas y diseñada para implementar un sistema de control gobernado por un MCCA. CERA se compone de cuatro capas (ver Figura 17): capa de servicios sensoriomotores, capa física, capa de misión y capa núcleo. Análogamente a las arquitecturas clásicas de subsunción, el nivel de abstracción es mayor en las capas superiores. Sin embargo, la definición de las capas en CERA no está asociada directamente con comportamientos específicos.

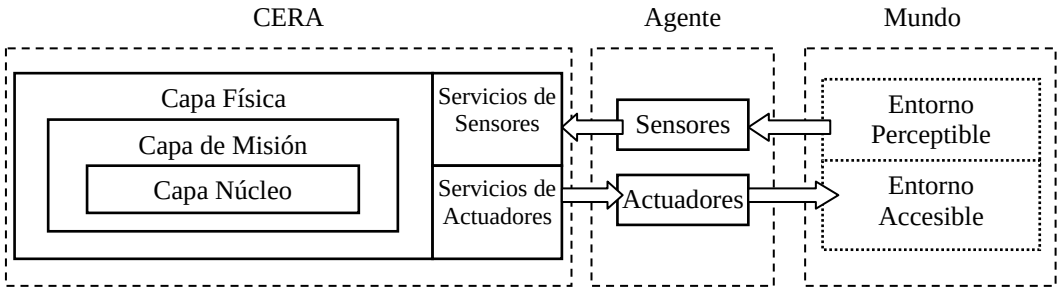


Figura 17. Diseño estructurado en capas de la arquitectura cognitiva CERA.

A continuación se especifican las funcionalidades asociadas a cada una de las capas de CERA:

- **Capa de servicios sensoriomotores.** Esta capa se compone de un conjunto de servicios de comunicación y de interfaz que implementan los métodos de

acceso a los datos de los sensores y el envío de comandos a los actuadores. Para poder dotar a un agente de un sistema de control autónomo total usando CERA cada sensor debe tener su correspondiente servicio de sensor; de forma análoga, cada actuador debe tener su correspondiente servicio de actuador. Estos servicios proporcionan a la capa física un interfaz de acceso uniforme a los mecanismos físicos (o simulados) del agente.

- **Capa física.** Esta capa contiene las representaciones de bajo nivel correspondientes a los sensores y los actuadores del agente. Adicionalmente, dependiendo de la naturaleza de los datos sensoriales adquiridos, la capa física realiza tareas de preparación y preprocesamiento de los datos. En esta capa también se realizan tareas análogas con los comandos dirigidos a los actuadores, asegurándose, por ejemplo, que los parámetros de los comandos se encuentran dentro de los márgenes de seguridad ("necesidades vitales" o requisitos de integridad funcional del agente). Aunque en este nivel no se realiza fusión de información sensorial, determinados datos sensoriales pueden agruparse y se calculan y anotan parámetros de contextualización de bajo nivel, como las posiciones relativas de los objetos detectados y las marcas de tiempo correspondientes.
- **Capa de misión.** La capa de misión produce y gestiona contenido sensoriomotor más elaborado y que está relacionado con sus misiones particulares (típicamente, una misión involucra la persecución de varias metas). Es en esta etapa cuando los contenidos simples adquiridos y preprocesados en la capa física se combinan para formar contenidos multimodales más complejos, que tienen significados específicos relacionados con las metas del agente. La capa de misión se puede modificar independientemente de otras capas de la arquitectura de forma que se adapte a las tareas específicas asignadas al agente (dominio de problema).
- **Capa núcleo.** Esta capa constituye el nivel más alto de control en CERA y contiene un conjunto de módulos que regulan funciones cognitivas superiores. La definición e interacción de estos módulos se puede ajustar de forma que se implemente una determinada variación de un MCCA. Algunos de los posibles módulos identificables son: el módulo de atención, el módulo de evaluación del estado propio, el módulo de gestión preconsciente, el módulo de gestión de la memoria y el módulo de auto-coordinación. Sin embargo, CERA está diseñada para permitir la definición personalizada de módulos de la capa núcleo. Los objetivos de estos módulos se describen más adelante, así como los mecanismos que usan para regular la forma en que operan las capas inferiores de CERA.

Para realizar un experimento usando la plataforma propuesta, es necesario establecer un valor concreto para estas cuatro variables:

- El MCCA que se quiere evaluar y su configuración específica.
- El agente (físico o simulado) que se va a emplear.
- La misión asignada al agente seleccionado.
- El entorno en el que se desenvolverá el agente seleccionado.

Este proceso de instanciación implica:

- La definición de los módulos de la capa núcleo.

- La implementación de interfaces específicos para los sensores y actuadores del agente seleccionado (capa de servicios sensoriomotores).
- Definición de las rutinas específicas de misión (capa de misión).

Para el caso particular de un estudio comparativo de diferentes MCCA, las variables entorno, misión y agente tienen que permanecer constantes. Por lo tanto, los únicos cambios necesarios a la hora de evaluar un nuevo MCCA han de realizarse exclusivamente en la capa núcleo.

Las capas física y de misión se caracterizan por la inspiración en las teorías cognitivas de la conciencia, donde un gran número de procesos paralelos compiten y colaboran en un espacio de trabajo compartido en busca de una solución global. De hecho, un agente controlado por la arquitectura CERA cuenta con dos espacios de trabajo compartido (o espacios de trabajo CRANIUM) organizados jerárquicamente y que operan en coordinación, ambos con el objetivo de encontrar dos soluciones globales e interrelacionadas: una se refiere a la percepción y otra a la acción (ver Figura 18). Como se explica en el siguiente apartado, CRANIUM se usa para la implementación de estos espacios de trabajo de acuerdo a los requisitos establecidos en la arquitectura CERA. En resumen, CERA ha de proporcionar respuesta de forma continua a las siguientes preguntas:

- ¿Cuál debe ser el próximo contenido de la percepción consciente del agente?
- ¿Cuál debe ser la siguiente acción a ejecutar?

Las arquitecturas clásicas de control de robots se centran en la segunda cuestión e ignoran la primera. En la presente tesis se mantiene la hipótesis de que es necesario un mecanismo para contestar de forma efectiva a la primera pregunta para así poder contestar a la segunda pregunta de forma similar a como lo hacen los humanos (ver Capítulo 6). En cualquier caso, ambas preguntas han de ser respondidas teniendo en consideración criterios de seguridad y la misión asignada al agente. Consecuentemente, se espera que CERA sea capaz de encontrar respuestas óptimas que eventualmente puedan conducir a la generación de un comportamiento equiparable al humano.

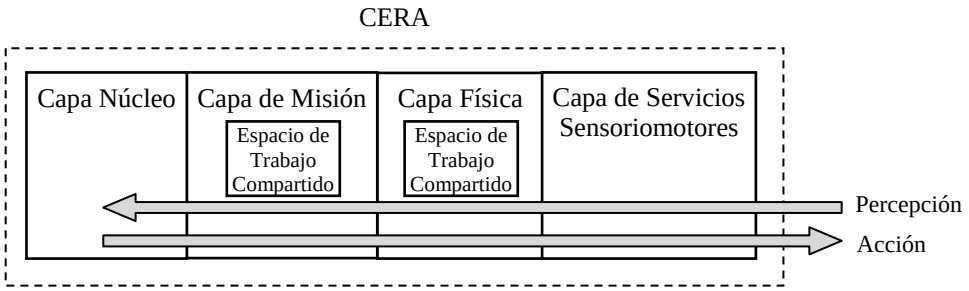


Figura 18. Espacios de trabajo compartido en la arquitectura cognitiva CERA.

4.3 CRANIUM: Un entorno de ejecución de procesadores especializados.

CRANIUM proporciona unos servicios software con los que CERA puede ejecutar multitud de procesos asíncronos de forma concurrente, pero coordinada. Además de

seguir las guías de diseño inspiradas en las principales teorías cognitivas de la conciencia tal y como se ha expuesto en el apartado anterior, CRANIUM también está inspirado en la forma en la que trabaja el cerebro humano a nivel de interacción funcional entre sus distintas áreas especializadas. Las regiones especializadas del cerebro procesan la información proveniente tanto de los sentidos como de otras regiones especializadas. De acuerdo con la Hipótesis de Acceso Global (Baars 2002), las conexiones neuronales entre las áreas especializadas del cerebro hacen posible que emerja una coordinación global. CRANIUM proporciona un servicio denominado Espacio de Trabajo CRANIUM (ETC) que permite reproducir esta dinámica de interacción funcional entre múltiples procesadores especializados.

Se puede considerar un ETC como una implementación particular de un *pandemonio*, en el sentido descrito por Dennett (1991), donde los *demonios* (o procesos especializados) compiten entre ellos para conseguir niveles altos de activación. Cada uno de estos demonios o procesos especializados se diseña para realizar una función específica sobre ciertos tipos de datos. En un momento dado, el nivel de activación de un procesador particular se calcula en base a una estimación heurística de en qué medida puede contribuir su salida a la solución global que se está buscando en el ETC. Los parámetros concretos que se usan para realizar esta estimación se establecen en la capa núcleo de CERA tal y como se explica más adelante. Como regla general, el funcionamiento de un ETC se regula constantemente gracias a comandos de modulación que se envían desde la capa núcleo de CERA.

En el marco propuesto se emplean dos ETC separados pero conectados (ver Figura 18). El ETC de nivel más bajo se localiza en la capa física de CERA, donde los procesadores especializados se alimentan de datos provenientes de los servicios de sensor de la capa inferior. Gracias a la dinámica de funcionamiento del ETC los procesadores especializados también se retroalimentan con la propia información que se genera en el espacio de trabajo compartido. El segundo ETC, que está localizado en la capa de misión de CERA, es el área de trabajo de procesadores especializados de más alto nivel, que toman como entrada tanto la información proveniente de la capa física como la información que se produce en el propio ETC. El flujo de información perceptual se organiza en paquetes llamados *perceptos simples*, *perceptos complejos* y *perceptos de misión* (ver Figura 19).

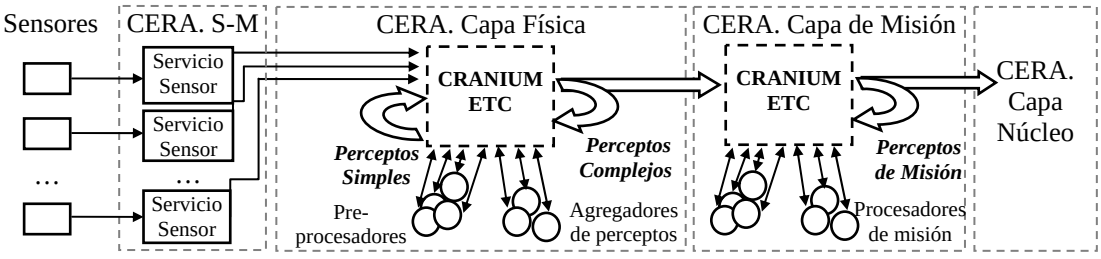


Figura 19. Flujo ascendente de procesos de percepción en CERA-CRANIUM.

A parte del flujo ascendente de procesos de percepción, CERA-CRANIUM también ofrece mecanismos para crear un flujo descendente que en última instancia genere las acciones del agente. Para articular estos procesos de acción, los ETC de la

capa física y de la capa de misión incluyen unas representaciones llamadas: *acciones simples*, *comportamientos simples* y *comportamientos de misión* (ver Figura 20).

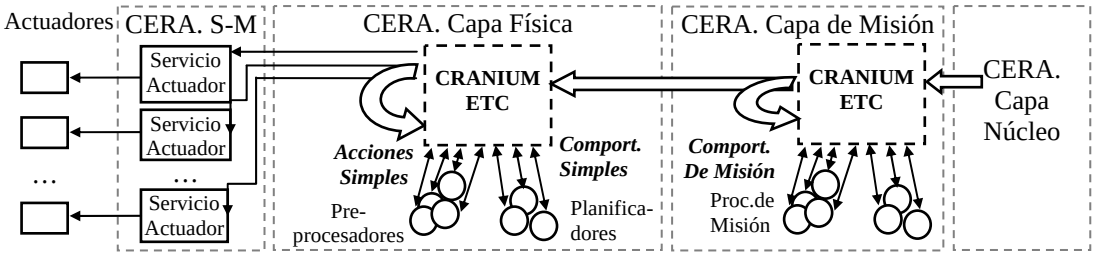


Figura 20. Flujo descendente de procesos de acción en CERA-CRANIUM.

Una de las diferencias principales entre los flujos ascendente y descendente en CERA-CRANIUM radica en el hecho de que mientras la complejidad y el significado de los perceptos aumentan iterativamente, los comportamientos de alto nivel son por el contrario descompuestos hasta que se obtiene una secuencia de acciones atómicas. Tal y como se describe más adelante, existen diferentes tipos de procesadores especializados que se pueden implementar y asociarse a un ETC.

Aunque en general cualquier procesador especializado podría tomar como entrada tanto perceptos como comportamientos, normalmente los procesadores están diseñados para:

- Generar perceptos más complejos a partir de perceptos simples.
- Generar secuencias de acciones simples a partir de definiciones de comportamientos complejos.
- Generar comportamientos en función de los perceptos recibidos.

Por ejemplo, se pueden generar rápidamente respuestas reactivas de alta prioridad en la capa física, típicamente sin ninguna intervención de las capas superiores. Si un percepto complejo que representa una amenaza física aparece en el ETC de la capa física, el procesador especializado a cargo de detectar este tipo de amenazas se activará y generará un comportamiento simple reactivo como respuesta (ver Figura 21). El comportamiento simple de evasión que se obtiene será seleccionado y se ejecutará la correspondiente secuencia de acciones en la capa de servicios sensoriomotores.

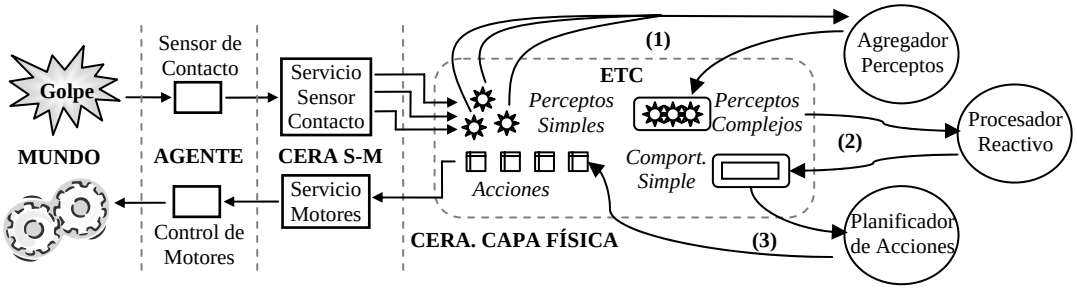


Figura 21. Mecanismo de reflejo en un robot móvil controlado por CERA-CRANIUM.

Un ETC proporciona un acceso compartido y global a una memoria de trabajo donde los procesadores especializados asociados pueden leer y escribir datos. En el caso particular descrito en la Figura 21, se generan tres perceptos simples que se añaden al ETC de la capa física como consecuencia de un impacto detectado casi simultáneamente en tres sensores de contacto contiguos de un robot móvil. El procesador agregador de perceptos lee estos tres perceptos simples y construye un nuevo percepto complejo a partir de los datos de los tres perceptos originales (los detalles del proceso de agregación se describen en el Apartado 4.5). El conjunto concreto de perceptos simples que se selecciona para formar el nuevo percepto complejo se determina gracias a la aplicación de múltiples criterios de contexto. Una vez que el nuevo percepto complejo se ha generado y publicado en el ETC, un procesador reactivo puede leerlo y detectar la existencia de una condición que requiera una respuesta reactiva rápida. Si tal condición se cumple, el procesador reactivo es capaz de construir un comportamiento simple diseñado para disminuir o evitar las posibles consecuencias negativas de la situación detectada. La presencia del comportamiento simple reactivo en el ETC desencadena la activación del procesador especializado planificador de acciones, que a su vez producirá la correspondiente secuencia de acciones simples.

Tal y como se desprende del ejemplo anterior, existen diversos tipos de procesadores especializados. Todos estos procesadores trabajan como programas asíncronos independientes que son capaces de suscribirse a un ETC, leer ciertos tipos de datos del ETC, realizar procesamientos concretos con estos datos y finalmente enviar los resultados del procesamiento de vuelta al ETC, donde la representación producida por el procesador se pone a disposición del resto de procesadores especializados. Básicamente, un ETC interactuando con sus procesadores especializados asociados constituye un mecanismo clásico de Inteligencia Artificial denominado *pizarra* (Nii 1986), en el que CERA juega el papel de controlador de la pizarra.

A continuación se describen los principales tipos de procesadores especializados definidos en CERA-CRANIUM:

- **Preprocesadores de Sensor.** Estos procesadores especializados se encargan de recoger los datos sensoriales que llegan a la capa física proveniente de la capa de servicios sensoriomotores. Usando estos datos construyen los perceptos simples correspondientes combinando las lecturas de los sensores con información contextual asociada. Los perceptos simples generados se envían automáticamente al ETC de la capa física de CERA. Para realizar esta tarea, los preprocesadores de sensor también adquieren información del sistema, como por ejemplo la posición relativa de los sensores móviles del agente o el tiempo actual que marca el cronómetro.

- **Preprocesadores de Acción.** Estos procesadores especializados preparan las acciones atómicas que han generado los planificadores de acciones (otro tipo de procesadores) para que puedan entrar en el ciclo de ejecución. Básicamente, los preprocesadores de acción generan una representación denominada acción simple, que incluye datos contextuales acerca de la acción correspondiente. Por ejemplo, cada acción simple contiene una marca de tiempo precisa correspondiente al momento en que la acción se creó (se planificó) y también contiene la marca de tiempo asignada para la ejecución por el gestor de acciones (*dispatcher*) de la capa física de CERA. Esta información se usa para abortar la ejecución de acciones que se han encolado esperando a ser ejecutadas durante demasiado tiempo. La información sensorial propioceptiva también se

incluye para poder adaptar las acciones a las posiciones actuales de los actuadores.

- **Agregadores de perceptos.** Estos procesadores especializados se encargan de crear perceptos complejos a partir de perceptos simples interrelacionados. Mientras que los perceptos simples representan información sensorial atómica, los perceptos complejos son combinaciones más elaboradas y con un nivel mayor de significado que los anteriores. Pueden considerarse múltiples criterios contextuales para la formación de los perceptos complejos. Consecuentemente, diferentes tipos de agregadores de perceptos se centran en diferentes parámetros para la selección de los perceptos simples que éstos son capaces de combinar. Tan pronto como un percepto complejo se genera, los agregadores de perceptos los envían al ETC, donde otros procesadores podrían usarlos como base para subsiguientes procesos de agregación.
- **Procesadores reactivos.** Este tipo de procesadores se localizan típicamente en la capa física de CERA donde son capaces de proporcionar una respuesta rápida a estímulos que se consideran dañinos o negativos para el agente. Estos procesadores monitorizan los perceptos simple y complejos que se generan, comprobando si se dan condiciones particularmente inseguras que requieran una respuesta inmediata. Si se cumplen tales condiciones, estos procesadores crean comportamientos simples pensados para mitigar el riesgo detectado.
- **Planificadores de acciones.** Estos procesadores especializados son capaces de tomar un comportamiento como entrada y generar la correspondiente secuencia de acciones atómicas que culminarían con la realización completa del comportamiento especificado. Gracias a los planificadores de acciones todos los comportamientos activos que se encuentran en el ETC se procesan y las correspondientes secuencias de acciones se envían de vuelta al espacio de trabajo para ser ejecutadas (en el caso de que finalmente sean seleccionadas).
- **Predictores sensoriales.** Estos procesadores especializados monitorizan incesantemente una fuente específica de datos sensoriales, interpretando la secuencia de datos (secuencia de perceptos) como una señal que se puede predecir a corto plazo. Cuando una entrada sensorial que se está analizando difiere significativamente de la predicción realizada, estos procesadores crean un percepto complejo de incongruencia (*mismatch*) que se envía al correspondiente ETC. Los perceptos complejos de incongruencia se usan para centrar la atención en percepciones poco usuales.

El hecho de tener un espacio de trabajo compartido, en el que convergen los flujos de percepción y de acción, facilita la implementación de múltiples lazos de retroalimentación (ver Figura 22). Estos lazos son muy útiles para generar un comportamiento adaptado y efectivo. El comportamiento simple ganador se enfrenta continuamente con las nuevas opciones generadas en la capa física, proporcionando así un mecanismo para la interrupción de comportamientos en progreso tan pronto como se considera que ya no son la mejor opción disponible.

En términos generales, CERA modula la activación y la inhibición de los procesos de percepción y de acción de acuerdo al MCCA implementado. La Figura 22 muestra una representación esquemática de los típicos lazos de retroalimentación y control que se crean en la arquitectura CERA-CRANIUM. Estos lazos se cierran cuando el agente

percibe las consecuencias de una acción, desencadenando respuestas adaptativas a diferentes niveles.

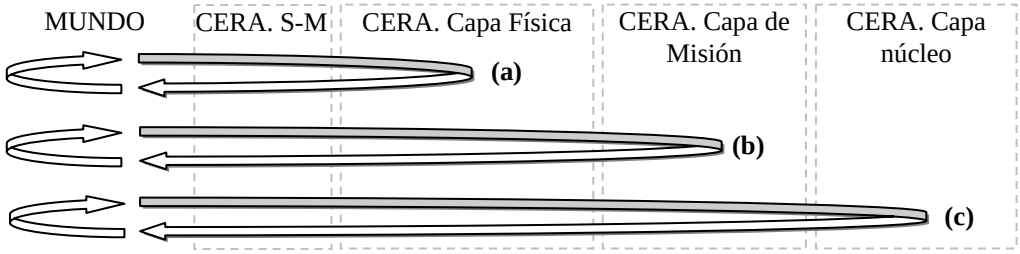


Figura 22. Diferentes lazos de control que se producen en CERA-CRANIUM.

La flecha (a) de la Figura 22 representa el lazo de retroalimentación producido cuando se desencadena un reflejo instintivo, como en el ejemplo descrito en la Figura 21. La flecha (b) corresponde a una situación en la que un comportamiento específico de misión se está realizando de forma inconsciente. Finalmente, la flecha (c) simboliza el lazo de control de más alto nivel, en el que una tarea se desarrolla de forma consciente. Estos tres tipos de lazos de control no se excluyen mutuamente. De hecho, los mismos perceptos contribuirán típicamente a lazos de control simultáneos que se producen a diferentes niveles.

Como se ha explicado anteriormente, la implementación del marco computacional especificado en CERA-CRANIUM tiene fuertes requisitos en términos de concurrencia y gestión de la entrada/salida asíncrona. La arquitectura software diseñada para dar soporte a este tipo de procesos se describe en el Apartado 4.4.

4.4 Arquitectura software de CERA-CRANIUM

Los conceptos sobre el diseño de CERA-CRANIUM descritos anteriormente se refieren a la arquitectura de nivel cognitivo. Sin embargo, construir un sistema de este tipo también requiere un diseño a nivel de la arquitectura software subyacente. Aplicar una buena estrategia de ingeniería del software permitirá disponer de un marco de aplicación robusto, fácil de mantener y de ampliar y reutilizar. Adicionalmente, el rendimiento y la escalabilidad son factores que no se pueden ignorar en este contexto debido a los requisitos computacionales asociados a la ejecución concurrente de un gran número de procesadores especializados. Teniendo estos factores en consideración, así como la necesidad de disponer de un simulador físico potente, se ha seleccionado la plataforma de desarrollo Robotics Developer Studio (RDS) (Microsoft, Corp. 2006) para la implementación de CERA-CRANIUM.

El soporte en tiempo de ejecución (runtime) de RDS se basa en dos componentes principales: el soporte de coordinación y concurrencia o CCR – *Concurrency and Coordination Runtime* (Richter 2006), que se usa para la programación asíncrona y la gestión de la concurrencia de los procesadores especializados; y el soporte de servicios descentralizados o DSS – Decentralized Software Services (Nielsen,

Chrysanthakopoulos 2006), que proporciona un marco para la implementación de arquitecturas orientadas a servicios.

Aplicando el paradigma de la orientación a servicios (Singh, Huhns 2005), cada capa de CERA se ha definido como un servicio independiente que, en caso de ser necesario, puede ejecutarse en un ordenador separado. Consecuentemente, la comunicación entre las capas se implementa usando el protocolo DSS (DSSP), que funciona sobre TCP/IP. El estilo de la arquitectura de orientación a servicios que se ha seguido para implementar CERA-CRANIUM se basa en el modelo REST – *Representational State Transfer* (Nielsen, Chrysanthakopoulos 2006), lo que implica que el comportamiento de cada servicio se define en términos de un *contrato de servicio*. Este contrato se refiere en esencia a un conjunto de operaciones que el servicio es capaz de realizar. Los mensajes que se envían entre los servicios invocan a estas operaciones y transportan una parte o todo el estado interno del servicio.

La Figura 24 proporciona una visión simplificada de los principales componentes arquitecturales de CERA-CRANIUM y su esquema de comunicación (los círculos entre los módulos indican que se produce comunicación asíncrona usando mensajes DSSP entre diferentes servicios). Los componentes que se comunican usando DSSP se pueden distribuir en diferentes máquinas. Asimismo, los componentes identificados en la Figura 24 constituyen elementos de compilación independientes.

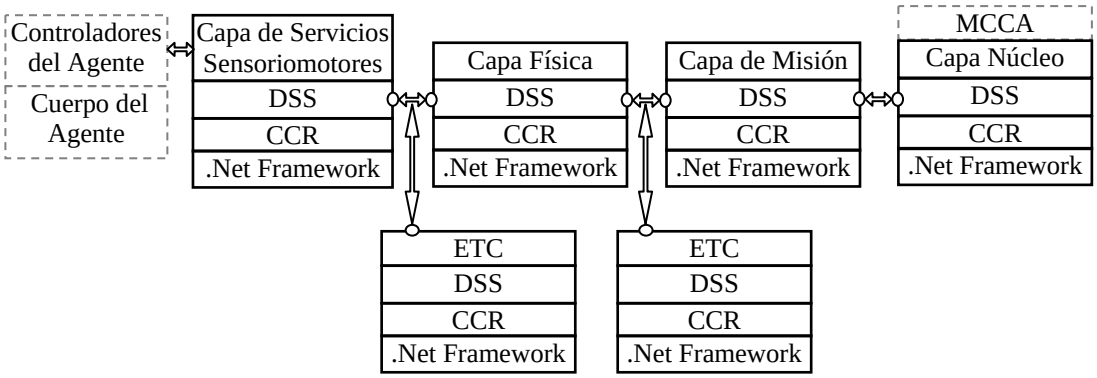


Figura 23. Arquitectura software de CERA-CRANIUM.

Como se puede observar en la Figura 24, los ETC también se han definido como servicios independientes cuyo estado interno está constituido en gran parte por la memoria de trabajo del sistema (el espacio de trabajo compartido). Cada uno de los procesadores especializados se implementa también como un servicio (ver Figura 24). Estos servicios son capaces de suscribirse al servicio ETC para recibir notificaciones de los nuevos perceptos que entran en la memoria de trabajo. Asimismo, los procesadores pueden enviar su salida al ETC de forma análoga.

Cada ETC contiene un gestor de hilos de alto rendimiento. El gestor asigna los hilos concurrentes a las tareas de asignación de perceptos a los procesadores especializados según su prioridad. Los procesadores especializados ejecutan sus tareas sobre los datos recibidos en servicios independientes. Gracias a esta arquitectura orientada a servicios se aprovechan todas las CPU disponibles en el sistema para ejecutar de forma asíncrona y concurrente todas las tareas descritas. Los servicios "Foco" representados en la Figura 24 contienen la selección de contenidos que pasa a la

capa superior. De esta forma, la capa de misión sólo accede en cada momento a una selección reducida de contenidos. Los contenidos seleccionados en el foco de atención depende de la señal de control emitida en forma de comandos desde la capa núcleo (ver Apartado 4.6).

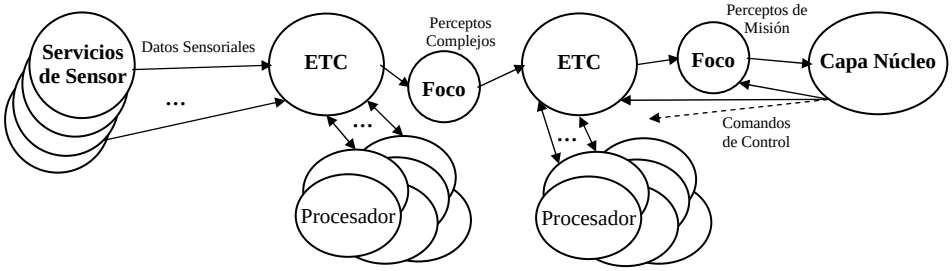


Figura 24. Comunicación entre los servicios principales de CERA-CRANIUM.

4.5 Representación del conocimiento en CERA-CRANIUM

La representación del conocimiento es una de las grandes cuestiones que permanecen sin resolver en el ámbito de la Inteligencia Artificial, y por consiguiente también en el campo de la Conciencia Artificial. Cualquier modelo de Conciencia Artificial debe proporcionar un mecanismo satisfactorio para la representación del conocimiento y la creación de estructuras con un significado ligado a la interacción del agente con el entorno en el que se desenvuelve (*symbol grounding*) (Haikonen 2007b).

En CERA-CRANIUM la información sensoriomotora se procesa de forma iterativa en los ETC, y ascendentemente a través de las capas de CERA, con el objetivo de crear conocimiento sobre el mundo cada vez con un significado de más alto nivel. Las lecturas de los sensores se procesan inicialmente gracias a la acción de los preprocesadores de sensor en la capa física de CERA. Estos preprocesadores crean los perceptos simples, que constituyen paquetes de información mono-modal combinados con parámetros de contextualización asociados. Estos parámetros de contextualización caracterizan el estímulo que se está procesando en términos de su posición relativa y el momento (marca de tiempo) en el que el sensor correspondiente lo ha detectado (ver Figura 25).

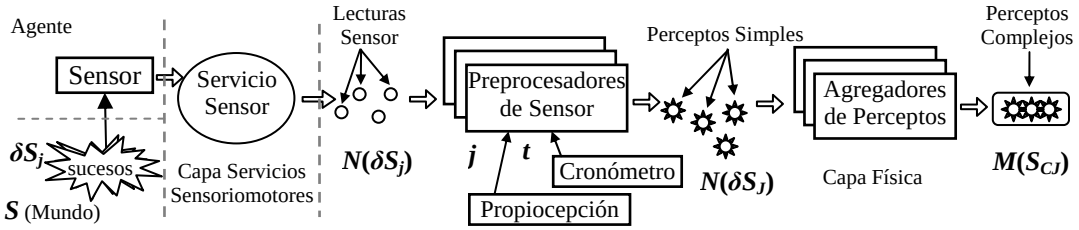


Figura 25. Generación de perceptos simples en la capa física de CERA.

La capa física de CERA contiene una serie de módulos diseñados para monitorizar variables físicas del agente. Por ejemplo, el módulo cronómetro implementa un reloj de precisión que representa la edad del agente con una resolución de un milisegundo. Adicionalmente, el módulo de propiocepción calcula la posición relativa de los sensores exteroceptivos. Los preprocesadores de sensor usan estos parámetros para calcular la localización (relativa al agente) del percepto que se está creando a partir de la información recibida del correspondiente sensor exteroceptivo. Para generar perceptos más selectivos y precisos se pueden añadir más parámetros de contextualización en CERA.

En la presente tesis se ha adoptado una notación análoga a la utilizada por Aleksander para su propuesta de los axiomas de la conciencia (Aleksander, Dunmall 2003), donde *S* es el mundo exterior accesible para los sensores del agente y δS_j representa un percepto mínimo. Un requisito necesario para el proceso de percepción es que estos perceptos mínimos tengan una correlación con un estado físico interno del agente: $N(\delta S_j)$. Siguiendo esta notación, *j* representa la posición relativa, es decir, la codificación de la localización donde se ha producido el percepto desde el punto de vista del agente (observador). Consecuentemente, *N(S)* es la representación completa del mundo que genera el agente internamente.

En la capa física de CERA los perceptos simples que producen los preprocesadores de sensor incluyen información contextual relativa denominada *índice-J* (*J* también incluye las representaciones necesarias para codificar localizaciones relativa o "vectores referentes *j*"). En otras palabras, los perceptos simples también se pueden denominar $N(\delta S_J)$, donde *J* es el conjunto de parámetros contextuales que incluyen *j* (posición relativa del estímulo) y *t* (marca de tiempo correspondiente al momento en que se ha producido el estímulo). Por lo tanto, *N(S)* es la unión de todas las representaciones $N(\delta S_J)$ producidas en CERA (ver Ecuación 1). En CERA, la representación de los parámetros contextuales relativos, *índice-J*, no se codifica usando redes de neuronas, sino representaciones explícitas en forma de vectores geométricos o variables enteras.

$$N(S) = \cup_J N(\delta S_J)$$

Ecuación 1. Representación del mundo en la capa física de CERA.

Todos los perceptos simples se producen en el ETC de la capa física de CERA, donde se hacen inmediatamente accesibles a todos los procesadores especializados de esta capa. Los procesadores agregadores de perceptos son capaces de combinar dos o más perceptos simples que estén disponibles en el espacio de trabajo compartido, creando a partir de ellos un percepto complejo. Los perceptos complejos pueden ser representaciones monomodales y multimodales dependiendo de la modalidad sensorial de los perceptos simples de los que proceden. En esencia, los perceptos complejos constituyen representaciones parciales del mundo más elaboradas que se han ensamblado e integrado como resultado de la aplicación de ciertas reglas de contextualización.

De forma similar a los perceptos simples, los recién generados perceptos complejos se publican inmediatamente en el ETC de la capa física de CERA, además también se envían al ETC de la capa de misión. Esto significa que los procesadores especializados de ambas capas pueden acceder a ellos. Determinados procesadores especializados podrían leer varios perceptos complejos para crear perceptos complejos más elaborados.

Para cada nuevo percepto complejo que se crea se calcula un *índice-J* compuesto (*índice-CJ*). Los agregadores de perceptos combinan los *índices-J* provenientes de los perceptos simples originales y construyen un *índice-CJ* integrado. Este *índice-CJ* representa los parámetros de contextualización del percepto complejo que se ha creado considerándolo como una unidad. Dado que los perceptos simples están indizados usando el *índice-J*, se pueden agrupar aplicando los criterios de contexto. Por ejemplo, un procesador especializado concreto podría seleccionar los perceptos simples que se han creado hace 10 segundos y corresponden a estímulos detectados en el lado izquierdo del agente, construyendo un percepto complejo con toda esa información. De acuerdo con esta definición, los perceptos complejos se pueden denominar $M(S_{CJ})$, es decir, un subconjunto de la representación del mundo que el agente mantiene internamente (ver Ecuación 2).

$$M(S_{CJ}) \subset N(S)$$

Ecuación 2. Subconjunto de la representación del mundo (percepto complejo).

Cuando un agregador de perceptos construye un nuevo percepto complejo se puede encontrar con problemas cuando los perceptos simples que se pretenden integrar son contradictorios. Esta situación puede ser debida, por ejemplo, al ruido inherente al funcionamiento de los sensores o a un fallo de hardware. Los perceptos simples que se han generado a partir de la información proveniente de sensores diferentes, pero relacionados gracias a la aplicación de un contexto común, pueden contener datos contradictorios. En tal caso es necesario aplicar una política para crear perceptos complejos con un significado coherente y que integren de forma efectiva la información procedente de los perceptos simples originales. Una opción es asignar niveles de confianza tanto a la información de los sensores como a los parámetros contextuales asociados. Otra opción complementaria consiste en la generación de perceptos de incongruencia (*mismatch*) que atraerán la atención de la capa núcleo hacia la situación inesperada que se ha detectado.

La capa de misión de CERA contiene un ETC en el que los perceptos complejos provenientes de la capa inferior se convierten en la entrada de los procesadores específicos de misión. Análogamente al funcionamiento de la capa física, los perceptos de la capa de misión se envían tanto a la ETC de esa capa como a la capa superior (la capa núcleo en este caso). Los perceptos de misión se podrían haber generado directamente usando un único ETC en el que se procesarían todos los tipos de perceptos existentes (simples, complejos y de misión). Sin embargo, el uso de ETC separados permite desacoplar los procesadores especializados específicos del cuerpo físico de un agente concreto de los procesadores especializados específicos de misión. Los detalles de algunos de los procesadores especializados de misión implementados se encuentran en el Apartado 7.4.2.

En CERA-CRANIUM la representación de las acciones se realiza de forma análoga a la representación descrita anteriormente para la información sensorial. Las acciones atómicas se definen como δB_I (Arrabales, Ledezma & Sanchis 2007), siendo el *índice-I* el conjunto de vectores de referencia que indica los parámetros del movimiento (como la dirección y la velocidad). Siguiendo la misma notación, $M(B_{CI})$ representa un comportamiento genérico, que se define como una secuencia de acciones atómicas. La notación *CI* se refiere a un índice integrado que representa el contexto que se espera alcanzar cuando el comportamiento se realice. Por ejemplo, si $M(B_{CI})$ se refiere a un

movimiento de cambio de sentido en un robot móvil, el *índice-CI* indicará la posición final esperada del agente una vez que el cambio de sentido se haya completado. Si el comportamiento de cambio de sentido se descompone en acciones atómicas, la secuencia de *índices-I* (correspondientes a la secuencia de pasos requeridos para realizar el cambio de sentido) representarán las posiciones y velocidades intermedias que componen la maniobra.

Aplicando de nuevo la misma notación que se ha empleado para los perceptos, $N(\delta B_I)$ se refiere a la representación que se mantiene en la capa física de CERA para las acciones atómicas del agente. De forma análoga al procesamiento ascendente que realizan los preprocesadores de sensores, los preprocesadores de acción realizan un procesamiento descendente para crear acciones simples a partir de las acciones atómicas incluyendo información contextual como el momento actual y la posición relativa de los actuadores del agente (ver Figura 26).

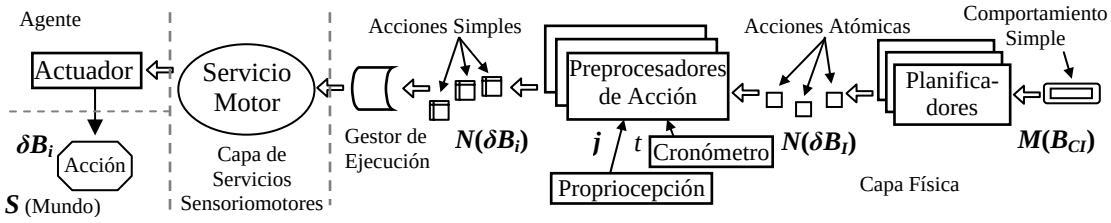


Figura 26. Generación de acciones simples en la capa física de CERA.

El gestor de ejecución de CERA controla la cola de ejecución en la que las secuencias de acciones simples que están activas esperan a ser ejecutadas. Las estructuras de datos llamadas acciones simples incluyen algunos parámetros que el gestor puede usar para aplicar determinadas políticas de gestión de la cola. Por ejemplo, las acciones derivadas de los comportamientos simples de más alta prioridad se ejecutan en primer lugar.

Los comportamientos específicos de misión se generan en la correspondiente capa de CERA y posteriormente se envían al ETC de la capa física donde se descomponen en secuencias de comportamientos simples. La activación de los comportamientos específicos de misión se basa en la aplicación de las metas establecidas en la capa de misión y los comandos enviados desde la capa núcleo. En general, el funcionamiento de la capa física y la capa de misión de CERA se regula gracias a comandos de modulación enviados desde la capa núcleo.

4.6 Mecanismos de modulación en CERA-CRANIUM

Los ETC de CRANIUM no son mecanismos pasivos de memoria a corto plazo. Su funcionamiento se regula en virtud de diversos parámetros que afectan la forma en que funciona el pandemonio. El valor de estos parámetros se establece por la acción de comandos de control que se envían a la capa física y la de misión desde la capa núcleo de CERA. Es decir, mientras que CRANIUM proporciona un mecanismo que permite combinar funciones especializadas y así generar representaciones con significado, CERA establece una estructura jerarquizada y modula los procesos de competición y

colaboración de acuerdo con el modelo de la conciencia especificado en la capa núcleo. Este mecanismo cierra el lazo de retroalimentación entre la capa núcleo y el resto de la arquitectura: la entrada de la capa núcleo (percepción) se moldea en base a su propia salida (señales de modulación de los ETC), la cual a su vez determina qué es lo que se percibe.

Todas las teorías de la conciencia hacen una distinción entre procesamiento explícito e implícito (ver Apartado 2.1). En CERA-CRANIUM todos los contenidos sensoriomotores que se procesan en los ETC son por defecto contenidos implícitos o inconscientes. La selección de un conjunto reducido de contenidos que estarán disponibles para el razonamiento explícito se produce gracias a la competición entre los diferentes procesadores especializados y perceptos. CERA-CRANIUM proporciona un mecanismo para modular estos procesos de competición por medio de la aplicación de contextos. Son los perceptos ganadores, es decir, lo seleccionados bajo el foco de atención, los que se consideran explícitos.

Como se ha adelantado en los apartados anteriores, los perceptos complejos se forman gracias a la combinación de perceptos simples. Esta combinación o integración de información se realiza básicamente gracias a la aplicación de contextos. En CERA-CRANIUM la aplicación de contextos se realiza paralelamente en dos flujos complementarios:

- **Proceso de contextualización ascendente** (o contextualización *bottom-up*). Consiste en la aplicación de contextos "nativos" (pre-configurados) en la capa física y la capa de misión, como por ejemplo los contextos espaciotemporales que permiten asociar automáticamente perceptos relacionados por su posición en el espacio y el momento en el que han tenido lugar.
- **Proceso de contextualización descendente** (o contextualización *top-down*). Consiste en la inducción de contextos activos mediante el envío de comandos de contexto enviados desde la capa núcleo a los ETC de la capa física y de la capa de misión (ver Figura 27).

La acción conjunta de estos dos procesos determina los contenidos que se crean en los ETC, es decir, establecen qué perceptos complejos y perceptos de misión se forman. Tanto el proceso de contextualización ascendente como el descendente permiten asociar perceptos que muestran afinidad o relación entre ellos. Sin embargo, el proceso descendente también afecta al nivel de activación de los perceptos creados, dando más prioridad a los perceptos más cercanos a los contextos inducidos.

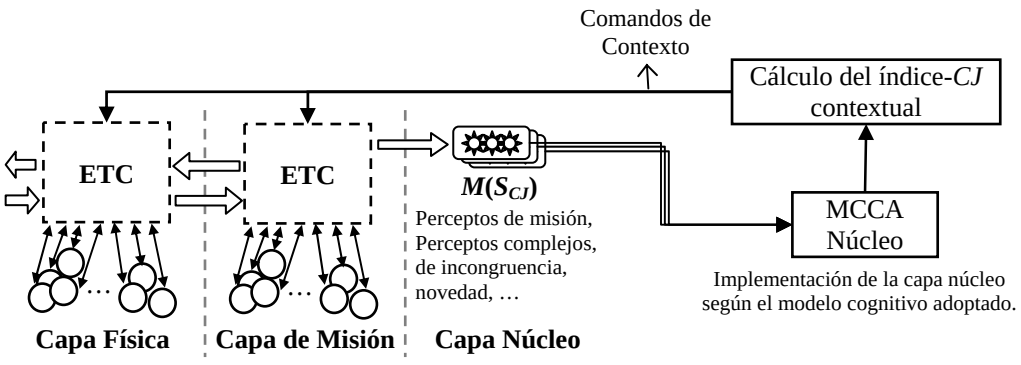


Figura 27. Modulación inducida desde la capa núcleo de CERA.

De acuerdo con el MCCA implementado la capa núcleo de CERA calcula periódicamente un *índice-CJ* contextual que representa una región de interés dentro del espacio sensoriomotor del agente. De hecho, este *índice-CJ* representa el sesgo (o modulación) que la capa núcleo induce a las capas inferiores (ver Figura 27). Los datos de entrada que se procesan en la capa núcleo con el objetivo de generar una señal de modulación adaptativa consisten en un conjunto de perceptos enviados desde "el foco" de la capa de misión. Estos perceptos pueden proporcionar información sobre contextos creados en las capas inferiores (*índices-CJ*), el nivel de activación, señales de novedad o incongruencia, nivel de cumplimiento de los objetivos, comportamientos que se están ejecutando actualmente, etc.

La tarea principal de la capa núcleo es calcular de forma adaptativa e iterativa un *índice-CJ* contextual. Este índice puede estar compuesto por diversos criterios de contextualización, que se pueden considerar como las dimensiones del espacio sensoriomotor del agente. Por ejemplo, *CJ = (j1, t1)* podría ser un vector perteneciente a un espacio sensoriomotor de dos dimensiones correspondientes a los parámetros contextuales espacio (*j*) y tiempo (*t*). A su vez, *j1* estaría compuesto por las tres dimensiones de un espacio tridimensional euclídeo representado en un sistema de referencia egocéntrico. Por lo tanto, un posible índice-*CJ* podría representarse con cuatro valores reales: *J = (j1(x1,y1,z1), t1)*. Los tres primeros, *x1*, *y1* y *z1*, representarían la localización relativa del contexto inducido y *t1* representaría el momento asociado. Aunque los criterios contextuales espaciotemporales son los más comunes y básicos, el *índice-CJ* puede estar compuesto de más dimensiones o criterios contextuales. Por ejemplo, en la implementación descrita en el Apartado 7.4.3 los *índices-CJ* incluyen además parámetros contextuales derivados del sensor de visión como el color y el movimiento.

Los ETC asignan un nivel de activación a los procesadores especializados, a los comportamientos (comportamientos simples y comportamientos de misión) y a los perceptos (perceptos simples, perceptos complejos y perceptos de misión). Los niveles de activación son variables muy dinámicas que se actualizan constantemente en los ETC. El papel de estos niveles de activación es doble: por un lado, los perceptos con un nivel de activación muy bajo dejan de procesarse, ahorrando de esta manera los limitados recursos computacionales; por otro lado, los perceptos con los niveles de activación más altos se procesan iterativamente hasta que se selecciona un ganador como contenido consciente. El funcionamiento de los procesadores especializados también depende de los niveles de activación de los mismos. Los procesadores con niveles de activación más bajos tienes menos probabilidad de ser ejecutados ya que se les da menos prioridad en el gestor de hilos del ETC.

Los niveles de activación de los procesadores especializados y de los perceptos se calculan en función de múltiples parámetros, siendo los criterios de contextualización los más importantes. Los comandos de contexto son mensajes que se envían a los ETC especificando un *índice-CJ*. Este *índice-CJ* contextual establece los criterios de contexto, como el tiempo o la localización relativa, que tienen que activarse en los espacios de trabajo. Por ejemplo, un *índice-CJ* podría hacer referencia a un segmento concreto del campo visual del agente definido en base a la amplitud angular existente entre los 8 y los 14 grados. El envío de un comando de contexto que contuviera un índice-*CJ* como este causaría el incremento del nivel de activación de aquellos perceptos cuyos *índices-CJ* indican que se han percibido cerca de esa posición específica. El nivel de activación que se asigna es inversamente proporcional a la distancia entre el *índice-CJ* del comando de contexto y el *índice-CJ* del percepto.

Cuando un ETC recibe un comando de contexto, los niveles de activación de todos los perceptos y procesadores asociados a ese espacio de trabajo se recalculan automáticamente. Se calcula la distancia entre el índice-*CJ* contextual especificado y los índices-*CJ* de los perceptos existentes, asignando consecuentemente nuevos valores de activación a los perceptos. El nivel de activación de los procesadores se asigna en base a la entrada que son capaces de procesar (el nivel de activación de la entrada potencial que podrían procesar).

Los contextos activos se establecen típicamente en la capa núcleo teniendo en cuenta las metas del agente y la retroalimentación obtenida de las capas inferiores. La activación de los perceptos también se basa en el mecanismo coincidencia-incongruencia-novedad (*match-mismatch-novelty*) propuesto por Haikonen (Haikonen 2007b). Por ejemplo, a un percepto de incongruencia se le asigna inicialmente un nivel de activación alto porque podría representar parte de una situación inesperada que puede requerir atención consciente. Una vez que esta señal de incongruencia (percepto de incongruencia) alcanza la capa núcleo, corresponde al MCCA decidir el sesgo contextual que ha de inducirse en las capas inferiores de CERA (por medio de un comando de contexto que especifique un índice-*CJ* dirigido hacia el origen del percepto inesperado).

El sesgo contextual que induce la capa núcleo determina los perceptos que se forman en los ETC. Consecuentemente, el proceso de percepción es un mecanismo claramente activo, contrariamente a lo que sería un sistema clásico de adquisición de datos que funciona de forma pasiva. Como también se asigna un nivel de activación a los comportamientos, la generación de comportamientos también se ve afectada en gran medida por los contextos activos en cada instante. En un momento dado, se generan en los ETC un número de posibles comportamientos. Sin embargo, sólo aquellos que tienen los niveles de activación más altos tienen una probabilidad significativa de ser seleccionados para su ejecución. El cálculo de los niveles de activación de los comportamientos también se basa en las distancia entre el *índice-CJ* contextual y los *índices-CI* de los comportamientos. Es decir, aquellos comportamientos que están dirigidos a la misma localización relativa que los contextos activos tendrán más probabilidades de ser seleccionados. De hecho, el *índice-CJ* contextual no sólo se usa para seleccionar comportamientos existentes, sino que también se usa para generar nuevos comportamientos dirigidos al foco de atención actual. La aplicación de estos mecanismos se ilustra en el Capítulo 7, donde se describe la implementación de una capa núcleo específica (en la selección de comportamientos también influye el modelo emocional implementado en la capa núcleo).

4.7 Modelo de Conciencia Artificial propuesto

Como se ha explicado en los apartados anteriores, la arquitectura CERA-CRANIUM está diseñada para soportar diversos modelos computacionales de la conciencia (o MCCA). Sin embargo, la inspiración tomada de las teorías cognitivas descritas en el Apartado 2.1.4, y en consecuencia el propio diseño software de la arquitectura, introducen un claro sesgo en cuanto a la definición de los MCCA que podrían implementarse de forma directa utilizando CERA-CRANIUM.

En este apartado se describe un modelo específico de Conciencia Artificial que se ha propuesto como marco tanto para la experimentación realizada en la presente tesis

como para su implementación completa en futuros trabajos de investigación (ver Apartado 8.2). Debido a la envergadura del modelo planteado, una implementación completa del mismo constituiría un proyecto de proporciones muy ambiciosas, quedando fuera del alcance de la presente tesis doctoral. Por lo tanto, la experimentación se ha realizado utilizando implementaciones parciales del modelo propuesto. Este modelo, denominado MC^3 (Modelo de la Cognición en CERA-CRANIUM), es un caso específico de MCCA que pretende cubrir un amplio espectro funcional, constituyendo así un modelo exhaustivo que define los principales mecanismos cognitivos asociados con la conciencia.

Con el objetivo de especificar el modelo mencionado se han identificado las principales funciones cognitivas asociadas con la conciencia. Este proceso de especificación está inspirado en la investigación científica de la conciencia y se basa específicamente en las teorías cognitivas descritas en el Apartado 2.1.4. Por ejemplo, Baars identifica nueve funciones básicas de la conciencia de acuerdo a la Teoría del Espacio de Trabajo Global (Baars 1988, Baars 1997): adaptación, aprendizaje, contextualización, acceso al yo, priorización, reclutamiento de procesadores inconscientes, toma de decisiones, detección de errores, auto-monitorización y optimización. Adicionalmente, en el modelo MC^3 también se han tenido en cuenta las funciones asociadas a las emociones tal y como las describe Damasio (Damasio 1999) (ver Apartado 2.1.4.3).

El modelo MC^3 se basa en la especificación de un conjunto de mecanismos funcionales interrelacionados que encajan en la arquitectura CERA-CRANIUM descrita en los apartados anteriores. En general, la aplicación de un nuevo MCCA en la arquitectura CERA-CRANIUM implica fundamentalmente el desarrollo de una nueva capa núcleo. Sin embargo, la realización de estos mecanismos no tiene que tomar necesariamente la forma de módulos independientes (situados exclusivamente en la capa núcleo), sino que la implementación de las funcionalidades puede realizarse a través de diversos módulos integrados o acoplados entre ellos a través de las diferentes capas de CERA. Asimismo, los propios mecanismos pueden interactuar entre ellos. En resumen, no tiene por qué existir una relación unívoca entre los mecanismos funcionales definidos y los módulos que finalmente puedan implementar (total o parcialmente) las correspondientes funcionalidades.

Dado que las posibles implementaciones del modelo MC^3 se basan en la arquitectura CERA-CRANIUM, normalmente la realización de cada uno de los mecanismos definidos requerirá el funcionamiento conjunto de varios módulos o subsistemas localizados en diferentes partes de la arquitectura cognitiva. Los mecanismos funcionales definidos en MC^3 son los siguientes:

- **Mecanismo de Atención**. Representa la función cognitiva de la atención, es decir, la capacidad de seleccionar contenidos específicos en el espacio sensoriomotor del agente.
- **Mecanismo de Valoración del Propio Estado**. Representa la funcionalidad que permite evaluar el funcionamiento del agente desde diferentes perspectivas: uso de recursos, consecución de las metas establecidas, etc.
- **Mecanismo de Búsqueda Global**. Permite un acceso global a cualquier contenido sensoriomotor del agente mediante la aplicación de determinados criterios de búsqueda.

- **Mecanismo de Gestión Preconsciente**. Establece qué contenidos sensoriomotores (de entre todos los que se están procesando implícitamente) pasan a formar parte del procesamiento explícito del agente.
- **Mecanismo de Contextualización**. Permite establecer contextos que definen, en base a determinados criterios, subconjuntos dentro del espacio sensoriomotor del agente.
- **Mecanismo de Predicción Sensoriomotora**. Representa la capacidad del agente de tener expectativas. Es decir, predecir los estados sensoriomotores en los que se encontrará el agente en un futuro inmediato en función de las condiciones actuales.
- **Mecanismo de Gestión de Memoria**. Permite tanto el almacenamiento como la recuperación de contenidos en los distintos tipos de memoria disponibles en el agente (memoria de trabajo, memoria episódica, memoria procedimental, etc.).
- **Mecanismos de Aprendizaje**. Permite la adquisición de nuevo conocimiento procedimental y declarativo.
- **Mecanismo de Auto-Coordinación**. Establece las políticas generales de funcionamiento y auto-regulación del agente.
- **Mecanismo de Comunicación de los Estados Mentales**. Permite que el agente pueda especificar a un observador externo (y a sí mismo) parte de su estado mental de forma precisa. La comunicación puede ser verbal o realizarse por cualquier otro medio o combinación de medios.

Los mecanismos descritos anteriormente han de integrarse de forma efectiva en la arquitectura cognitiva CERA-CRANIUM para completar un sistema de control autónomo de un agente. Sin embargo, es posible realizar implementaciones parciales de tal sistema de control en las que no estén presentes algunos de los mecanismos especificados. De hecho, todas las implementaciones utilizadas en la experimentación (ver Capítulo 7) son realizaciones parciales del modelo MC^3 en las que, por ejemplo, no hay mecanismos complejos de aprendizaje ni sistema de memoria a largo plazo.

De la descripción de la arquitectura CERA-CRANIUM ofrecida en los apartados anteriores se puede deducir que algunos de los mecanismos definidos en el modelo MC^3 ya están en gran parte presentes o sólo necesitan pequeñas ampliaciones. En general, CERA-CRANIUM ya ofrece un soporte básico para varios de estos mecanismos. Consecuentemente, para la realización de los mismos sólo resta desarrollar la parte de control necesaria en la capa núcleo de CERA (ver Tabla 3).

En el caso del mecanismo de atención, los procesos de modulación ya existentes en CERA-CRANIUM (ver Apartado 4.6) proporcionan la base necesaria para desarrollar la funcionalidad asociada a la atención. Sin embargo, corresponde a la capa núcleo la tarea de emitir los comandos de contexto adecuados para modular el funcionamiento de los ETC de las capas inferiores de forma adaptativa. La completa implementación del mecanismo de atención implica que el agente sea capaz de prestar atención a un objeto o suceso particular, modulando consecuentemente los procesos de percepción, comportamiento y aprendizaje.

Análogamente, la implementación del mecanismo de valoración del propio estado se basará en la infraestructura presente en las capas inferiores de CERA, aunque la

valoración del estado global del agente ha de realizarse en la capa núcleo. En contraposición, otros mecanismos como la contextualización y la búsqueda global ya se encuentran disponibles completamente en CERA-CRANIUM sin necesidad de desarrollar ampliaciones en la capa núcleo. Precisamente, la capacidad de establecer criterios de contextualización y poder acceder globalmente a todos los contenidos disponibles en la memoria de trabajo son las características principales de los ETC.

El mecanismo de gestión preconsciente aborda el problema de la dualidad entre el procesamiento implícito y el procesamiento explícito. Según el modelo adoptado, el conocimiento explícito se gestiona en un único hilo secuencial, mientras que el conocimiento implícito se reparte entre multitud de procesadores paralelos inconscientes. El procesamiento implícito, que en CERA-CRANIUM se produce en los ETC, se caracteriza por ser inaccesible para el razonamiento consciente. Por el contrario, el procesamiento explícito, que en CERA-CRANIUM tiene lugar esencialmente en la capa núcleo, opera sobre contenidos de alto nivel (que aparecen como resultado de procesamiento inconsciente que va construyendo estos contenidos iterativamente gracias al funcionamiento de los ETC). Por lo tanto, el mecanismo de gestión preconsciente, es decir, el interfaz entre los procesos implícitos y los procesos explícitos, está ya presente en CERA-CRANIUM, aunque requiere una implementación completa de la capa núcleo para poder funcionar de forma efectiva.

El mecanismo de gestión preconsciente también se encarga de la interrelación entre los flujos cognitivos ascendente y descendente. El flujo cognitivo ascendente consiste en la creación iterativa de perceptos cada vez más complejos, que empiezan siendo muchos, pequeños, poco significativos e implícitos (perceptos simples), para terminar siendo pocos, complejos, integrados, muy significativos y explícitos (perceptos de misión seleccionados por el mecanismo de atención). El flujo cognitivo descendente consisten en la generación de una secuencia coordinada de acciones motoras (secuencia de acciones simples) a partir de representaciones abstractas de comportamientos (comportamientos complejos).

Los mecanismos de aprendizaje pueden tener lugar tanto a nivel de procesamiento implícito como a nivel de procesamiento explícito. En otras palabras, los mecanismos de aprendizaje pueden operar concurrentemente en las diferentes capas de CERA-CRANIUM. La interacción entre los diversos procesos de aprendizaje que ocurren simultáneamente a diferentes niveles se produce gracias a otros mecanismos existentes en el modelo. Por ejemplo, las funcionalidades de contextualización y búsqueda global proporcionan una forma de recuperar recursos necesarios desde el dominio del inconsciente. A su vez, el mecanismo de atención permite seleccionar los contenidos más relevantes de entre todos los recuperados, ocurriendo todo esto de forma transparente desde el punto de vista de la capa núcleo. Al igual que ocurre en la mente humana, desde el punto de vista del sujeto, los contenidos conscientes simplemente aparecen, la introspección no permite observar los mecanismos que dan lugar a la creación de estos contenidos.

En las implementaciones basadas en CERA-CRANIUM, los mecanismos de aprendizaje pueden operar simultáneamente en el dominio del inconsciente (esencialmente en el funcionamiento de los ETC) y en el dominio de la conciencia (esencialmente en el funcionamiento de la capa núcleo). También es posible tener mecanismos de aprendizaje sólo en una capa. Asimismo, se pueden combinar diferentes estrategias de aprendizaje dentro de una misma capa de la arquitectura cognitiva. Incluso se podrían aplicar técnicas de aprendizaje para optimizar el funcionamiento de

los propios mecanismos de atención, predicción, etc. Sin embargo, una implementación básica de CERA-CRANIUN puede funcionar sin ningún mecanismo de aprendizaje implementado (utilizando reglas básicas pre-programadas).

La predicción sensorial se basa en un proceso de monitorización que permite la predicción (inconsciente) de la información que captan los sensores. Cuando lo percibido difiere significativamente de lo predicho, la información correspondiente debe hacerse consciente con el objetivo de lidiar de forma efectiva con la situación imprevista. En CERA-CRANIUM este mecanismo está presente en forma de procesadores especializados específicos (predictores sensoriales) que monitorizan determinados tipos de perceptos (ver Apartado 4.3).

En cuanto al mecanismo de gestión de memoria, la arquitectura básica CERA-CRANIUM descrita en el presente trabajo sólo incluye la memoria de trabajo (implementada mediante los ETC). Por lo tanto, las implementaciones derivadas de la versión actual de CERA-CRANIUM no cuentan con sistemas de memoria a largo plazo. La memoria modal, es decir, la que sólo almacena un determinado tipo de contenidos, está representada por los perceptos simples. La memoria multimodal, que puede representar diversos tipos de contenidos conjuntamente, se realiza por medio de los perceptos complejos y los perceptos de misión.

El mecanismo de auto-coordinación se sitúa típicamente en la capa núcleo de CERA-CRANIUM, pues es el encargado de gestionar las políticas de alto nivel requeridas para alcanzar las metas del agente. Las principales funcionalidades asociadas con este mecanismo son la toma de decisiones (de forma explícita), el control de alto nivel y la planificación y priorización de tareas. La implementación básica de CERA-CRANIUM descrita en los apartados anteriores no incluye un mecanismo específico de auto-coordinación. Dado que este mecanismo es exclusivo de la capa núcleo, su implementación puede variar significativamente de un MCCA a otro. En el Capítulo 7 se describen las implementaciones concretas desarrolladas para la experimentación con CERA-CRANIUM.

El mecanismo de comunicación de estados mentales es útil desde dos puntos de vista: por un lado, permite la experimentación basada en los principios de la Fenomenología Sintética (ver Apartado 2.2.6); por otro lado, permitiría también que el agente desarrolle interacciones sociales complejas con otros agentes. Para implementar una versión básica de este mecanismo se ha desarrollado un módulo denominado CERA Viewer, que permite inspeccionar las representaciones que llegan a la capa núcleo.

La Tabla 4 resume como se han diseñado los mecanismos definidos en MC^3 para poder ser implementados en la arquitectura CERA-CRANIUM. En general, el diseño de los mecanismos tiene en cuenta la interacción existente entre las diferentes capas de la arquitectura cognitiva, consecuentemente no es posible localizar de forma unívoca una determinada funcionalidad en un módulo específico de CERA-CRANIUM. Tal y como se sintetiza en la Tabla 4 la implementación de los mecanismos definidos en MC^3 está distribuida en los diversos módulos de la arquitectura. Este diseño pone de manifiesto la integración y el acoplamiento de los mecanismos especificados en el modelo.

El diseño del modelo MC^3 para su implementación en CERA-CRANIUM busca la sinergia que potencialmente puede emerger a partir de la interacción de las diversas funciones asociadas con la conciencia. La experimentación realizada (ver Capítulo 7) trata de poner de manifiesto cómo la interacción entre estas funciones puede dar lugar a algunos comportamientos similares a los atribuibles a los seres conscientes.

Tabla 3. Implementación de los mecanismos MC^3 en CERA-CRANIUM.

Mecanismo MC^3	*Capa Física*	*Capa de Misión*	*Capa Núcleo*
Atención	Modulación del ETC.	Modulación del ETC.	Envío de comandos de contexto.
Valoración del Propio Estado	Valoración de la consecución de metas físicas.	Valoración de la consecución de metas de misión.	Valoración global y activación de metas globales.
Búsqueda Global	Reclutamiento de procesadores en el ETC. Acceso global a los contenidos del ETC.	Reclutamiento de procesadores en el ETC. Acceso global a los contenidos del ETC.	Obtención de contenidos de las capas inferiores.
Gestión Preconsciente	Selección de contenidos ganadores que pasan a la capa superior.	Selección de contenidos ganadores que pasan a la capa superior.	Sólo los contenidos seleccionados en la capa inferior se procesan explícitamente.
Contextualización	Asociación, agregación y selección de contenidos en base a los criterios de contexto activos.	Asociación, agregación y selección de contenidos en base a los criterios de contexto activos.	Elaboración de criterios de contexto que modulan el funcionamiento de las capas inferiores.
Predicción Sensoriomotora	Procesadores especializados calculan expectativas y se crean perceptos de incongruencia o novedad.	Procesadores especializados calculan expectativas y se crean perceptos de incongruencia o novedad.	Se usan las notificaciones de incongruencia y novedad en la toma de decisiones.
Gestión de Memoria	Los contenidos del ETC constituyen la memoria de trabajo.	Los contenidos del ETC constituyen la memoria de trabajo.	La aplicación de contextos determina los contenidos cuya activación disminuye y finalmente desaparecen de los ETC.
Aprendizaje	Diversos mecanismos de aprendizaje pueden operar a este nivel.	Diversos mecanismos de aprendizaje pueden operar a este nivel.	Diversos mecanismos de aprendizaje pueden operar a este nivel.
Auto-Coordinación	La aplicación de metas de nivel físico junto con los comandos de la capa núcleo regulan el funcionamiento del ETC.	La aplicación de metas de misión junto con los comandos de la capa núcleo regulan el funcionamiento del ETC.	Definición de metas globales y generación de señales de modulación de las capas inferiores.
Comunicación de los Estados Mentales	El procesamiento es implícito y los contenidos no son interpretables a alto nivel.	El procesamiento es implícito y los contenidos no son interpretables a alto nivel.	Especificación de los contenidos explícitos que se usan para la toma de decisiones.

Es importante remarcar que no se ha definido ningún mecanismo o módulo específico etiquetado como "conciencia", ya que una de las hipótesis de trabajo en la concepción del modelo MC^3 es que la Conciencia A (conciencia de acceso) emerge como resultado de la interacción de los mecanismos descritos anteriormente. Asimismo,

se mantiene que estos mecanismos son la base para el potencial desarrollo de la Conciencia P (conciencia fenomenológica).

La capa núcleo de CERA proporciona un marco para la comunicación e integración a alto nivel de todos los componentes del modelo MC^3. Adicionalmente, la capa núcleo alberga la función volitiva explícita del agente. Es decir, es en la capa núcleo donde se decide de forma explícita el próximo comportamiento que ha de ejecutarse. Sin embargo, el comportamiento final del agente se produce como resultado de las interacciones complejas existentes entre las diferentes capas de la arquitectura cognitiva. Aunque se establezca una consigna desde la capa superior, que se basa en las metas globales establecidas a ese nivel, el flujo descendente de CERA-CRANIUM impone un control jerárquico que contempla también las metas de nivel físico y las metas de misión.

Las metas de la capa física típicamente se manifiestan en forma de reflejos que se desencadenan al detectarse ciertos estímulos a ese nivel. Como en el ejemplo de la Figura 21, en el que un percepto es procesado por un procesador reactivo y se desencadena una respuesta refleja. Las metas de la capa física están codificadas en los propios procesadores reactivos que procesan los estímulos de nivel físico. Es decir, para cambiar las metas de la capa física es necesario modificar los procesadores asociados al ETC de la capa física y/o modificar los niveles de activación de estos procesadores. Normalmente, las metas de nivel físico tienen la prioridad más alta, puesto que se refieren a estímulos obtenidos directamente del entorno del agente (a través de mínimos procesos interpretativos) y suelen requerir una respuesta más inmediata como en el ejemplo de la Figura 21. Sin embargo, como se muestra en los resultados experimentales (Capítulo 7), la modulación inducida desde la capa núcleo puede hacer que metas de nivel superior alcancen niveles de activación más altos, provocando que los propios “reflejos” del agente tengan en ocasiones menor prioridad que comportamientos específicos de la capa de misión (provocando así un veto a los reflejos de nivel físico).

En la capa de misión tiene lugar un mecanismo análogo, pero esta vez los estímulos percibidos son perceptos de misión. Por lo tanto, las metas codificadas en los procesadores especializados de la capa de misión están relacionadas con sucesos significativos a este nivel. Es decir, las metas de misión representan soluciones parciales o totales del dominio de problema específico para el que se ha diseñado el agente. Por ejemplo, si el agente tiene como misión la creación de mapas de entornos desconocidos, los perceptos complejos correspondientes a espacios no explorados pueden desencadenar reacciones o tendencias dirigidas a explorar dichos espacios.

Las metas de la capa núcleo (o meta-objetivos) representan metas cognitivas de propósito general. Típicamente, estas metas tienen la prioridad más baja dado que no actúan directamente sobre los mecanismos sensoriomotores del agente. Sin embargo, la aplicación de las metas de la capa núcleo establece la señal de modulación que se induce en las capas inferiores. Consecuentemente, estos meta-objetivos definen a largo plazo el comportamiento del agente. El nivel de activación de las diferentes metas de la capa de núcleo se puede calcular en base al mecanismo de evaluación del propio estado.

El cálculo de los niveles de activación de las metas es esencial ya que es normal que durante el funcionamiento del agente las metas de diferentes niveles entren en conflicto. Es decir, induzcan comportamientos enfrentados. Por ejemplo, en un agente CERA-CRANIUM diseñado para la exploración y mapeo de espacios desconocidos,

podrían coexistir las siguientes metas (tal implementación se describe en el Apartado 7.4.2):

- Nivel Físico: “Evitar golpes contra las paredes” (M1).
- Nivel de Misión: “Explorar zonas desconocidas” (M2).
- Nivel de Núcleo: “Maximizar el resultado de la valoración del propio estado” (M3).

La meta M1 induce un comportamiento que mantiene el agente a una distancia prudencial de las paredes para evitar contacto durante maniobras de giro, etc. La meta M2 induce un comportamiento que dirige al agente hacia zonas no exploradas. La meta M3 constituye un meta-objetivo que permite resolver conflictos entre las metas de los niveles inferiores. Dado que la valoración del propio estado se basa en la consecución de metas, el meta-objetivo M3 está encaminado a variar los niveles de activación de las metas de niveles inferiores de forma que se mantenga el estado global más positivo posible. Dada una situación en la que explorar una nueva zona desconocida implique pasar muy cerca de una pared o incluso rozarla, los objetivos M1 y M2 entrarán en conflicto. Sus niveles de activación actuales determinarán la acción final del agente. A su vez, estos niveles dependerán de la aplicación de la meta M3, que provocará la creación de comandos de contexto dirigidos bien hacia la zona inexplorada o bien hacia otro lugar. Esta decisión dependerá del estado actual del agente. Es decir, dependiendo del desempeño obtenido hasta el momento puede merecer la pena o no sacrificar una meta para intentar obtener beneficio de otra y así aumentar la valoración global del propio estado (los experimentos descritos en el Apartado 7.4.2 incluyen una explicación más detallada de este tipo de mecanismos).

En general, acerca de las metas en CERA-CRANIUM es importante destacar que:

- Las metas no están definidas explícitamente en la arquitectura, sino que son el resultado de la interacción entre diversos procesadores especializados. La identificación de las metas de cada nivel es un proceso interpretativo ajeno al propio funcionamiento de la arquitectura cognitiva.
- Las metas descritas anteriormente operan principalmente al nivel de las capas en las que están definidas. Es decir, la interacción entre las metas de distinto nivel no se realiza directamente, sino que depende de los niveles de activación de los procesadores especializados involucrados en cada caso.
- Las metas globales de la capa núcleo (o meta-objetivos) deben ser siempre independientes del dominio. Es decir, todas las metas identificadas en la capa núcleo se referirán al funcionamiento general de la arquitectura. Todas las representaciones específicas de un dominio deben definirse inicialmente en la capa de misión (aunque puedan ser interpretadas posteriormente desde la capa núcleo).

4.8 Convirtiendo datos sensoriales en qualia

El problema de la Fenomenología Sintética (ver Apartado 2.2.6) se puede abordar usando como marco de experimentación la arquitectura cognitiva CERA-CRANIUM. En la presente tesis se ha adoptado un enfoque práctico para la modelización de los

qualia en el dominio de los agentes artificiales. Concretamente se usa la expresión qualia artificiales para hacer referencia a los perceptos que un agente percibe de forma explícita (ver Capítulo 6). Tal y como apunta Haikonen (2009), los humanos exploramos el mundo a través de perceptos explícitos que aparecen como qualia (experiencias subjetivas). Consecuentemente, no es posible tener percepción consciente sin qualia.

En términos de la arquitectura CERA-CRANIUM, las capas inferiores proporcionan los mecanismos necesarios para la adquisición de datos sensoriales, su procesamiento, composición y selección. Todos estos procesos tienen lugar de forma encubierta, es decir, no son accesibles al razonamiento explícito. Sólo una selección específica de perceptos, que tienen un alto nivel de elaboración y significado, están disponibles de forma abierta para su uso en el proceso de razonamiento explícito. Estos perceptos explícitos, aunque están adaptados y correlacionados con el mundo real gracias a los mecanismos subyacentes de CERA-CRANIUM, no representan cualidades reales del mundo exterior, sino que son la impresión que ha creado el sistema de percepción multicapa de la arquitectura cognitiva. En resumen, la entrada de la capa núcleo no consiste en estímulos externos, sino que consiste en una representación de las reacciones de las capas inferiores a dichos estímulos. En este nivel, el papel funcional de los perceptos explícitos y sus *índices-CJ* asociados consiste en proporcionar la información necesaria para crear la ilusión de que ellos representan directamente las cualidades del mundo exterior. Sin embargo, como es sabido, las propiedades de los qualia no coinciden necesariamente con las propiedades físicas del mundo real.

Como se ha explicado en los apartados anteriores, la arquitectura CERA-CRANIUM está diseñada para integrar los sistemas sensoriales exteroceptivos y propioceptivos de tal forma que se puedan calcular *índices-CJ* para cada percepto. Gracias a las propiedades de localización espacial de los vectores de referencia *j* los perceptos se pueden procesar como si estuvieran situados en el mundo exterior. Esta capacidad de procesar los perceptos como si estuvieran situados en el mundo exterior, en lugar de inspeccionar directamente los datos sensoriales, podría ser la base para la creación de qualia artificiales en una máquina. En cualquier caso, los mecanismos presentes en CERA-CRANIUM son útiles para realizar modelos o simulaciones de los contenidos de la experiencia consciente.

Es preciso aclarar que la afirmación anterior no implica que una máquina controlada por CERA-CRANIUM posea estados fenomenológicos, simplemente se afirma que es posible modelar y especificar los contenidos que formarían la experiencia consciente (qualia) en el caso de que la máquina fuera consciente. Es decir, CERA-CRANIUM constituye una plataforma válida para la experimentación en Fenomenología Sintética.

5 *ConsScale*: Una Escala para Medir la Conciencia Artificial

5.1 Introducción

La medición precisa del nivel de conciencia de una criatura sigue siendo un reto científico sin solución. Como se ha puesto de manifiesto anteriormente (ver Apartado 2.3), no existen en la actualidad medidas aceptadas y consensuadas por la comunidad científica para realizar una evaluación objetiva del nivel de conciencia de una máquina. Dado que para el desarrollo de esta tesis la posibilidad de contar con mecanismos prácticos de evaluación es crucial, se ha desarrollado una escala denominada *ConsScale* para medir el desarrollo de la Conciencia Artificial.

Tal y como indican Seth et al. (2008), el uso de este tipo de herramientas de medida no sólo es útil para evaluar el progreso de la investigación, sino que también puede ayudar a discernir cuáles son las líneas de investigación más prometedoras. Aunque se han propuesto diversos enfoques, la definición de una métrica precisa para evaluar el nivel de conciencia en sistemas artificiales es un problema sin resolver. Una de las principales dificultades radica en la propia caracterización del término conciencia, que como se ha visto, puede hacer alusión a múltiples perspectivas diferentes. Por ejemplo, desde el punto de vista de la fenomenología, conciencia P (Carruthers 2000), la conciencia se podría medir en función de la intensidad de las experiencias conscientes. Sin embargo, desde el punto de vista de la conciencia funcional, conciencia A (Baars 2002), la conciencia se podría medir en base a los contenidos de la mente que están disponibles para el procesamiento explícito. Estas apreciaciones ponen de manifiesto el hecho de que diferentes teorías tratan de explicar la conciencia desde diferentes puntos de vista, que en ocasiones pueden ser contradictorios (Atkinson, Thomas & Cleeremans 2000). Esta situación implica la definición de medidas que son sólo válidas en el contexto de la teoría que defienden. Aunque las teorías actuales sobre la conciencia ofrecen una disparidad de explicaciones para la conciencia P, existen bastantes denominadores comunes en cuanto a otros aspectos de la conciencia. En esta tesis se plantea la posibilidad de aprovechar esta particularidad con el objetivo de definir una medida, que aunque no sea definitiva, se caracterice por representar el consenso existente acerca de lo que se espera que sea la primera generación de implementaciones de Conciencia Artificial.

Puesto que en la actualidad no existe lo que podría llamarse una “gran teoría unificada de la conciencia”, no cabe esperar que se pueda definir una medida de equiparable completitud. Sin embargo, es posible adoptar una visión que contribuya a la definición y mejora reiterativa de medidas que vayan integrando incrementalmente el

nuevo conocimiento que se obtiene acerca de la conciencia. Es más, se espera que este proceso también ayude a evaluar la validez de las hipótesis consideradas en el proceso mismo de medición, de forma que se proporcione una valiosa retroalimentación acerca de la verdadera naturaleza de la conciencia. Por ejemplo, si de acuerdo a la medida considerada se demuestra la existencia de casos en los que se asigna un nivel alto de conciencia a implementaciones que en realidad no muestran un comportamiento consistente con ese nivel, las hipótesis subyacentes han de ser revisadas.

En esencia, la escala propuesta está orientada a la resolución de los llamados "problemas fáciles" de la conciencia (Chalmers 1995), soslayando temporalmente la aplicación de medidas especulativas centradas exclusivamente en el "problema duro" (ver Apartado 1.1.6). Al abordar en primer lugar los llamados problemas fáciles, se conseguirá estar en una mejor posición para enfrentarse más tarde al problema duro de la conciencia. Incluso es posible, que la concepción actual que se tiene del problema duro cambie drásticamente cuando se den por resueltos los problemas fáciles (Dennett 1996).

Aunque en los humanos se pueden dar estados fenomenológicos incluso en ausencia de cualquier comportamiento externo asociado (por ejemplo, cuando se "sueña estando despierto"), el desarrollo de la conciencia tiene su origen en una interacción directa del cuerpo con el entorno (Humphrey 1999). Los estados fenomenológicos que no tienen comportamientos asociados no tienen sentido a menos que el sujeto tenga ciertas capacidades cognitivas adquiridas anteriormente. Consecuentemente, asumiendo que los mismos principios del desarrollo cognitivo son de aplicación también en el contexto de la Conciencia Artificial, no tiene sentido diseñar máquinas capaces de producir estados fenomenológicos si ni siquiera son capaces en primer lugar de solventar los problemas fáciles. Por lo tanto, el esfuerzo en el desarrollo de una medida útil en el ámbito de la Conciencia Artificial ha de centrarse en la evaluación de las capacidades cognitivas asociadas a la conciencia.

Se espera que el desarrollo de implementaciones de Conciencia Artificial capaces de lidiar de forma efectiva con los llamados problemas fáciles contribuya al conocimiento sobre la generación de qualia artificiales (ver Capítulo 6). De esta forma se podrá avanzar en el estudio de la existencia de estados fenomenológicos en máquinas. Este conocimiento podría usarse entonces para definir nuevas medidas integradas que consideren la generación de estados fenomenológicos en sistemas artificiales.

A continuación se detalla el alcance y la utilidad esperada de la escala (Apartado 5.2), se describen los niveles conceptuales que forman la escala (Apartado 5.3), luego se explica el cálculo del índice cuantitativo asociado (Apartado 5.4) y también se ilustra el uso de la representación gráfica de los índices de *ConsScale* (Apartado 5.5). Finalmente, se especifican los métodos de aplicación de la escala (Apartado 5.7).

5.2 Motivación y alcance de la escala ConsScale

ConsScale se ha definido para servir como una escala de referencia en el campo de la Conciencia Artificial. La escala está diseñada para que pueda ser aplicada a cualquier implementación, independientemente de la tecnología empleada y del dominio de aplicación. Con el objetivo de elaborar una escala práctica, que sea útil para evaluar el nivel de conciencia de un sistema artificial, en la presente tesis se ha adoptado un punto

de vista funcionalista, basado principalmente en las capacidades cognitivas de un sujeto. Se ha considerado la conciencia como una "gran función", resultado de la integración de diversas capacidades cognitivas. Por supuesto, este enfoque constituye una visión parcial de la conciencia, pues, como se ha mencionado, la escala propuesta no trata directamente los aspectos fenomenológicos.

Se ha planteado una escala práctica, enfocada a la medición de la integración de diversas capacidades cognitivas asociadas con la conciencia. Se espera que el análisis profundo de las implementaciones de Conciencia Artificial, utilizando tanto la escala propuesta como otras medidas de conciencia (ver Apartado 2.3.1) muestre si existe una correlación entre los niveles de *ConsScale* y los niveles de conciencia P predichos por otras medidas (ver Apartado 2.3.2).

Las definiciones intuitivas de conciencia a menudo involucran conceptos como la percepción, las emociones, la atención, el auto-reconocimiento, la teoría de la mente, la voluntad, etc. Debido a que este tipo de definiciones de la conciencia implican la combinación de múltiples conceptos y propiedades, es difícil discernir qué se entiende por un ser consciente y cómo se podría evaluar la "cantidad" de conciencia existente en una máquina o un programa informático.

Examinando en el dominio de la biología los ejemplos de seres conscientes más evolucionados, como los grandes simios y los humanos, se observa una gran complejidad en las interacciones existentes entre diversas funciones cognitivas. La falta de comprensión detallada de los mecanismos neuronales asociados a estos procesos, hace que la tarea de ingeniería inversa sea virtualmente inalcanzable y por lo tanto no sea fácil establecer métricas precisas extrapolables al ámbito de la Conciencia Artificial. En este contexto, *ConsScale* se propone como un intento de tratar de forma efectiva el problema de la evaluación de modelos de la conciencia a nivel cognitivo. Concretamente, *ConsScale* define un camino concreto para el desarrollo de la conciencia, desde la "conciencia cero" hasta un nivel de conciencia equiparable al existente en los humanos, o incluso superior.

En este camino se definen fases de desarrollo clave para el proceso progresivo de construcción de máquinas conscientes. Por lo tanto, la escala también se puede considerar como una hoja de ruta, en la que cada nivel de conciencia caracteriza tipos de máquinas conscientes cada vez más avanzadas. Además de la definición de una serie de niveles de conciencia, la escala propuesta permite el cálculo de una medida cuantitativa de conciencia. La caracterización de un sistema usando *ConsScale* se completa con la confección de un gráfico radial que representa el grado de satisfacción de cada uno de los niveles de la escala.

Es importante destacar que la escala propuesta no se ha planteado como un método alternativo a los métodos de medida actuales. En realidad viene a ocupar el hueco existente en cuanto a la medición de las capacidades cognitivas asociadas con la conciencia. Su enfoque eminentemente funcional hace que la escala sea adecuada para su aplicación en el dominio de los agentes artificiales. Sin embargo, para obtener una caracterización total de una implementación de conciencia artificial, teniendo en cuenta también los aspectos fenomenológicos, *ConsScale* puede combinarse o ampliarse con otras técnicas basadas en los principios de la fenomenología sintética.

Tratar con un concepto tan complejo como es la conciencia requiere tener en consideración diversas dimensiones interrelacionadas. Aunque se considere la conciencia desde un punto de vista puramente funcional, la complejidad de los múltiples procesos cognitivos y la dinámica asociada a sus interacciones hace extremadamente

difícil el análisis, modelado y diseño de sistemas basados en Conciencia Artificial. Si bien es cierto que la cuestión sobre la posibilidad de crear máquinas con una conciencia equiparable a la humana sigue abierta (ver Apartado 2.2), muchas capacidades cognitivas normalmente asociadas a la conciencia y presentes en los humanos son la inspiración de muchos sistemas artificiales.

Las líneas de investigación actuales en el área de sistemas cognitivos concluyen que múltiples habilidades cognitivas han de combinarse en un agente con el objetivo de lograr una mente artificial equiparable a la humana. De hecho, la hipótesis que se mantiene en la presente tesis es que la conciencia, al menos desde el punto de vista funcional, podría emerger como resultado de la sinergia producida en la interacción de múltiples procesos cognitivos integrados. En este sentido, se pueden plantear diversas cuestiones de diseño: ¿cuál es la combinación correcta de habilidades cognitivas? ¿Qué capacidades cognitivas son esenciales para la conciencia? ¿Qué otras habilidades no son imprescindibles? ¿Cuáles se deberían implementar en primer lugar?

Con el objetivo de aclarar estos aspectos se propone la escala *ConsScale* como una hoja de ruta de inspiración biológica para el desarrollo de máquinas conscientes. Observando como la conciencia ha podido evolucionar en los organismos biológicos se ha intentado identificar una serie de fases de desarrollo, o niveles de conciencia. La caracterización de cada nivel de conciencia tiene dos componentes elementales: uno arquitectural y otro basado en el comportamiento. La arquitectura se refiere a los componentes básicos del sistema y el comportamiento se caracteriza como el resultado de la interacción de las habilidades cognitivas que es posible observar exteriormente.

Aplicando el marco propuesto se puede evaluar una implementación de Conciencia Artificial y asignarle un nivel de conciencia determinado. La escala no está orientada a ningún tipo particular de agentes, estando diseñada tanto para su aplicación en agentes físicamente situados como para agentes software (como por ejemplo, un agente web). En el caso de estos últimos, actividades como el envío de un mensaje se consideran al mismo nivel que una acción física en términos de análisis del comportamiento.

Mediante el análisis de la arquitectura del agente y la implementación de habilidades cognitivas, *ConsScale* define una lista ordenada de niveles de conciencia. La escala no sólo proporciona una medida cualitativa de la conciencia, si no que también proporciona un mecanismo para calcular una medida cuantitativa precisa. También se contempla la generación de una representación gráfica para una mejor caracterización de las implementaciones analizadas. El proceso genérico seguido para evaluar una implementación y obtener el nivel de desarrollo cognitivo de la misma se muestra en la Figura 28.

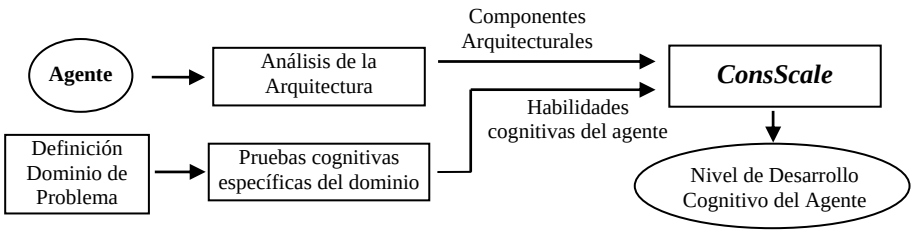


Figura 28. Obtención del nivel de desarrollo cognitivo de un agente usando *ConsScale*.

Los niveles de referencia establecidos en la escala *ConsScale* se pueden usar como una guía o una hoja de ruta para la construcción de máquinas conscientes. Aunque la escala está inspirada en la filogenia observada en el reino animal y en la ontogenia de la especie humana, los niveles definidos son discutibles, siendo posibles otras distribuciones de las capacidades cognitivas. Se espera que la retroalimentación obtenida de la comunidad científica proporcione información útil para la revisión de la escala en aquellos puntos donde se considere necesario.

Aunque no se conoce la forma precisa en la que el cerebro humano produce la conciencia, hay características específicas que se pueden analizar e incluso ordenar de acuerdo a una perspectiva evolutiva, desde las funciones más antiguas filogenéticamente hasta características más modernas observadas sólo en humanos. Aunque este enfoque filogenético aplicado al desarrollo de la conciencia no ha sido demostrado rigurosamente, es innegable que existe una jerarquía de capacidades cognitivas asociada a la evolución. Considerando la pregunta propuesta por Dennett (1997b), ¿qué clases de mentes existen?", la escala propuesta puede contribuir, si no con una respuesta, al menos con una redefinición y una contextualización de la pregunta en el ámbito de la Conciencia Artificial.

5.3 Niveles de Conciencia en ConsScale

ConsScale permite el estudio comparativo, proporcionando una medida cualitativa representada por un nivel de desarrollo cognitivo asociado a la conciencia. Los niveles se definen como una progresión incremental, es decir, cada nivel incluye a todos los niveles inferiores. Se dice que un agente cumple con un determinado nivel *si y solo si* reúne todos los requisitos de ese nivel y además todos los requisitos de todos los niveles inferiores.

Cada nivel de conciencia definido en la taxonomía propuesta se ilustra con una analogía biológica, caracterizada por un nivel aproximadamente equivalente de conciencia. La filogenia animal y la ontogenia humana son la inspiración evolutiva básica de esta escala. Los niveles de conciencia establecidos en *ConsScale* comprenden desde los agentes más sencillos hasta aquellos agentes que presumiblemente son capaces de desarrollar comportamientos equiparables a los humanos (o incluso más avanzados). Uno de los problemas de este enfoque de inspiración evolutiva es que cualquier tipo posible de conciencia que sea radicalmente distinto del ejemplo humano no sería correctamente considerado en la escala. Sin embargo, como la mayoría de las implementaciones en el campo de la conciencia artificial tratan de imitar los mecanismos del cerebro humano, la escala sería válida al menos para este tipo de implementaciones.

A continuación se describen los niveles de conciencia de *ConsScale*. Cada nivel se define en base a unos requisitos arquitecturales y de comportamiento. Por lo tanto, la descripción de los niveles consiste en la especificación de los componentes arquitecturales y las habilidades cognitivas características de cada nivel. Los componentes arquitecturales se refieren a componentes abstractos que pueden ser identificados en cualquier implementación (ver Apartado 5.3.1). Las habilidades cognitivas se refieren a las capacidades de la implementación que operan a un nivel superior al puramente sensoriomotor (ver Apartado 5.3.2).

5.3.1 Definición de Arquitectura Abstracta

Con el objetivo de caracterizar la estructura y diseño interno de un sistema artificial, es necesario formalizar la especificación de sus componentes básicos y la relación existente entre ellos. Un agente situado interacciona con el mundo que lo rodea recogiendo información tanto de su propio cuerpo como de su entorno, procesándola y actuando en consecuencia. De acuerdo con la definición de Wooldridge (1999) de una arquitectura abstracta para los agentes inteligentes, y teniendo en cuenta el aspecto relativo al cuerpo de un agente, se han identificado un conjunto de módulos arquitecturales esenciales: sensores, coordinación sensoriomotora, estado interno (memoria) y actuadores. Estos módulos son la base fundamental para el desarrollo de los siguientes procesos: percepción, razonamiento y acción. La cognición y el aprendizaje se pueden desarrollar en un agente como resultado de la combinación de los procesos anteriores durante la interacción con el mundo exterior y con su propio estado interno.

Se han identificado los siguientes componentes arquitecturales básicos:

- Cuerpo (***B***). La encarnación en un cuerpo es una característica principal de un agente situado (Dobbyn, Stuart 2003). El cuerpo de un agente puede ser físico o simulado por software (así como su entorno). La presencia de un cuerpo se caracteriza por la existencia de una frontera entre el propio cuerpo del agente y su entorno (***E***). El resto de componentes del agente se localiza normalmente dentro de estas fronteras. Es importante hacer la distinción entre el cuerpo del agente (o la planta, si se ve el problema desde la perspectiva de la teoría de control [Sanz, López & Hernández 2007]) y el entorno. Mientras que el primero es controlado directamente, el segundo se controla indirectamente. La distinción del cuerpo de un agente es importante ya que determina qué sensores se pueden usar, cómo funcionan los actuadores y en última instancia como los procesos de percepción y acción se ven afectados por la encarnación física del agente. Tener un cuerpo activo es esencial para la adquisición de conciencia.

- Sensores (***S***). Los sensores de un agente se encargan de recoger la información del entorno (sensores exteroceptivos o S_{ext}) o del propio cuerpo del agente (sensores propioceptivos o S_{propio}).

- Actuadores (***A***). Para poder interaccionar con el entorno el agente usa sus actuadores. El comportamiento final del agente se compone de las acciones que realizan en última instancia los actuadores.

- Coordinación Sensoriomotora (***R***). Tanto en los agentes puramente reactivos como en los deliberativos se puede decir que es el módulo de coordinación sensoriomotora el que se encarga de generar un comportamiento concreto en función de los estímulos externos y el estado interno del agente.

- Memoria (***M***). El estado interno del agente está representado tanto por su propia estructura como por la información almacenada. La memoria es el mecanismo por el cual el agente almacena tanto la información percibida como el conocimiento generado. Se considera que incluso los agentes que no mantienen un estado interno de forma explícita tienen un estado mínimo representado por su propia estructura. Por ejemplo, un conjunto pre-programado de reglas de coordinación sensoriomotoras.

- Mecanismo de Atención (***Att***). Mecanismo que permite dirigir ***S*** y ***A*** hacia un subconjunto específico del espacio sensoriomotor ($\boldsymbol{E_i}$).
- Mecanismo de Gestión de Múltiples Contextos en Memoria ($\boldsymbol{M^n}$). Mecanismo de representación de múltiples contextos en memoria.
- Mecanismo de Auto-Evaluación del Propio Estado (***SsA***). Mecanismo por el cual el agente es capaz de inspeccionar y evaluar sus propios parámetros de funcionamiento y consecución de objetivos.
- Mecanismo para la Representación del Yo (***I***). Mecanismo mediante el cual el agente mantiene un modelo actualizado del yo (*self*).
- Mecanismo para la Representación de Otros Yos (***O***). Mecanismo mediante el cual el agente es capaz de asignar un modelo del yo a otros agentes.
- Mecanismo de Comunicación Precisa (***AR***). Mecanismo mediante el cual el agente proporciona información precisa acerca de su estado interno.
- Mecanismo de Comunicación Verbal Precisa (***AVR***). Mecanismo mediante el cual el agente comunica verbalmente información precisa acerca de su estado interno.
- Mecanismo para Coordinar Múltiples Flujos de Conciencia ($\boldsymbol{R^n}$). Mecanismo mediante el cual el agente puede ejecutar y sincronizar varios hilos de procesamiento consciente, todos ellos relativos al mismo modelo del yo y al mismo cuerpo.

Todos estos componentes se definen como los bloques básicos de una arquitectura abstracta. Es decir, no se considera ninguna implementación particular ni ningún mecanismo sensoriomotor concreto (ver Figura 29).

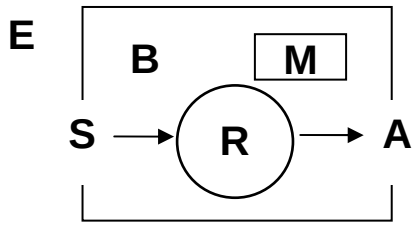


Figura 29. Componentes básicos de la arquitectura abstracta de un agente artificial.

Usando la arquitectura abstracta definida anteriormente se describen una serie de niveles de desarrollo cognitivo asociados a la conciencia. Estos niveles se definen de genéricamente, de forma que puedan referirse a cualquier agente dado, independientemente de los detalles de implementación y el dominio de aplicación. Tal y como indica Wooldridge (1999), se pueden obtener diferentes tipos de agente dependiendo de la forma concreta en la que se instancie la arquitectura abstracta. Con respecto a este punto, es importante resaltar que ningún componente específico de la arquitectura abstracta descrita se define como el responsable de la producción de conciencia. En contraposición, se mantiene la hipótesis de que la conciencia puede emerger a partir de la interacción que se produce entre los diferentes procesos presentes en el agente.

De acuerdo con la notación que se ha adoptado, se podría decir que diferentes funciones de coordinación sensoriomotora dan lugar a diferentes tipos de agentes, desde agentes puramente reactivos, pasando por agentes BDI (*Belief-Desire-Intention* – Creencia, Deseo, Intención) (Rao, Georgeff 1991), hasta los más avanzados agentes conscientes. Mientras que la coordinación sensoriomotora de los agentes reactivos se caracteriza por la aplicación de una correspondencia directa entre estímulos y acciones, la toma de decisiones en los agentes BDI se basa en el estado interno del agente, que contiene representaciones correspondientes a creencias, deseos e intenciones.

De forma análoga, en el marco de la arquitectura abstracta definida pueden existir diversos mecanismos de aprendizaje dependiendo de los componentes disponibles y la relación existente entre los mismos. Por ejemplo, los agentes puramente reactivos no pueden desarrollar procesos de aprendizaje debido a que no tienen memoria ni capacidad alguna de razonamiento.

5.3.2 Definición de Habilidades Cognitivas

En términos computacionales, la conciencia se podría considerar como un hilo secuencial único que integra la información sensorial concurrente y multimodal y coordina la acción voluntaria. Por lo tanto, la conciencia está estrechamente relacionada con la coordinación sensoriomotora (O'Regan, Noë 2001, O'Regan 2007). Consecuentemente, la escala propuesta pretende establecer una clasificación para agentes artificiales de acuerdo a la realización de funciones cognitivas asociadas a la conciencia desde el punto de vista de la coordinación sensoriomotora. En esta escala se define una serie de niveles que se caracterizan por determinadas funciones que un agente puede desempeñar.

Para determinar las funciones características que han de incluirse como parte de cada nivel se han considerado las capacidades básicas necesarias para la confección de una historia (o "película mental") a partir de información sensorial. En otras palabras, se ha considerado que un agente consciente es aquel que es capaz de generar imágenes mentales que sirven para producir comportamientos adapatativos en un medio complejo, dinámico y con numerosas fuentes de incertidumbre.

Es importante destacar que mientras que los posibles contenidos del espacio sensoriomotor (información sensorial y capacidades de acción) son dependientes de varios aspectos como el dominio de problema, las capacidades físicas del agente y la riqueza de las representaciones que forman su estado interno, las funciones cognitivas son, por el contrario, de aplicación general. Es decir, siempre se requieren las mismas funciones subyacentes para la producción de la conciencia. De hecho, cada tipo específico de organismo está diseñado para percibir distintas realidades en el mundo, de forma que se limita el espectro de contenido disponible para su procesamiento consciente. Por ejemplo, algunos organismos biológicos tienen la capacidad de percibir el campo magnético de la Tierra, mientras que otros, como los humanos, son completamente "ciegos" en este sentido. Dado que la escala propuesta está pensada para ser aplicada en implementaciones de Conciencia Artificial, en vez de analizar los contenidos específicos de la conciencia, se consideran las características funcionales generales que son necesarias para la producción de conciencia mediante el uso del espacio sensoriomotor disponible en cada caso.

De entre todas las capacidades cognitivas que un agente inteligente podría desarrollar potencialmente, el siguiente grupo de funciones caracteriza específicamente el comportamiento de un agente consciente:

- **La Función Reguladora de las Emociones.** Las emociones juegan un papel fundamental en la generación y modulación del comportamiento, incluso en los organismos que no tienen autoconciencia (Damasio 1999). En organismos con niveles más altos de conciencia, como los humanos, los sentimientos conscientes (representaciones autoconscientes de las emociones), posibilitan la interacción y competición entre respuestas emocionales y racionales (Damasio, Everitt & Bishop 1996, Koenigs et al. 2007). El aprendizaje efectivo en entornos complejos y desestructurados requiere estas capacidades cognitivas (ver Figura 30).
- **La Función Ejecutiva (FE).** El término Función Ejecutiva (o Control Ejecutivo) abarca todos los procesos responsables del control de la acción a alto nivel. Concretamente, se refiere a la gestión necesaria para mantener una meta especificada mentalmente y ser capaz de alcanzar dicha meta incluso en presencia de alternativas distractoras (Perner, Lang 1999). La capacidad de atención es una característica esencial de la FE. Representa la habilidad de un agente para dirigir la percepción y la acción, es decir, para seleccionar, de entre toda la información accesible en la mente, los contenidos que se almacenan en la memoria de trabajo. La planificación, la coordinación y el cambio de contexto (la habilidad de intercalar la ejecución de tareas diferentes) son también procesos clave que se incluyen dentro de la definición de FE (ver Figura 31).
- **La Teoría de la Mente[10] (TdM).** La Teoría de la Mente es la capacidad de atribuir estados mentales e intencionales a uno mismo y a otros (Vygotsky 1980). Desde el punto de vista del desarrollo cognitivo humano, Lewis propone distinguir entre cuatro estadios en el proceso de adquisición de la TdM (Lewis 2003): (1) "Yo sé", (2) "Yo sé que yo sé", (3) "Yo sé que tú sabes" y finalmente (4) "Yo sé que tú sabes que yo sé". Sin lugar a dudas, las funciones asociadas con la TdM son necesarias para las relaciones y el aprendizaje social (ver Figura 32).
- **La Capacidad de Aprendizaje.** Los agentes inteligentes se caracterizan por su capacidad para aprender y adaptarse a variaciones de los problemas para los que se han diseñado (Wooldridge 1999). En un contexto más ambicioso, se considera la Inteligencia Artificial General (Wang, Goertzel & Franklin 2008), para referirse a aquellos agentes que potencialmente puedan aprender y adaptarse incluso a entornos totalmente nuevos y desestructurados. En los organismos biológicos la capacidad de aprendizaje parece estar directamente correlacionada con el nivel de conciencia (Grossberg 2003, Andrade 1996) (ver Figura 33).

En esta tesis se mantiene la hipótesis de que la integración efectiva de todas estas funciones cognitivas es un requisito imprescindible para la construcción de una mente consciente artificial. Sin embargo, cada una de las funciones descritas anteriormente se pueden implementar independientemente o incluso se pueden integrar parcialmente con

[10] El concepto de *Teoría de la Mente* en este contexto no se refiere al campo de estudio filosófico del mismo nombre, sino que se refiere a una capacidad cognitiva que puede desarrollar un sujeto y que consiste en ser capaz de atribuir estados mentales intencionales a uno mismo y a otros individuos.

otras funciones cognitivas, dando lugar de esta forma a diferentes niveles de implementación o desarrollo de conciencia artificial (tal y como se describe en los siguientes apartados).

Con el objetivo de organizar todas estas funciones de forma incremental se ha considerado la sinergia que se produce por su interrelación. Esta organización particular será la base para la definición de una escala de niveles de conciencia artificial. Específicamente, se considera la siguiente organización (composición y orden) inspirada en la evolución de las capacidades cognitivas en la naturaleza:

- Desde el punto de vista del desarrollo de las emociones:
  - Emoción.
  - Emoción + Sentimiento.
  - Emoción + Sentimiento + Sentimiento Consciente (Damasio 1995).
  - Emoción + Sentimiento + Sentimiento Consciente + Emociones Simuladas.
- Desde el punto de vista del desarrollo de la percepción y la acción:
  - Percepción.
  - Percepción + Adaptación.
  - Percepción + Adaptación + Atención.
  - Percepción + Adaptación + Atención + Capacidad de Cambio de Contexto.
  - Percepción + Adaptación + Atención + Capacidad de Cambio de Contexto + Planificación.
  - Percepción + Adaptación + Atención + Capacidad de Cambio de Contexto + Planificación + Imaginación.
- Desde el punto de vista del desarrollo de la Teoría de la Mente:
  - “Yo sé”.
  - “Yo sé” + “Yo sé que yo sé”.
  - “Yo sé” + “Yo sé que yo sé” + “Yo sé que tú sabes”.
  - “Yo sé” + “Yo sé que yo sé” + “Yo sé que tú sabes” + “Yo sé que tú sabes que yo sé”.
- Desde el punto de vista del aprendizaje:
  - Aprendizaje por prueba y error.
  - Aprendizaje por prueba y error + Aprendizaje por refuerzo.
  - Aprendizaje por prueba y error + Aprendizaje por refuerzo + Aprendizaje Abstracto.
  - Aprendizaje por prueba y error + Aprendizaje por refuerzo + Aprendizaje Abstracto + Uso de herramientas.
  - Aprendizaje por prueba y error + Aprendizaje por refuerzo + Aprendizaje Abstracto + Uso de herramientas + Aprendizaje social.

El objetivo de la ordenación propuesta es representar en la escala la sinergia que emerge de la composición de estas funciones cognitivas, dando lugar a niveles de conciencia potencialmente más altos según estas funciones se integran de forma efectiva en un agente situado.

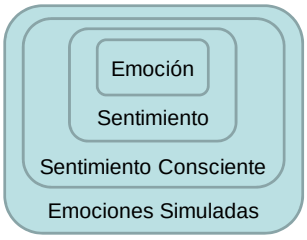


Figura 30. Jerarquía cognitiva de *ConsScale* desde el punto de vista de las emociones.

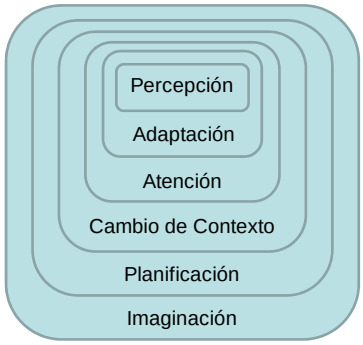


Figura 31. Jerarquía cognitiva de *ConsScale* desde el punto de vista de la percepción.

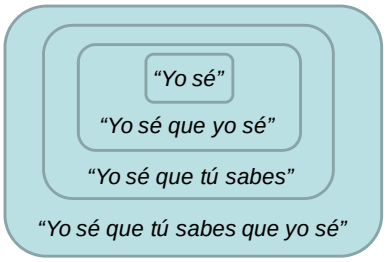


Figura 32. Jerarquía cognitiva de *ConsScale* desde el punto de vista de la TdM.

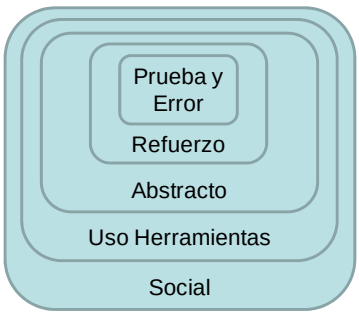


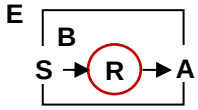
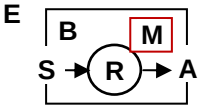
Figura 33. Jerarquía cognitiva de *ConsScale* desde el punto de vista del aprendizaje.

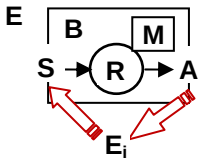
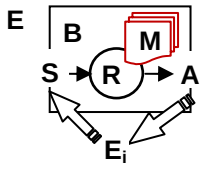
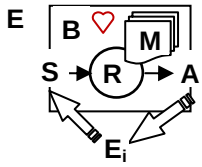
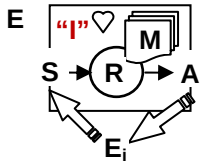
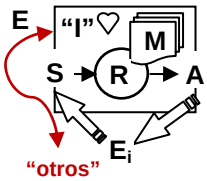
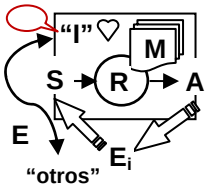
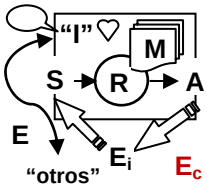
A continuación se describen los niveles de desarrollo cognitivo definidos en *ConsScale* (ver Tabla 4). Esta especificación de la escala consiste en una lista ordenada de niveles de conciencia potencialmente alcanzables en agentes artificiales. La escala se ha definido en base a:

- Arquitecturas de referencia (arquitectura abstracta característica de cada nivel basada en la definición dada en el Apartado 5.3.1).
- Habilidades cognitivas (funciones cognitivas características de cada nivel).
- Comportamiento asociado (comportamientos característicos que demuestran la existencia de ciertas funciones cognitivas).

A forma de analogía ilustrativa, se han asignado ejemplos provenientes de la filogenia biológica y de la ontogenia humana a cada uno de los niveles de conciencia definido en *ConsScale*. La Tabla 5 contiene una lista de las habilidades cognitivas definidas para cada uno de los niveles de *ConsScale*. La caracterización concreta de cada uno de los niveles se expone en los siguientes apartados (del Apartado 5.3.3 al 5.3.15), indicando qué tipo de sinergia existe entre las diversas funciones cognitivas y los tipos de comportamientos que éstas son capaces de generar.

Tabla 4. Resumen de los niveles de conciencia en *ConsScale*.

Nivel de Conciencia	**Arquitectura Abstracta**	**Descripción Breve**	**Comporta-miento típico**	**Filogenia**	**Humano**
Nivel -1. Sin Cuerpo Definido (*Disembodied*).		Los límites del cuerpo del agente no están bien definidos. Se puede confundir con el entorno.	Ninguno. No es un agente situado.	Aminoácido como parte de una proteína.	n/a
Nivel 0. Aislado (*Isolated*).		Existe una distinción clara entre el cuerpo y el entorno, pero no hay procesamiento autónomo alguno.	Ninguno. No es un agente situado.	Cromosoma aislado.	n/a
Nivel 1. Pre-Funcional. (*Decontrolled*).		El agente tiene sensores y actuadores, pero no existe relación funcional entre ellos.	Ninguno. No es un agente situado.	Bacteria muerta.	n/a
Nivel 2. Reactivo. (*Reactive*).		El agente muestra respuestas reactivas fijas. ***R*** establece la salida de ***A*** en base a una función fija que toma ***S*** como única entrada.	Reflejos. El aprendizaje se da sólo a nivel evolutivo (entre generaciones).	Virus.	n/a
Nivel 3. Adaptativo. *(Adaptive).*		Las acciones se determinan de forma dinámica en función de la información adquirida por ***S*** y la memoria del agente.	Una capacidad básica de aprendizaje y la propiocepción permiten comportamientos de orientación y posicionamiento	Lombriz de tierra.	Bebé de 1 mes.

Nivel de Conciencia	**Arquitectura Abstracta**	**Descripción Breve**	**Comporta-miento típico**	**Filogenia**	**Humano**
Nivel 4. Atencional ***(****Attentional*).		Un mecanismo de atención selecciona contenidos específicos ***E_i*** de ***S*** y ***M***. Aparecen las emociones.	La capacidad de dirigir la atención hacia un ***E_i*** seleccionado junto con el aprendizaje permite los comportamientos de ataque y huída.	Pez.	Bebé de 5 meses.
Nivel 5. Ejecutivo. (*Executive*).		El agente puede intercalar la persecución de múltiples metas ya que estas se representan explícitamente en la memoria	La capacidad de cambio de contexto permite la consecución de varias metas de forma simultánea. Aprendizaje emocional.	Mamífero cuadrúpedo.	Bebé de 9 meses.
Nivel 6. Emocional. *(Emotional).*		Emociones globales y estables. Estadio 1 del desarrollo de TdM: “Yo sé”.	Las emociones globales permiten una auto-evaluación y modulan el comportamiento.	Mono.	Bebé de 1 año.
Nivel 7. Autoconsciente (*Self-Conscious*).		Estadio 2 del desarrollo de TdM: “Yo sé que yo sé”.	La auto-referencia hace posible la planificación avanzada. Uso de herramientas.	Mono.	Bebé de año y medio.
Nivel 8. Empático. *(Empathic*).		Estadio 3 del desarrollo de TdM: “Yo sé que tú sabes”.	Fabricación de herramientas. Comportamiento social.	Chimpancé.	Bebé de 2 años.
Nivel 9. Social (*Social*).		Estadio 4 del desarrollo de TdM: “Yo sé que tú sabes que yo sé”.	Comunicación precisa. Capacidades lingüísticas. Capacidad para desarrollar una cultura.	Humano.	Niño de 4 años.
Nivel 10. Androide. *(Human-Like*).		Conciencia como la humana. Entorno adaptado (***E_c***).	Comunicación verbal precisa. Comportamiento modulado por la cultura (***E_c***).	Humano.	Adulto.

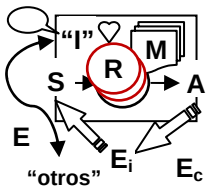
Nivel de Conciencia	Arquitectura Abstracta	Descripción Breve	Comporta-miento típico	Filogenia	Humano
Nivel 11. Super-Consciente. *(Super-Conscious)*.		Varios flujos de conciencia con el mismo yo.	Capacidad de sincronizar y coordinar varios flujos de conciencia.	n/a	n/a

Tabla 5. Resumen de las habilidades cognitivas definidas en *ConsScale*.

Nivel	Descripción	Habilidades Cognitivas
-1	Sin cuerpo definido.	Ninguna.
0	Aislado.	Ninguna.
1	Pre-funcional.	Ninguna.
2	Reactivo.	**$CS_{2,1}$**: Capacidad de producir respuestas motoras fijas o reactivas (reflejos).
3	Adaptativo.	**$CS_{3,1}$**: Adquisición autónoma y adaptativa de nuevas respuestas reactivas. **$CS_{3,2}$**: Uso de los sensores propioceptivos para generar respuestas adaptativas. **$CS_{3,3}$**: Selección de información sensorial relevante. **$CS_{3,4}$**: Selección de información motora relevante. **$CS_{3,5}$**: Selección de información relevante en memoria. **$CS_{3,6}$**: Evaluación (positiva o negativa) de objetos o sucesos seleccionados. **$CS_{3,7}$**: Selección de los contenidos que necesitan ser almacenados en memoria.
4	Atencional.	**$CS_{4,1}$**: Aprendizaje por prueba y error. Re-evaluación de los objetos o sucesos seleccionados. **$CS_{4,2}$**: Comportamiento dirigido hacia determinados objetivos, como seguimiento y escape o acercamiento y alejamiento. **$CS_{4,3}$**: Evaluación del propio rendimiento en la consecución de una meta simple. **$CS_{4,4}$**: Capacidad básica de planificación: cálculo de las *n* siguientes acciones secuenciales. **$CS_{4,5}$**: Representación espacial relativa (*depictive*) de los objetos percibidos.
5	Ejecutivo	**$CS_{5,1}$**: Habilidad para cambiar de contexto entre múltiples tareas. **$CS_{5,2}$**: Persecución de múltiples metas. **$CS_{5,3}$**: Evaluación del rendimiento en la consecución de múltiples metas. **$CS_{5,4}$**: Aprendizaje por refuerzo autónomo. **$CS_{5,5}$**: Capacidad avanzada de planificación teniendo en cuenta todas las metas activas. **$CS_{5,6}$**: Capacidad para generar contenidos mentales seleccionados con significados basados en la interacción del agente con el medio (*grounded meaning*) y que integren diversas modalidades sensoriales en perceptos explícitos diferenciados (ver Apartado 2.1.3.5).

6	Emocional	**$CS_{6,1}$**: Evaluación del estado propio (emociones globales). **$CS_{6,2}$**: Las emociones globales causan efectos en el cuerpo del agente. **$CS_{6,3}$**: Representación de los efectos de las emociones en el organismo (sentimientos). **$CS_{6,4}$**: Capacidad para mantener un mapa preciso y actualizado del esquema corporal. **$CS_{6,5}$**: Aprendizaje abstracto (generalización de lecciones aprendidas). **$CS_{6,6}$**: Capacidad para representar un flujo de perceptos integrados que incluyan el estado propio.
7	Autoconsciente	**$CS_{7,1}$**: Representación de la relación entre el yo y la percepción. **$CS_{7,2}$**: Representación de la relación entre el *yo* y la acción. **$CS_{7,3}$**: Representación de la relación entre el *yo* y los sentimientos. **$CS_{7,4}$**: Capacidad de auto-reconocimiento. **$CS_{7,5}$**: Planificación avanzada incluyendo al *yo* como actor en los planes. **$CS_{7,6}$**: Uso de estados imaginados en la planificación (imaginación). **$CS_{7,7}$**: Aprendizaje del uso de herramientas. **$CS_{7,8}$**: Capacidad para representar y auto-comunicarse el contenido mental explícito (flujo continuo interno de perceptos – imágenes mentales internas).
8	Empático	**$CS_{8,1}$**: Habilidad para modelar a otros individuos como yos subjetivos e intencionales. **$CS_{8,2}$**: Aprendizaje por imitación de un individuo homólogo. **$CS_{8,3}$**: Habilidad para colaborar con otros individuos en la persecución de una meta común. **$CS_{8,4}$**: Planificación social (utilizando planes sociales, que tienen en cuenta a los otros individuos participantes en el plan). **$CS_{8,5}$**: Habilidad para crear nuevas herramientas. **$CS_{8,6}$**: Las imágenes mentales internas están enriquecidas con contenidos relacionados con los modelos de otros y la relación entre el yo y otros yos.
9	Social	**$CS_{9,1}$**: Habilidad para desarrollar estrategias Maquiavélicas, como la mentira o la astucia. **$CS_{9,2}$**: Aprendizaje social (aprendizaje de nuevas estrategias Maquiavélicas). **$CS_{9,3}$**: Habilidades avanzadas de comunicación (comunicación precisa del contenido mental). **$CS_{9,4}$**: Los grupos son capaces de desarrollar una cultura. **$CS_{9,5}$**: Habilidad para modificar y adaptar el entorno a las necesidades del agente.
10	Androide	**$CS_{10,1}$**: Reporte verbal preciso. Capacidades lingüísticas avanzadas. **$CS_{10,2}$**: Habilidad para pasar el Test de Turing. **$CS_{10,3}$**: Los grupos son capaces de desarrollar una civilización y cultura y tecnología avanzadas.
11	Super-consciente	**$CS_{11,1}$**: Habilidad para gestionar varios flujos de conciencia.

Para formalizar la propuesta de medida basada en las capacidades cognitivas descritas anteriormente (ver Tabla 5) se ha aplicado la teoría del orden (Stanley 2000). Las relaciones entre los diferentes CS se han formalizado considerando un conjunto finito parcialmente ordenado (o *poset*), que se puede representar mediante su diagrama de Hasse asociado (ver Figura 34). La jerarquía de habilidades cognitivas que se ha definido está basada en una *relación binaria de orden estricto* representada por el símbolo "<" y que indica dependencia cognitiva. Por lo tanto, el conjunto de todos los CS en *ConsScale* (CSS) parcialmente ordenado en función de la relación de dependencia cognitiva se puede considerar como un poset (CSS, <).

Por ejemplo, la relación binaria de orden estricto $CS_{6,4} < CS_{7,4}$ (representada en la Figura 34 por una flecha ascendente desde el nodo $CS_{6,4}$ al nodo $CS_{7,4}$) significa que $CS_{7,4}$ *engloba a* $CS_{6,4}$. En otras palabras, *la capacidad de auto-reconocimiento* ($CS_{7,4}$) requiere la capacidad de *mantener un mapa preciso y actualizado del esquema corporal* ($CS_{6,4}$). Es decir, existe una relación de dependencia cognitiva e inclusión entre estas dos funciones. De forma análoga se han identificado otras relaciones de dependencia entre el resto de las habilidades cognitivas definidas en *ConsScale* tal y como se ilustra en la Figura 34.

Como regla general, la definición de las habilidades cognitivas CS y su jerarquía asociada satisface la propiedad de que ninguna función requiere el cumplimiento de otra función de nivel superior (esta propiedad puede observarse gráficamente en la Figura 34, dónde todas las flechas del diagrama representan relaciones en sentido ascendente). El conjunto CSS no es un conjunto totalmente ordenado porque no todas las funciones cognitivas son *comparables* (en el sentido matemático establecido por la relación binaria "<" definida anteriormente). Las relaciones de dependencia se han establecido considerando la ontogenia humana y la filogenia biológica tal y como se ha explicado anteriormente.

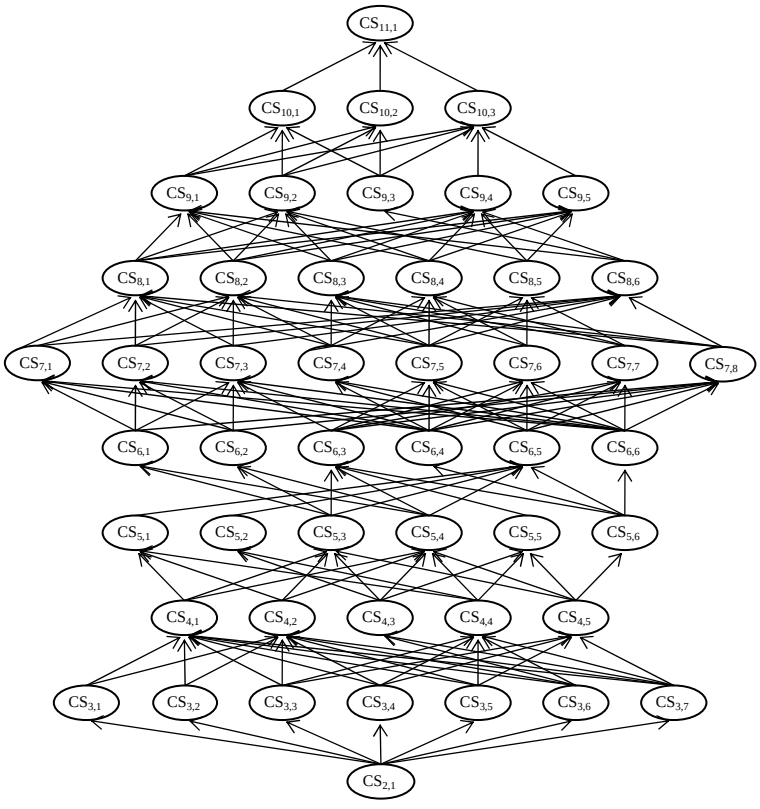


Figura 34. Diagrama de Hasse del poset CCS.

El poset (CSS, <) se compone de múltiples subconjuntos interrelacionados que representan el desarrollo, dependencia y composición incremental de funciones cognitivas específicas. Considerando por ejemplo la función de TdM se puede observar que el siguiente orden parcial está incluido en CSS:

$$CS_{6,1\text{-}6} < CS_{7,1\text{-}5} < CS_{8,1\text{-}4} < CS_{9,1\text{-}2}$$

La relación especificada anteriormente pone de manifiesto que el poset definido cumple con la propiedad transitiva. Es decir, el Estadio 4 del desarrollo de la TdM (*"yo sé que tú sabes que yo sé"*) depende cognitivamente del Estadio 3 *("yo sé que tú sabes"*), que a su vez depende del Estadio 2 (*"yo sé que sé"*). Aplicando la propiedad transitiva se desprende que el Estadio 4 también depende cognitivamente del Estadio 2. Se pueden realizar análisis análogos para el resto de las relaciones de orden establecidas en la jerarquía definida. En el caso del orden parcial correspondiente a la TdM, todas las relaciones de dependencia se derivan de las siguientes relaciones binarias de orden:

$CS_{6,1\text{-}6}$ (*"yo sé"*) < $CS_{7,1\text{-}5}$ (*"yo sé que sé"*)
$CS_{7,1\text{-}5}$ (*"yo sé que sé"*) < $CS_{8,1\text{-}4}$ (*"yo sé que tú sabes"*)
$CS_{8,1\text{-}4}$ (*"yo sé que tú sabes"*) < $CS_{9,1\text{-}2}$ (*"yo sé que tú sabes que yo sé"*)

Además, como la relación binaria de orden establecida es estricta (no existe la posibilidad de que $CS_{i,j} \leq CS_{k,m}$), el poset (CSS, <) también satisface las propiedades de irreflexibilidad y asimetría. Estas propiedades, junto con el hecho de todas las funciones CS definidas en un mismo nivel son incomparables, permite una caracterización formal de los niveles de desarrollo cognitivo de la conciencia definidos en *ConsScale*.

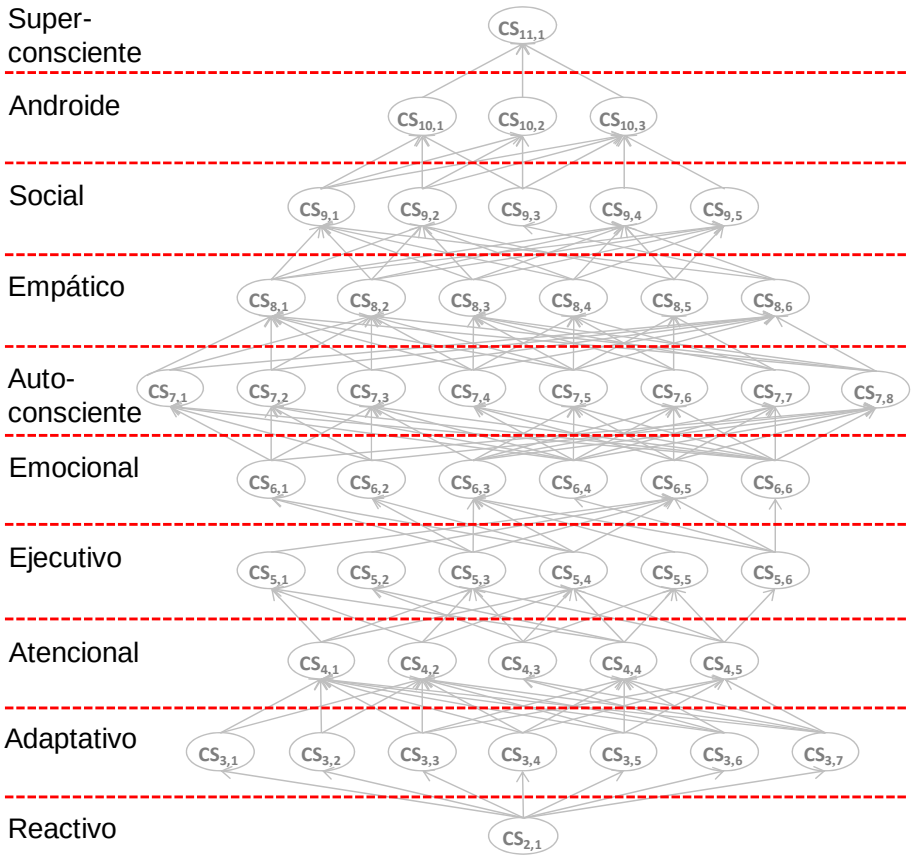


Figura 35. Niveles de *ConsScale* en relación a la jerarquía cognitiva CSS.

La definición del poset (CSS, <) permite asimismo identificar subconjuntos parcialmente ordenados correspondientes a grupos de habilidades cognitivas como los

expuestos anteriormente. Por ejemplo, considerando específicamente las habilidades relativas a la TdM puede identificarse el subconjunto marcado en la Figura 36.

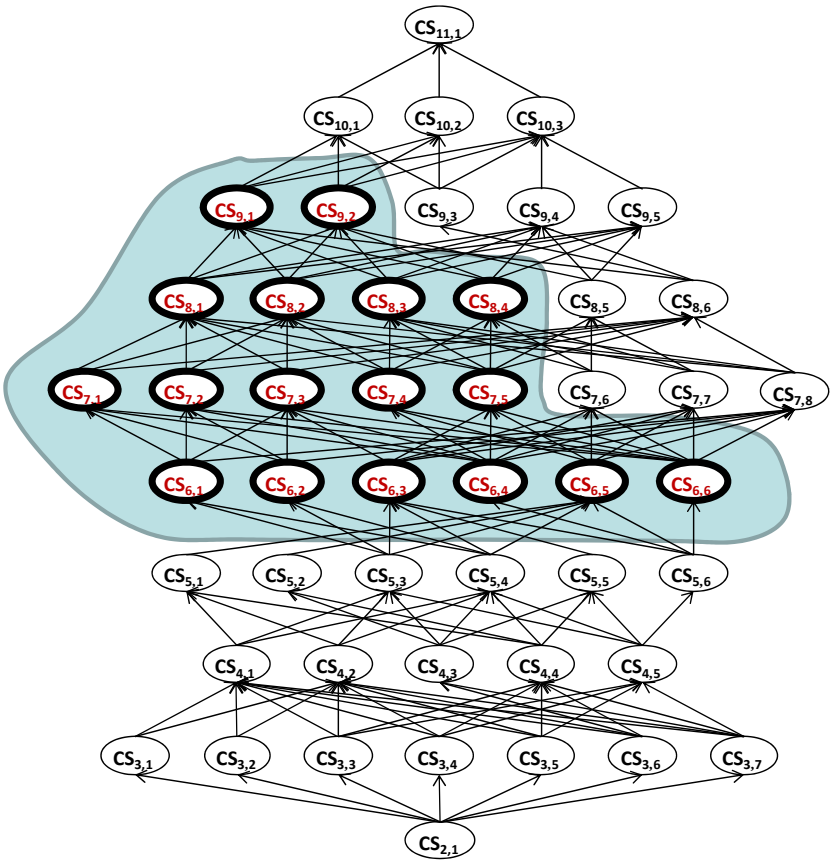


Figura 36. Subconjunto TdM en el diagrama de Hasse del poset CCS.

A continuación se describen en detalle las principales características de cada uno de los niveles definidos en la escala *ConsScale*.

5.3.3 Nivel -1 (Sin cuerpo definido o "*Disembodied*")

5.3.3.1 Descripción

Este es un nivel inicial de referencia que se corresponde con implementaciones muy simples que ni siquiera tienen una definición clara de las fronteras que separan el agente del entorno que lo rodea. En otras palabras, este nivel se refiere a implementaciones que no pueden ser consideradas como agentes y se pueden confundir fácilmente con el resto del entorno.

Desde el punto de vista del desarrollo cognitivo propuesto en *ConsScale*, las entidades de nivel -1 podrían considerarse como "proto-agentes". Aunque este nivel no se use de forma práctica en la evaluación de posibles implementaciones, su definición sirve para poner de manifiesto la importancia del cuerpo (**B**) como requisito básico para la descripción de un agente situado.

No hay habilidades cognitivas características de este nivel. Una analogía tomada del mundo biológico para los individuos de este nivel sería la consideración de un aminoácido como parte de una proteína.

El resto de la escala consiste en un conjunto de doce niveles (del 0 al 11, ambos inclusive), en la que los niveles más altos incluyen como parte de su propia definición a todos los niveles inferiores. Usando estos niveles, se puede caracterizar cualitativamente el desarrollo cognitivo de un agente artificial.

5.3.3.2 Arquitectura Característica

Teniendo en cuenta los componentes definidos en la arquitectura abstracta de referencia, un agente de nivel -1 ni siquiera posee un cuerpo (**B**) claramente definido (ver Figura 37).

Figura 37. Arquitectura característica de un agente de nivel *ConsScale* -1.

5.3.3.3 Habilidades Cognitivas

No hay habilidades cognitivas definidas para el nivel -1 de *ConsScale*.

5.3.4 Nivel 0 (Aislado o "*Isolated*")

5.3.4.1 Descripción

Al igual que el nivel -1, el nivel 0 es un nivel de referencia conceptual que permite resaltar la importancia de la interacción con el entorno (*situatedness*), cuya base se encuentra en la existencia de un cuerpo.

En este nivel, aunque hay una clara distinción entre la propia implementación (su cuerpo) y el entorno, hay una falta total de procesamiento autónomo y no existen sistemas sensoriales ni motores. Por lo tanto, una entidad de nivel 0 consiste únicamente en un cuerpo inerte que no presenta ninguna funcionalidad ni interacción activa con el medio (excepto la inevitablemente provocada por las propiedades físicas de la entidad).

En este nivel tampoco hay habilidades cognitivas características. Un cromosoma aislado podría considerarse como una analogía válida para este nivel.

5.3.4.2 Arquitectura Característica

La arquitectura abstracta del nivel 0 de *ConsScale* se caracteriza por la existencia de un único componente: el cuerpo, que no tiene ninguna función asociada. Es decir, se trata de un cuerpo inerte (ver Figura 38). Como se ha mencionado en la definición de la arquitectura abstracta (Apartado 5.3.1) puede tratarse de un cuerpo físico real o de un cuerpo simulado por ordenador.

Figura 38. Arquitectura característica de un agente de nivel *ConsScale* 0.

5.3.4.3 Habilidades Cognitivas

Al tratarse únicamente de un cuerpo inerte, no hay habilidades cognitivas definidas para este nivel.

5.3.5 Nivel 1 (Pre-funcional o "*Decontrolled*")

5.3.5.1 Descripción

Este nivel se refiere a aquellas implementaciones en las que los subsistemas correspondientes a sensores y actuadores están presentes, pero o bien no funcionan o no existe una relación funcional entre ellos. Como ni la adquisición de información ni la generación de acciones funcionan o no están relacionadas, todavía no se pueden definir habilidades cognitivas características a este nivel. El cuerpo inerte de una bacteria muerta podría ser una analogía plausible tomada del mundo biológico.

5.3.5.2 Arquitectura Característica

De nuevo, la arquitectura correspondiente al nivel 1 ilustra un tipo de agente conceptual, que no suele darse en la realidad (ver Figura 39). Aunque están presentes los subsistemas de adquisición de información (sensores) y los actuadores, éstos no realizan operación alguna o la realizan de forma descontrolada.

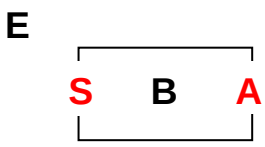


Figura 39. Arquitectura característica de un agente de nivel *ConsScale* 1.

5.3.5.3 Habilidades Cognitivas

Puesto que no existe relación funcional entre los sensores y los actuadores, no hay habilidades cognitivas definidas para este nivel.

5.3.6 Nivel 2 (Reactivo o "Reactive")

5.3.6.1 Descripción

En este nivel, tanto los subsistemas de sensores como los actuadores son funcionales y están relacionados entre ellos mediante funciones predefinidas. El comportamiento de estos agentes está caracterizado por la producción de respuestas reactivas fijas en base a los datos de entrada capturados por los sensores. Por lo tanto, la única capacidad cognitiva característica de este nivel es la de un agente situado que responde al entorno siempre con los mismos reflejos. Una analogía biológica para este nivel podría ser un virus.

El nivel 2 de *ConsScale* se corresponde con un agente reactivo clásico que carece de memoria explícita y tampoco dispone de mecanismos de aprendizaje. Es a partir del nivel 2, cuando los agentes empiezan a utilizar el propio entorno que les rodea como medio para cerrar el bucle de retroalimentación entre la acción y la percepción. Por lo tanto, todos los agentes de nivel 2 o superior pueden considerarse agentes situados.

Aunque la escala se centra principalmente en la evaluación de agentes individuales, es importante destacar que incluso en el nivel 2 pueden existir procesos adicionales de aprendizaje y adaptación en el plano evolutivo (suponiendo que los agentes sean capaces de replicarse, mutar y evolucionar). Por ejemplo, aunque las reglas de control reactivas son fijas en un único individuo de nivel 2, podría aparecer un proceso de adaptación de las respuestas reactivas a lo largo de sucesivas generaciones en una población de agentes de nivel 2.

5.3.6.2 Arquitectura Característica

La arquitectura característica de un agente de nivel 2 es la correspondiente a un agente reactivo clásico. Los componentes principales son los sensores, los actuadores y un componente adicional (**R** o coordinación sensoriomotora) que establece las acciones a realizar en función de los estímulos recibidos (ver Figura 40).

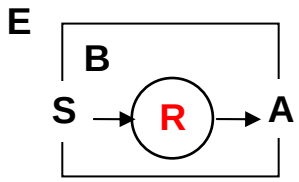


Figura 40. Arquitectura característica de un agente de nivel *ConsScale* 2.

5.3.6.3 Habilidades Cognitivas

Para el nivel 2 de *ConsScale* se define una única habilidad cognitiva genérica:

- **$CS_{2,1}$**: Capacidad de producir respuestas motoras fijas o reactivas (reflejos).

5.3.7 Nivel 3 (Adaptativo o "Adaptive")

5.3.7.1 Descripción

A este nivel las acciones del agente se generan dinámicamente en función tanto de la memoria como de la información que se obtiene del entorno por medio de los sensores. Las habilidades cognitivas características de este nivel son la capacidad básica para aprender nuevos reflejos y el uso de sensores propioceptivos para la generación de comportamientos básicos como los de orientación y posicionamiento. La lombriz de tierra sería una analogía biológica ilustrativa para este nivel.

Un agente de nivel 3 se corresponde con la forma más simple de un agente deliberativo. En este nivel, el estado interno del agente se mantiene en un sistema de memoria. La coordinación sensoriomotora (**R**) se genera en función de la información percibida y la información recordada. En este nivel, también se considera la presencia de sensores propioceptivos, aunque esto por sí solo no es suficiente para generar autoconciencia. Los mecanismos de percepción propioceptiva permiten que parte del estado interno del agente forme parte de la entrada de la función de coordinación sensoriomotora.

Los agentes de nivel 3 también tienen mecanismos de aprendizaje que les permiten descubrir nuevos comportamientos reactivos. Es decir, la respuesta a un determinado estado del entorno no es fija, sino que es una función de la información adquirida por **S** y el estado interno del agente (**M**).

El nivel 3 también se puede ver como una evolución del nivel 2 en la que ha aparecido la capacidad de aprender nuevos reflejos.

5.3.7.2 Arquitectura Característica

La arquitectura típica del nivel 3 de *ConsScale* está formada por los siguientes componentes principales: sensores (incluyendo sensores propioceptivos), actuadores, coordinación sensoriomotora y sistema de memoria (ver Figura 41).

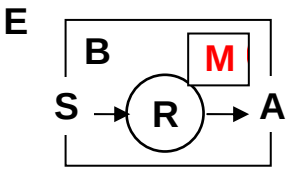


Figura 41. Arquitectura característica de un agente de nivel *ConsScale* 3.

5.3.7.3 Habilidades Cognitivas

En el nivel 3 de *ConsScale* se definen las siguientes habilidades cognitivas de carácter genérico:

- **$CS_{3,1}$**: Adquisición autónoma y adaptativa de nuevas respuestas reactivas.
- **$CS_{3,2}$**: Uso de los sensores propioceptivos para generar respuestas adaptativas.
- **$CS_{3,3}$**: Selección de información sensorial relevante.
- **$CS_{3,4}$**: Selección de información motora relevante.
- **$CS_{3,5}$**: Selección de información relevante en memoria.
- **$CS_{3,6}$**: Evaluación (positiva o negativa) de objetos o sucesos seleccionados.
- **$CS_{3,7}$**: Selección de los contenidos que necesitan ser almacenados en memoria.

5.3.8 Nivel 4 (Atencional o "*Attentional*")

5.3.8.1 Descripción

A este nivel el comportamiento del agente está modulado por la influencia que ejerce un mecanismo de atención. La atención selecciona contenidos específicos del repertorio total de contenidos disponibles a través de los sensores y de la memoria. Además, los contenidos seleccionados se evalúan positiva o negativamente, constituyendo esto la semilla de las emociones. Las capacidades cognitivas típicas de un agente de nivel 4 permiten la producción de comportamientos de ataque y de escape (o de acercamiento y alejamiento). Los peces podrían constituir una analogía biológica para este nivel.

Gracias al mecanismo de atención, los procesos de aprendizaje se pueden dirigir explícitamente hacia determinados objetos o sucesos. Además, los procesos de aprendizaje implícito, como la adquisición de nuevos reflejos (característica del nivel anterior), también se dan al mismo tiempo.

Los agentes de nivel 4 son capaces de dirigir la atención (usando el componente **Att**) a un subconjunto seleccionado de los contenidos percibidos del entorno (**E_i**), mientras que otras variables ambientales, que son adquiridas en **S**, no son procesadas explícitamente en **R**. Los objetos o sucesos seleccionados por el foco de atención son evaluados automáticamente en base a las metas del propio agente. Esto permite que las

respuestas subsiguientes del agente hacia esos estímulos se adapten (este mecanismo se puede considerar la base para el desarrollo de las emociones (Damasio 1999)).

Los agentes atencionales son capaces de desarrollar comportamientos dirigidos, como los de ataque o escape, y también son capaces de implementar mecanismos de aprendizaje por prueba y error. La habilidad de prestar atención hacia estímulos específicos permite la formación de comportamientos dirigidos. Por ejemplo, un agente puede desarrollar comportamientos claramente relacionados con objetivos específicos, como la persecución o la huída. Adicionalmente, los agentes de nivel 4 tienen mecanismos primitivos para las emociones ya que los objetos a los que se presta atención son evaluados elementalmente como positivos o negativos. Una emoción positiva desencadena un comportamiento de acercamiento o un vínculo hacia el objeto seleccionado. Por el contrario, una emoción negativa desencadena un comportamiento de alejamiento y refuerzo de las fronteras entre el agente y el objeto seleccionado. Adicionalmente, aparece en este nivel una nueva relación entre la memoria y las emociones. Como se ha demostrado con los organismos biológicos, las emociones influyen en gran medida en la selección de los contenidos que se almacenan en memoria (LaBar, Cabeza 2006, Schacter 1989). En resumen, un agente de nivel 4 puede considerarse una evolución de un agente de nivel 3 en el que ha aparecido la capacidad de atención.

5.3.8.2 Arquitectura Característica

La arquitectura característica del nivel 4 contiene un componente adicional a los ya existentes en el nivel 3, el mecanismo de atención o **Att** (ver Figura 42).

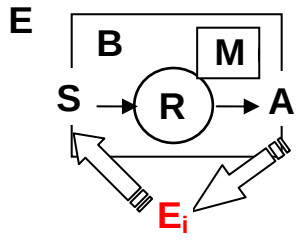


Figura 42. Arquitectura característica de un agente de nivel *ConsScale* 4.

5.3.8.3 Habilidades Cognitivas

En el nivel 4 de *ConsScale* se definen las siguientes habilidades cognitivas de carácter genérico:

- **$CS_{4,1}$**: Aprendizaje por prueba y error. Re-evaluación de los objetos o sucesos seleccionados.
- **$CS_{4,2}$**: Comportamiento dirigido hacia determinados objetivos, como seguimiento y escape o acercamiento y alejamiento.
- **$CS_{4,3}$**: Evaluación del propio rendimiento en la consecución de una meta simple.

- **$CS_{4,4}$**: Capacidad básica de planificación: cálculo de las *n* siguientes acciones secuenciales.
- **$CS_{4,5}$**: Representación espacial relativa (*depictive* [Aleksander, Dunmall 2003]) de los objetos percibidos.

5.3.9 Nivel 5 (Ejecutivo o "*Executive*")

5.3.9.1 Descripción

Los agentes de este nivel son capaces de intercalar múltiples metas ya que son capaces de almacenar en memoria distintos conjuntos de trabajo. Las habilidades cognitivas características de este nivel son el cambio de contexto (habilidad para pasar de una tarea a otra) y el aprendizaje emocional básico. La capacidad de cambio de contexto implica que el agente puede suspender la realización de una tarea dada para retomarla más tarde, estableciendo prioridades entre todas las tareas pendientes. El agente puede perseguir múltiples metas, asignando más tiempo y esfuerzo a aquellas que reportan mayores beneficios emocionales. Los mamíferos cuadrúpedos son una analogía biológica ilustrativa para este nivel.

Un agente de nivel 5 se caracteriza por una capacidad de razonamiento más compleja y una representación más rica del estado interno que permite la implementación de mecanismos de cambio de contexto. La consecución de múltiples metas se consigue gracias a un mecanismo de coordinación que es capaz de mover el foco de atención de forma efectiva de una tarea a otra. Un agente de nivel 5 también está dotado de un mecanismo que permite evaluar el propio desempeño en la consecución de las metas pendientes. Se trata del mecanismo de auto-evaluación del propio estado (**SsA**), que puede ser identificado como las emociones (Ciompi 2003, Damasio 1999).

La presencia de emociones asociadas a objetos, sucesos, y ahora también a las propias acciones del agente, permite el desarrollo de procesos de aprendizaje por refuerzo. Un agente de nivel 5 es aquel que exhibe capacidades de cambio de contexto y aprendizaje emocional básico (aprendizaje por refuerzo). Otras características de un agente ejecutivo son la capacidad de planificación avanzada y la aplicación de las emociones al cambio de contexto: el agente tiende a asignar más tiempo y esfuerzo a aquellas tareas que resultan más gratificantes para él. En resumen, un agente de nivel 5 se puede considerar como una evolución de un agente de nivel 4 en la que ha aparecido la capacidad de perseguir múltiples metas e intercalar de forma efectiva la realización de múltiples tareas.

5.3.9.2 Arquitectura Característica

La arquitectura abstracta correspondiente al nivel 5 de *ConsScale*, incluye el componente $\mathbf{M^n}$ (mecanismo de gestión de múltiples contextos en memoria) a la arquitectura definida para el nivel anterior (ver Figura 43).

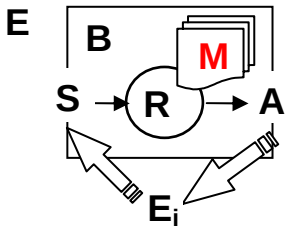


Figura 43. Arquitectura característica de un agente de nivel *ConsScale* 5.

5.3.9.3 Habilidades Cognitivas

En el nivel 5 de *ConsScale* se definen las siguientes habilidades cognitivas de carácter genérico:

- **$CS_{5,1}$**: Habilidad para cambiar de contexto entre múltiples tareas.
- **$CS_{5,2}$**: Persecución de múltiples metas.
- **$CS_{5,3}$**: Evaluación del rendimiento en la consecución de múltiples metas.
- **$CS_{5,4}$**: Aprendizaje por refuerzo autónomo.
- **$CS_{5,5}$**: Capacidad avanzada de planificación teniendo en cuenta todas las metas activas.
- **$CS_{5,6}$**: Capacidad para generar contenidos mentales seleccionados con significados basados en la interacción del agente con el medio (grounded meaning) y que integren diversas modalidades sensoriales en perceptos explícitos diferenciados (ver Apartado 2.1.3.5).

5.3.10 Nivel 6 (Emocional o "*Emotional*")

5.3.10.1 Descripción

Este nivel se caracteriza por la capacidad de desarrollar el estadio 1 de la Teoría de la Mente o TdM: "Yo sé". La TdM es la capacidad de atribuir estados mentales a uno mismo y a otros sujetos (Vygotsky 1980). En el nivel 6 de *ConsScale* los sentimientos aparecen como representaciones de los cambios que experimenta el organismo debido a las emociones (Damasio 1999). El sentido de "yo sé" aparece en el agente gracias a la representación de las relaciones existentes entre las emociones y los estados del organismo. La habilidad cognitiva característica del nivel 6 es la capacidad de aprendizaje emocional. El agente generaliza las lecciones aprendidas y las aplica a su comportamiento global. Además, las emociones se asignan también al propio *yo* y al proceso de auto-monitorización, produciendo una auto-evaluación continua que da lugar al "yo sé". Los monos son una analogía biológica plausible para el nivel 6.

El nivel 6 es el primer nivel de la escala *ConsScale* en el que se puede considerar que un agente es hasta cierto punto consciente (pero no autoconsciente). La principal característica de este nivel es, como se ha mencionado anteriormente, el desarrollo del estadio 1 de la TdM. Aparecen las emociones complejas, que se construyen como

combinaciones de las emociones básicas presentes en los niveles anteriores. Estas emociones no evalúan sólo estímulos externos o el propio estado interno, sino que se generan emociones de fondo que evalúan la relación del agente con su entorno.

Como en este nivel los agentes cuentan con una representación precisa de su estructura corporal y sus capacidades físicas, es más fácil establecer una separación entre el *yo* y el resto del mundo. Por ejemplo, un robot antropomórfico de nivel 6 actualizará su modelo del yo al descubrir la diferencia existente entre tocar con la mano un objeto extraño y tocar su propio cuerpo. En el segundo caso, se produce una notificación de los sensores de contacto de la zona del cuerpo que toca la mano al mismo tiempo que se ejecuta la acción. Esta capacidad permite establecer una distinción clara entre los objetos que se pueden controlar directamente (el propio cuerpo) y los que no.

La sensación correspondiente al "yo sé" aparece en el agente gracias a la representación que se genera de la respuesta del organismo a las emociones (Damasio 1999). Existe un modelo implícito del *yo* que permite asociar las emociones con el desempeño del propio agente. Usando estas representaciones el agente es capaz de generalizar en sus procesos de aprendizaje. Mientras que un agente de nivel 5 sólo aprende las reglas específicas de una tarea y adaptar su comportamiento consecuentemente, un agente de nivel 6 es capaz de usar las emociones complejas para aprender lecciones básicas que se pueden generalizar a todo su comportamiento, independientemente de la tarea que se realice (generalización de lecciones aprendidas). Por ejemplo, desarrollar un miedo a manejar cargas de más de 500 kilogramos podría hacer que un hipotético robot de carga de nivel 6 evitase tanto las tareas conocidas como nuevas tareas que involucren el levantamiento y transporte de ese tipo de cargas. Es decir, el agente habría aprendido la lección de que en general (él) no es capaz de manejar cargas tan pesadas.

5.3.10.2 Arquitectura Característica

La arquitectura abstracta correspondiente al nivel 6 de *ConsScale* se caracteriza por la suma del componente **SsA** (mecanismo de auto-evaluación del propio estado) a la arquitectura definida para el nivel anterior (ver Figura 44).

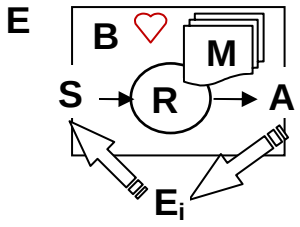


Figura 44. Arquitectura característica de un agente de nivel *ConsScale* 6.

5.3.10.3 Habilidades Cognitivas

En el nivel 6 de *ConsScale* se definen las siguientes habilidades cognitivas de carácter genérico:

- **$CS_{6,1}$**: Evaluación del estado propio (emociones globales).
- **$CS_{6,2}$**: Las emociones globales causan efectos en el cuerpo del agente.
- **$CS_{6,3}$**: Representación de los efectos de las emociones en el organismo (sentimientos).
- **$CS_{6,4}$**: Capacidad para mantener un mapa preciso y actualizado del esquema corporal.
- **$CS_{6,5}$**: Aprendizaje abstracto (generalización de lecciones aprendidas).
- **$CS_{6,6}$**: Capacidad para representar un flujo de perceptos integrados que incluyan el estado propio.

5.3.11 Nivel 7 (Autoconsciente o "*Self-Conscious*")

5.3.11.1 Descripción

La autoconciencia se adquiere cuando el agente es capaz de desarrollar el estadio 2 de desarrollo de la TdM: "yo sé que yo sé". La presencia de un modelo explícito del *yo* en el agente hace posible el auto-reconocimiento. De hecho, los mecanismos de aprendizaje ahora pueden operar en el dominio del propio futuro anticipado. El agente puede planear acerca de sí mismo (ya que el propio agente puede ser parte del plan) y luego aprender si el plan fue eficiente para él o no. Aprender a usar herramientas es una capacidad cognitiva característica de este nivel, ya que ser actor en el plan es un factor clave para el uso de herramientas (Sasaki et al. 2008). Los humanos de 18 meses de edad son una analogía biológica plausible para este nivel.

Un agente de nivel 7 es capaz de desarrollar pensamientos de orden superior (Rosenthal 2005), es decir, pensamientos sobre pensamientos (ver Apartado 2.1.4.5), y más específicamente pensamientos sobre uno mismo (Metzinger 2003) (ver Apartado 2.1.5.1). Consecuentemente, los agentes de nivel 7 se encuentran en el estadio 2 del desarrollo de la TdM: "yo sé que yo sé". Esto requiere la presencia de un modelo explícito del *yo*, que a su vez permite la planificación avanzada incluyendo al propio yo en los planes.

En el nivel 7 los mecanismos de aprendizaje operan también en el dominio del futuro anticipado. El agente puede realizar planes sobre sí mismo, y después de ejecutarlos, evaluar si el plan confeccionado fue beneficioso para el agente o no. Es más, un agente de nivel 7 tiene capacidad de imaginación. Es decir, puede planificar en base a los resultados de simulaciones internas, que son capaces de predecir los resultados de las posibles acciones del agente.

Al existir un símbolo explícito para el *yo*, el agente de nivel 7 es capaz de reconocerse a sí mismo. Por lo tanto, el test de comportamiento característico para este

nivel sería la prueba del espejo (ver Apartado 2.3.2). El comportamiento de los agentes de este nivel también se caracteriza por la capacidad de usar herramientas.

5.3.11.2 Arquitectura Característica

La arquitectura correspondiente al nivel 7 de *ConsScale* se caracteriza por la adición del componente **I** (mecanismo para la representación del *yo*) a la arquitectura definida en el nivel anterior (ver Figura 45).

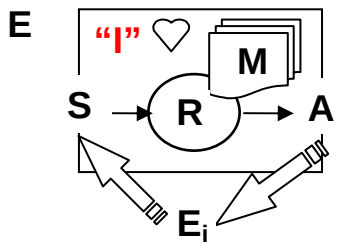


Figura 45. Arquitectura característica de un agente de nivel *ConsScale* 7.

5.3.11.3 Habilidades Cognitivas

En el nivel 7 de *ConsScale* se definen las siguientes habilidades cognitivas de carácter genérico:

- **$CS_{7,1}$**: Representación de la relación entre el *yo* y la percepción.
- **$CS_{7,2}$**: Representación de la relación entre el *yo* y la acción.
- **$CS_{7,3}$**: Representación de la relación entre el *yo* y los sentimientos.
- **$CS_{7,4}$**: Capacidad de auto-reconocimiento.
- **$CS_{7,5}$**: Planificación avanzada incluyendo al *yo* como actor en los planes.
- **$CS_{7,6}$**: Uso de estados imaginados en la planificación (imaginación) (Aleksander, Dunmall 2003).
- **$CS_{7,7}$**: Aprendizaje del uso de herramientas.
- **$CS_{7,8}$**: Capacidad para representar y auto-comunicarse el contenido mental explícito (flujo continuo interno de perceptos – imágenes mentales internas).

5.3.12 Nivel 8 (Empático o “Empathic”)

5.3.12.1 Descripción

La intersubjetividad es la característica principal de este nivel, en el que el agente no sólo está dotado de un modelo interno mejorado que incluye el *yo*, sino que también tiene la habilidad de modelar a otros como *yos* intencionales. El desarrollo del estadio 3 de TdM, “yo sé que tú sabes”, hace posible los comportamientos sociales. Los chimpancés son una analogía biológica ilustrativa para este nivel.

En el nivel 8 las representaciones internas del agente se enriquecen al contemplarse la intersubjetividad. Además del modelo del *yo*, característico del nivel anterior, el individuo mantiene de forma análoga modelos actualizados de otros individuos (*otros*). Es decir, el agente asigna a otros individuos un modelo de subjetividad (se aplica un modelo del *yo* también a otros individuos). Esta capacidad es la base de la interacción social compleja.

Tanto la intersubjetividad como la capacidad de mantener un esquema corporal actualizado (la cual estaba presente ya en el nivel 6 de *ConsScale*), son necesarias para el aprendizaje del uso de herramientas mediante imitación e incluso para la fabricación de nuevas herramientas.

Los agentes de nivel 8 también se caracterizan por ser capaces de colaborar con otros agentes en la persecución de una meta común. Este tipo de agentes pueden desarrollar planes que tienen en cuenta la dimensión social. Es decir, la relación existente entre el modelo del *yo* y los modelos de otros individuos subjetivos. La necesidad de estas habilidades cognitivas se ha puesto de manifiesto en el diseño de agentes BDI que sean capaces de colaborar entre ellos y/o con humanos (Rao, Georgeff & Sonenberg 1992).

5.3.12.2 Arquitectura Característica

La arquitectura del nivel 8 de *ConsScale* se caracteriza por la existencia de un componente **O** (mecanismo para la representación de otros *yos*) que se suma a los componentes ya existentes en el nivel anterior (ver Figura 46).

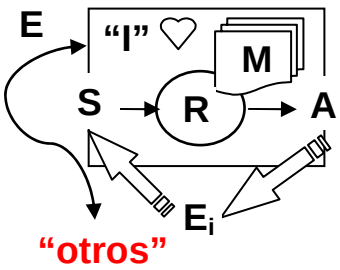


Figura 46. Arquitectura característica de un agente de nivel *ConsScale* 8.

5.3.12.3 Habilidades Cognitivas

En el nivel 8 de *ConsScale* se definen las siguientes habilidades cognitivas de carácter genérico:

- **$CS_{8,1}$**: Habilidad para modelar a otros individuos como yos subjetivos e intencionales.
- **$CS_{8,2}$**: Aprendizaje por imitación de un individuo homólogo.
- **$CS_{8,3}$**: Habilidad para colaborar con otros individuos en la persecución de una meta común.
- **$CS_{8,4}$**: Planificación social (utilizando planes sociales, que tienen en cuenta a los otros individuos participantes en el plan).
- **$CS_{8,5}$**: Habilidad para crear nuevas herramientas.
- **$CS_{8,6}$**: Las imágenes mentales internas están enriquecidas con contenidos relacionados con los modelos de otros y la relación entre el yo y otros yos.

5.3.13 Nivel 9 (Social o "*Social*")

5.3.13.1 Descripción

En este nivel el modelo interno de otros *yos* se perfecciona gracias al desarrollo total de la capacidad de TdM, "yo sé que tú sabes que yo sé". Esto significa que el comportamiento característico de este nivel viene definido por el desarrollo de estrategias Maquiavélicas sofisticadas (o inteligencia social), entre las que se incluyen comportamientos sociales como la mentira, la astucia o el liderazgo. Además, los agentes de nivel 9 cuentan con capacidades lingüísticas y comunicación precisa. Los agentes de este nivel serían capaces de desarrollar una cultura propia. Los humanos de 4 años de edad son la analogía biológica para este nivel.

En el nivel 9, la TdM está totalmente desarrollada, por lo tanto los agentes están fuertemente influidos por su entorno social. Además, el desarrollo de una cultura proporciona nuevas posibilidades de aprendizaje. Un agente social A (en términos de la escala *ConsScale*) sería consciente de que otro agente B podría conocer las creencias, los deseos y las intenciones de A. Este tipo de agentes también se caracterizan por sus capacidades avanzadas de comunicación, que junto con el desarrollo de la TdM, les hace capaces de, por ejemplo, contar mentiras intencionadamente. Existen modelos matemáticos de la dinámica de la inteligencia Maquiavélica que se podrían usar potencialmente para descubrir este tipo de comportamientos en agentes artificiales (Gavrilets, Vose 2006).

5.3.13.2 Arquitectura Característica

La arquitectura del nivel 9 de *ConsScale* se caracteriza por el componente **AR** (mecanismo de comunicación precisa) que se integra con los componentes ya existente en el nivel anterior (ver Figura 47).

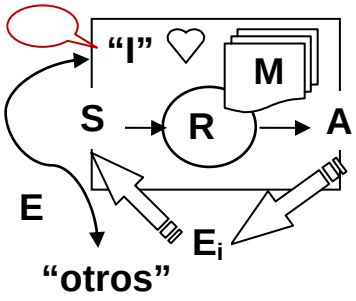


Figura 47. Arquitectura característica de un agente de nivel *ConsScale* 9.

5.3.13.3 Habilidades Cognitivas

En el nivel 9 de *ConsScale* se definen las siguientes habilidades cognitivas de carácter genérico:

- **$CS_{9,1}$**: Habilidad para desarrollar estrategias Maquiavélicas, como la mentira o la astucia.
- **$CS_{9,2}$**: Aprendizaje social (aprendizaje de nuevas estrategias Maquiavélicas).
- **$CS_{9,3}$**: Habilidades avanzadas de comunicación (comunicación precisa del contenido mental).
- **$CS_{9,4}$**: Los grupos son capaces de desarrollar una cultura.
- **$CS_{9,5}$**: Habilidad para modificar y adaptar el entorno a las necesidades del agente.

5.3.14 Nivel 10 (Androide o "*Human-Like*")

5.3.14.1 Descripción

Tal y como indica el nombre de este nivel, la analogía biológica correspondiente es el ser humano adulto. Las características de este nivel son la capacidad para formar una cultura compleja y la comunicación verbal precisa. Esto implica la capacidad de uso de herramientas externas complejas para el aprendizaje. La fluidez entre la inteligencia social y la inteligencia técnica permite la extensión del conocimiento usando medios externos (como la comunicación escrita). Los avances tecnológicos son posibles en este

nivel. Los agentes de nivel 10, al igual que los humanos, son capaces de modificar su entorno de forma extrema.

El nivel 10 representa a la clase de agentes dotados con un nivel de conciencia equiparable al humano. Esto implica que los grupos de este tipo de agentes pueden formar una cultura compleja o formar parte de la cultura humana. El Test de Turing, en cualquiera de sus formas o variantes, es una prueba obvia para los agentes de este nivel.

5.3.14.2 Arquitectura Característica

La arquitectura del nivel 10 de *ConsScale* se caracteriza por la inclusión del componente **AVR** (mecanismo de comunicación precisa) que se integra con el resto de componentes ya definidos para el nivel anterior (ver Figura 48).

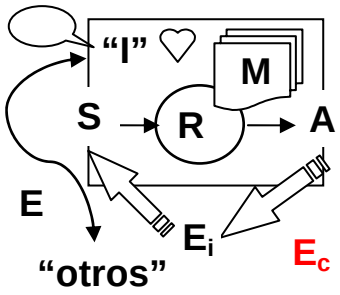


Figura 48. Arquitectura característica de un agente de nivel *ConsScale* 10.

5.3.14.3 Habilidades Cognitivas

En el nivel 10 de *ConsScale* se definen las siguientes habilidades cognitivas de carácter genérico:

- **$CS_{10,1}$**: Reporte verbal preciso. Capacidades lingüísticas avanzadas.
- **$CS_{10,2}$**: Habilidad para pasar el Test de Turing.
- **$CS_{10,3}$**: Los grupos son capaces de desarrollar una civilización y cultura y tecnología avanzadas.

5.3.15 Nivel 11 (Super-Consciente o "*Super-Conscious*")

5.3.15.1 Descripción

Este último nivel está caracterizado por la habilidad de sincronizar y coordinar varios flujos de conciencia en el mismo *yo* físico. No hay ejemplos conocidos de esta habilidad en el mundo biológico.

Un agente super-consciente es aquel capaz de manejar internamente varios flujos de conciencia, coordinando al mismo tiempo un único cuerpo y su correspondiente sistema de atención concurrente. Los agentes de este tipo requieren un mecanismo de coordinación entre los distintos flujos de conciencia y un mecanismo de acceso sincronizado a los recursos físicos. Un agente de este tipo sería capaz, por ejemplo, de mantener varias conversaciones conscientes simultáneas utilizando diferentes líneas de comunicación. Además, podría usar la información obtenida de una conversación en cualquiera de las otras de forma prácticamente inmediata.

5.3.15.2 Arquitectura Característica

La arquitectura del nivel 11 de *ConsScale* se caracteriza por la inclusión del componente $\mathbf{R^n}$ (mecanismo para gestionar múltiples flujos de conciencia) en la arquitectura abstracta definida para el nivel anterior (ver Figura 49).

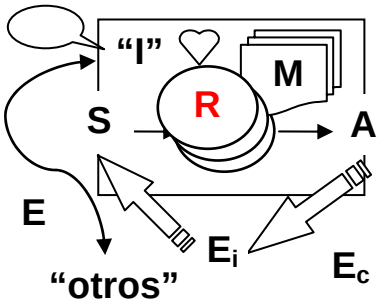


Figura 49. Arquitectura característica de un agente de nivel *ConsScale* 11.

5.3.15.3 Habilidades Cognitivas

En el nivel 11 de *ConsScale* se define una sola habilidad cognitiva de carácter genérico:

- $\mathbf{CS_{11,1}}$: Habilidad para gestionar varios flujos de conciencia.

5.4 CQS. El Índice Cuantitativo de Conciencia

Además de la medida cualitativa proporcionada en la escala *ConsScale* mediante los trece niveles conceptuales de conciencia descritos anteriormente, se ha definido una medida cuantitativa llamada CQS (*ConsScale Quantitative Score*). Tener una medida cuantitativa asociada a la escala es beneficioso desde los siguientes puntos de vista: por un lado, diferentes implementaciones de Conciencia Artificial se pueden evaluar con los mismos criterios independientemente de su dominio de aplicación, permitiendo de esta forma los estudios comparativos. Por otro lado, se podrían definir funciones de adecuación (*fitness*) que permitan aplicar enfoques basados en computación evolutiva al diseño de modelos de Conciencia Artificial.

El índice CQS está diseñado para cubrir todas las posibilidades que puedan existir en una implementación de Conciencia Artificial, es decir, contemplar todas las posibles combinaciones de cumplimiento de las capacidades cognitivas ($CS_{i,j}$) definidas. Al mismo tiempo, el índice CQS también mantiene las restricciones originales de la escala conceptual estructurada en niveles. Es decir, este índice es capaz de proporcionar un valor numérico para cada nivel conceptual, permitiendo también asignar valores intermedios a las implementaciones que no se corresponden exactamente con uno de los niveles canónicos definidos. Esto permite evaluar de forma precisa el grado de desarrollo cognitivo de cualquier implementación que se evalúe.

5.4.1 El cálculo del índice CQS

El índice CQS se calcula en tres pasos. En primer lugar, ha de calcularse una puntuación específica para cada nivel de la escala (esta puntuación se denomina L_i, siendo *i* el número del nivel de *ConsScale* correspondiente en cada caso). Básicamente, los índices L_i representan el grado de cumplimiento que la implementación que se está analizando presenta con respecto al nivel *i*. Suponiendo que una implementación cumple con todos los requisitos de arquitectura de un nivel, hay que considerar el número de habilidades cognitivas ($CS_{i,j}$) presentes en el agente. Aunque el número total de habilidades cognitivas consideradas para cada nivel es diferente, el cálculo de L_i se hace de forma equivalente para todos los niveles. Es decir, L_i tiene que proporcionar una medida del grado de cumplimiento del nivel i, independientemente del número de capacidades cognitivas definidas para cada nivel. Como se ha mencionado anteriormente (en los apartados 5.3.3, 5.3.4 y 5.3.5), los tres primeros niveles de la escala (-1, 0 y 1) son simplemente niveles de referencia y no tienen asociada ninguna habilidad cognitiva. Por lo tanto L_{-1}, L_0 y L_1 son siempre cero y no hay necesidad de realizar ningún cálculo para obtener estos tres primeros valores. Para el resto de los niveles (del nivel 2 al 11) ha de resolverse la Ecuación 3.

$$L_i = \begin{cases} 0 & si\ ncsf\ es\ 0 \\ \dfrac{\left(ncsf + (J - J_i)\right)^3}{10^3} & si\ ncsf\ no\ es\ 0 \end{cases}$$

Ecuación 3. Cálculo del índice L_i.

Donde **ncsf** (*number of cognitive skills fulfilled*) es el número de habilidades cognitivas ($CS_{i,j}$) que el agente tiene implementadas de forma efectiva. ***J*** es el número máximo de habilidades cognitivas consideradas para cualquier nivel, y $\boldsymbol{J_i}$ es el número total de habilidades cognitivas definidas para el nivel *i*. El índice L_i se ha diseñado como una curva exponencial con el objetivo de proporcional un valor significativo correlacionado con la sinergia existente entre las habilidades cognitivas de un mismo nivel.

Los valores posibles de L_i varían entre 0 y 1. Un valor nulo significa que el agente no cumple ningún requisito cognitivo del nivel *i*; el valor máximo, 1, significa que el agente cumple completamente los requisitos cognitivos correspondientes al nivel *i*. La

Figura 50 representa dos curvas con los valores posibles de L_i en función del número de habilidades cognitivas presentes en supuestos niveles con 6 y 10 habilidades cognitivas definidas respectivamente. En ambos gráficos, el eje X se corresponde con el número de habilidades cognitivas presentes en el agente (*ncfs*) y el eje Y es el valor correspondiente de L_i, que va desde 0 hasta 1.

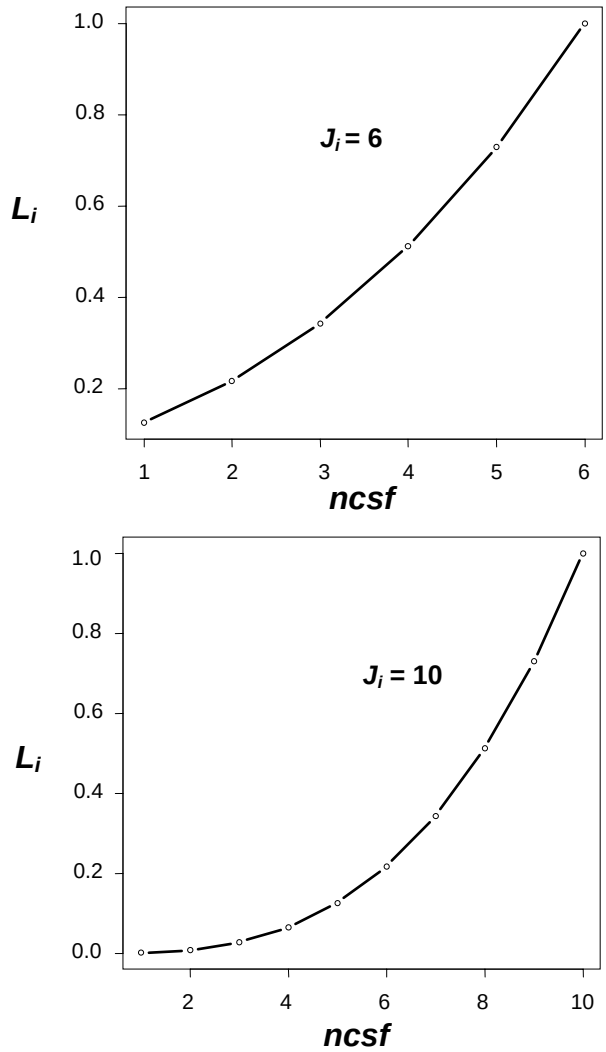


Figura 50. Representación gráfica del índice L_i para J_i=6 y J_i=10.

Una vez calculados los valores L_i para todos los niveles de la escala, se puede obtener un índice acumulado denominado CLS (*Cumulative Level Score*) utilizando la Ecuación 4.

$$CLS = \sum_{i=2}^{11} \left(\frac{L_i}{i-1} \right)^2$$

Ecuación 4. Cálculo del índice CLS.

El índice CLS combina las puntuaciones de todos los niveles en una medida única que sigue una progresión logarítmica. La Figura 51 representa los posibles valores del índice acumulado CLS para agentes que cumplan con los niveles 1 a 11. El eje X representa los niveles de *ConsScale* y el eje Y el valor de CLS asociado. Los posibles valores del índice CLS van de 0 a 1.55.

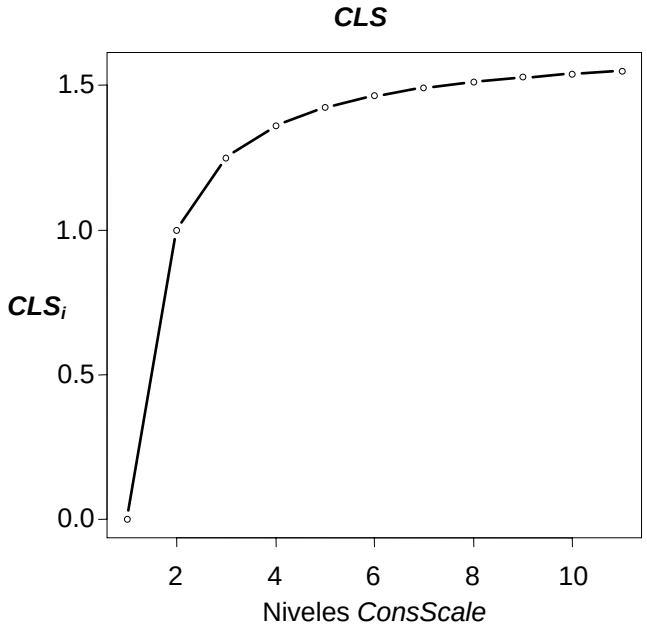


Figura 51. Posibles valores del índice CLS.

El diseño logarítmico del índice acumulado CLS impide que el significado de escala se distorsione debido al efecto combinado de bajas puntuaciones L_i para niveles inferiores y altas puntuaciones L_i en niveles superiores. Es decir, el CLS asegura que la medida cuantitativa siga el mismo sesgo que se establece para la medida cualitativa. Es decir, se penaliza a aquellas implementaciones que satisfaciendo un nivel dado no cumplen con todos los requisitos de los niveles inferiores; favoreciendo a aquellas otras implementaciones, que aun no mostrando habilidades de niveles tan altos, sí que siguen de forma completa la progresión de desarrollo cognitivo definida en *ConsScale*.

En cualquier caso, un agente sólo puede ser considerado como nivel *i* ***si y solo si*** también cumple con todos los niveles inferiores. Por ejemplo, si un agente llamado A cumple con los requisitos de los niveles 2, 3, 5, 6 y 7, pero no cumple completamente los requisitos del nivel 4, su nivel *ConsScale* nominal será el 3 (aunque su CQS reflejará que tiene un desarrollo cognitivo mayor al de un agente básico de nivel 3, que no tendría ninguna habilidad correspondiente a niveles superiores). En cambio, un agente B que cumpla completamente los requisitos de los niveles 2, 3 y 4, aunque no puntúe nada en niveles superiores, sería considerado como nivel 4 de acuerdo a la definición de los niveles de *ConsScale*.

Finalmente, la obtención del índice CQS se realiza calculando la función exponencial descrita en la Ecuación 5.

$$CQS = \frac{e^{(CLS^5/K)} + a}{10}$$

Ecuación 5. Cálculo del índice CQS.

Donde **K** y **a** son constantes definidas específicamente para normalizar los valores de puntuación en el rango de 0 a 1000. Consecuentemente, la puntuación mínima es 0 (que corresponde a los niveles -1, 0 y 1 de *ConsScale*) y el máximo nivel de conciencia se representa con una puntuación de 1000 en el índice CQS (que se corresponde con el nivel 11 de *ConsScale*).

La Figura 52 muestra una representación gráfica de los valores que el índice CQS puede tomar para los agentes clasificados canónicamente en los niveles 1 a 11. El resto de los agentes obtendrían una puntuación determinada en la curva exponencial CQS entre dos niveles de referencia consecutivos.

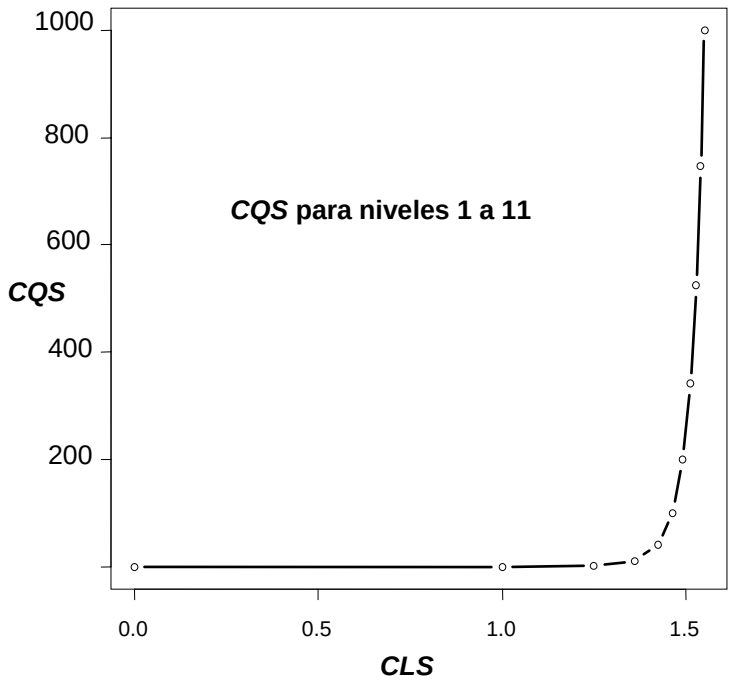


Figura 52. Posibles valores del índice CQS.

Las constantes K y a (0.97062765 y -1 respectivamente) se calculan resolviendo el sistema especificado en la Ecuación 6.

$$\left\{ \begin{array}{l} \dfrac{e^{0/K} + a}{10} = 0 \\ \dfrac{e^{c^5/K} + a}{10} = 1000 \end{array} \right\}$$

Ecuación 6. Cálculo de constantes para la normalización del índice CQS.

Donde **c** es el valor máximo que CLS puede tomar (1.549768).

Como parte de las herramientas asociadas a *ConsScale* se ha desarrollado una calculadora que permite el cálculo automático del índice CQS. La calculadora *ConsScale* está disponible en versión Web y puede consultarse en la dirección www.consscale.com (ver Figura 53). El usuario de la calculadora *ConsScale* sólo necesita introducir los componentes arquitecturales de un agente y las capacidades cognitivas que éste cumple para obtener la puntuación CQS asociada.

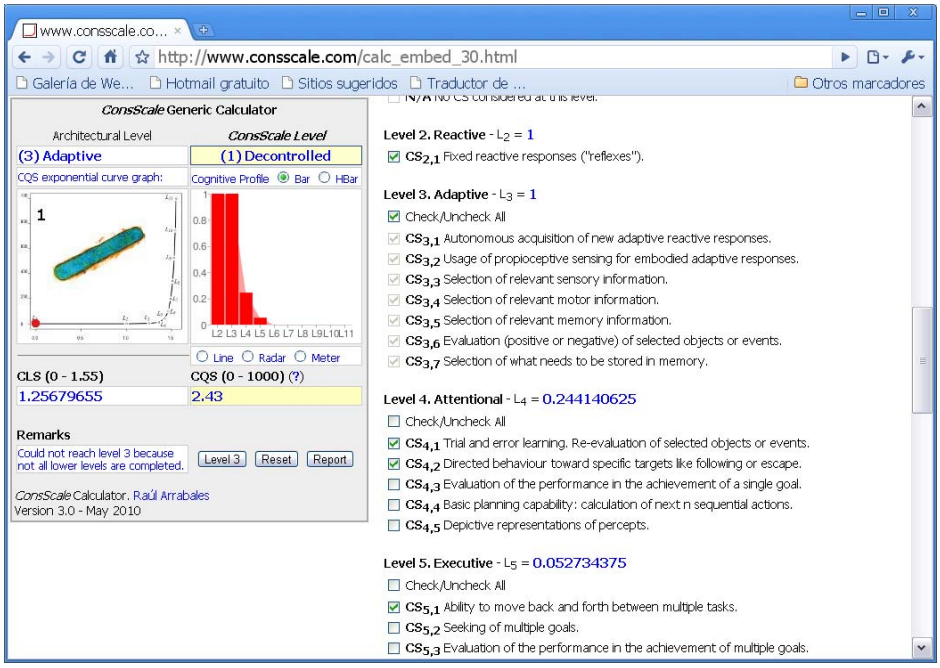


Figura 53. Aplicación Web para calcular el índice CQS.

5.4.2 Significado del índice CQS

La curva CQS representada en la Figura 52 abarca un rango continuo de valores desde 0 hasta 1000 de forma exponencial, representando la sinergia acumulada mediante la adición e integración de capacidades cognitivas a lo largo de los diferentes niveles de la escala. Según se incrementa la capacidad cognitiva global, la puntuación correspondiente al índice CQS también se incrementa exponencialmente, llegando a alcanzar valores significativos a partir de los niveles 4 y 5 (ver Tabla 6).

Tabla 6. Relación entre valores de CQS y niveles conceptuales de ConsScale.

Nivel	*Descripción*	*CQS*
1	*Pre-Funcional*	0.00
2	*Reactivo*	0.18
3	*Adaptativo*	2.22
4	*Atencional*	12.21
5	*Ejecutivo*	41.23
6	*Emocional*	101.08
7	*Autoconsciente*	200.03
8	*Empático*	341.45
9	*Social*	524.54
10	*Androide*	745.74
11	*Super-Consciente*	1000.00

Cuando los niveles se consideran como propiedades discretas, el nivel cualitativo de un agente se puede deducir a partir de su puntuación CQS. Por ejemplo, un CQS de 2.83 indica que una implementación se ha clasificado como nivel 3 (porque el valor de su CQS se encuentra entre el valor mínimo para el nivel 3 y el mínimo para el nivel 4). Sin embargo, dado que su puntuación es mayor que el CQS canónico para el nivel 3 (2.22), se puede inferir que el agente en cuestión también cumple algunos otros requisitos de niveles superiores.

En definitiva, aunque todas las capacidades cognitivas presentes en el agente cuentan, el índice CQS premia especialmente aquellas implementaciones que siguen el camino para el desarrollo de la conciencia especificado en *ConsScale*. Este aspecto del diseño del índice CQS hace que se dé más valor al desarrollo de implementaciones de inspiración biológica frente a otras posibles trayectorias de desarrollo de máquinas conscientes. La caracterización del índice basada en la hoja de ruta particular propuesta en *ConsScale* es útil, ya que no se considera una combinación arbitraria de diferentes habilidades cognitivas, sino que se plantea un enfoque evolutivo en el que las sinergias entre los componentes del agente se van incrementando según se realizan implementaciones más complejas (ver Apartado 5.3.2).

La medida CQS es consistente con la concepción de los niveles diseñada en *ConsScale*. Por ejemplo, para un agente situado es mejor tener capacidad de adquisición de información más adaptación más atención, en vez de tener sólo capacidades de cambio de contexto y planificación, pero sin tener las anteriores (en el caso de que tal implementación fuera posible de alguna forma).

5.5 Representación Gráfica del Nivel de Conciencia en ConsScale

Además de los niveles conceptuales definidos en la escala y el índice CQS, también se define como parte de *ConsScale* una herramienta adicional para la caracterización del desarrollo cognitivo de un agente. Se trata de una representación gráfica que puede tomar la forma de diagrama en estrella o diagrama de barras y que permite visualizar de forma rápida y precisa el desarrollo de un agente en cada uno de los aspectos definidos en los niveles de *ConsScale*. Aunque una medida cuantitativa simple como el índice CQS es útil para una evaluación rápida y sucinta, carece de capacidades de representación detallada. Por esta razón se propone el uso complementario de una representación gráfica de los perfiles cognitivos (RGPC).

Para poder representar el perfil cognitivo de un agente en términos de la escala *ConsScale*, es necesario considerar los valores particulares de los índices L_i. Tanto el índice CLS como el CQS son parámetros unidimensionales, calculados a partir del vector multidimensional L_i ($i \in \{2\text{-}11\}$). Por lo tanto, para preservar la riqueza multidimensional derivada de los niveles de *ConsScale*, se usan los índices L_i como base de la representación gráfica.

En aras de una mayor claridad, los niveles -1 (sin cuerpo definido), 0 (aislado) y 1 (pre-funcional) se han excluido de la representación gráfica propuesta. También se puede excluir el nivel 11, cuando se pretende colocar el nivel de conciencia humano como el máximo posible. Inicialmente, se ha decidido emplear diagramas de radar (también conocidos como diagramas de estrella) ya que son particularmente adecuados para realizar una representación compacta y significativa de los valores L_i (ver Figura 54). Los diagramas de radar representan en cada eje el grado de cumplimiento de cada nivel de *ConsScale*. Los posibles valores de cada eje L_i van de 0.0 (cuando no se satisface ningún $CS_{i,j}$) hasta 1.0 (cuando se satisfacen todos los $CS_{i,j}$). La Figura 54 representa un diagrama de radar vacío, donde todos los L_i son 0.

Para las tareas de análisis comparativo de diversos modelos se ha considerado el uso de diagramas de barras horizontales (ver Figura 55), ya que permiten una comparación más directa de varios perfiles cognitivos (como se ilustra en el Apartado 7.5.2). La naturaleza jerárquica de la escala queda bien representada usando un diagrama de barras horizontales en el que los niveles inferiores se colocan abajo y los niveles superiores arriba. Análogamente al diagrama en estrella, en este diagrama cada barra representa el nivel de cumplimiento del correspondiente nivel de *ConsScale*. A continuación se muestran unos ejemplos básicos de la representación gráfica propuesta.

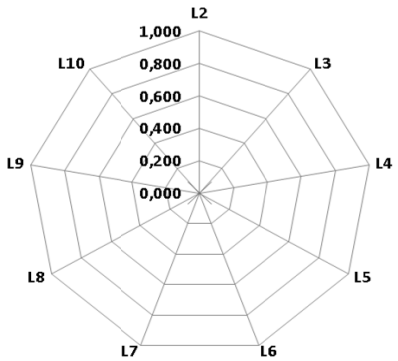


Figura 54. Representación gráfica en estrella de un perfil cognitivo vacío.

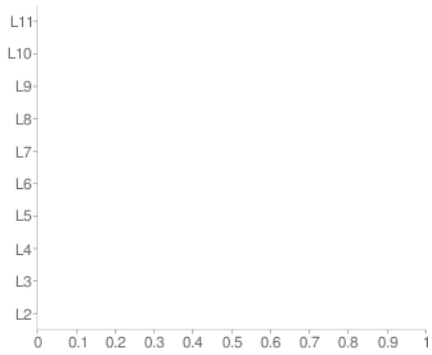


Figura 55. Diagrama de barras correspondiente a un perfil cognitivo vacío.

La Figura 56 representa una implementación de nivel 3 (adaptativa). Es interesante remarcar que aunque este agente tiene características de niveles más altos, sólo se puede considerar como nivel 3. Sin embargo, su CQS (3.77) es más alto que el de una criatura de nivel 3 puro (que sería 2.22). Comparando la RGPC con el índice CQS, se puede observar que la representación gráfica ofrece información adicional acerca de los niveles en los que el agente consigue mejores puntuaciones. Aunque los diagramas en estrella y los diagramas de barras representan la misma información, puede apreciarse que los diagramas de barras reflejan mejor la relación jerárquica de los niveles. Tanto la Figura 56 como la Figura 57 representan al mismo agente. Sin embargo, es más fácil apreciar la diferencia real entre niveles no adyacentes en el caso del diagrama de barras (Figura 57).

Figura 56. Representación en estrella del perfil cognitivo de un agente de nivel 3.

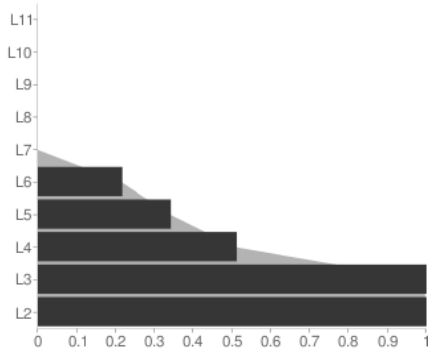


Figura 57. Representación con barras del perfil cognitivo de un agente de nivel 3.

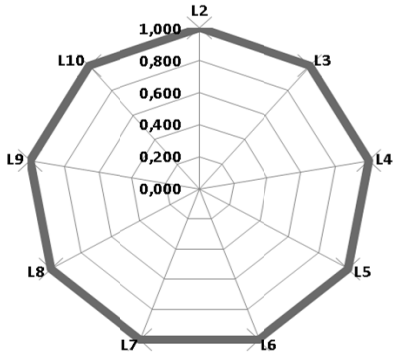


Figura 58. Representación en estrella del perfil cognitivo de un agente de nivel 10.

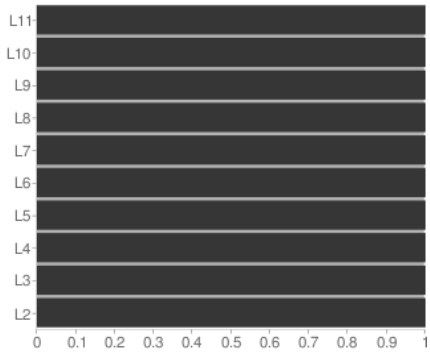


Figura 59. Representación en barras del perfil cognitivo de un agente de nivel 10.

La Figura 58 representa una criatura de nivel 11 (super-consciente). Como los niveles de 2 a 11 tienen un valor 1 en su coeficiente L_i, el CQS correspondiente es de 1000. La Figura 59 representa a la misma criatura usando un diagrama de barras horizontales.

5.6 ConsScale como una Hoja de Ruta

La definición de esta escala en el marco de la investigación en Conciencia Artificial no es útil sólo para la evaluación y el estudio comparativo, sino que también se puede ver como una propuesta en cuanto al camino a seguir para la construcción de máquinas conscientes. *ConsScale* se puede considerar como una posible hoja de ruta con hitos identificables como retos ingenieriles. Enfrentarse al reto de la comprensión y desarrollo de la conciencia desde la perspectiva de la ingeniería implica definir un plan concreto. La mayoría de los esfuerzos que se están realizando actualmente se centran en aspectos particulares de la conciencia (ver Apartado 2.2), sin embargo, el modo en el que las actuales líneas de investigación deberían converger no está claro.

El hecho de tomar organismos biológicos como inspiración ha sido un enfoque obvio durante décadas. Pero esta inspiración toma a menudo la forma de funciones o mecanismos concretos, como la atención o las emociones, perdiendo la perspectiva global de un sistema cognitivo complejo. Una de las propuestas básicas de esta tesis doctoral es considerar la conciencia como el integrador que da lugar a la unidad de una mente. Considerar las habilidades cognitivas específicas como componentes

individuales o independientes no ayudará a largo plazo en el diseño de máquinas equiparables a los humanos. La idea planteada en esta tesis es redefinir la conciencia (al menos desde el punto de vista de la ingeniería) como una super-función, la cual está compuesta de diversas capacidades cognitivas integradas.

En este contexto, *ConsScale* sugiere un camino concreto para el desarrollo progresivo de agentes artificiales conscientes. Se podrían definir otras hojas de ruta diferentes, pero viendo como la conciencia ha evolucionado en la naturaleza, parece que el enfoque propuesto es razonable. Por ejemplo, tiene sentido construir primero máquinas empáticas (nivel 8) con el objetivo de poder construir más tarde agentes sociales (nivel 9).

Actualmente no es posible saber si la hoja de ruta propuesta en *ConsScale* es una biena apuesta. Por supuesto, la escala está abierta a mejoras, que se pueden ir incorporando a medida que se asimile y se consensue la retroalimentación obtenida por parte de la comunidad científica dedicada al estudio de la conciencia. Por el momento, *ConsScale* proporciona un enfoque plausible para la adición incremental de capacidades cognitivas en un sistema artificial. Adicionalmente, la definición y evolución de este tipo de métricas orientadas a la conciencia ha contribuido a determinar el estado del arte y a definir objetivos concretos en el campo de la Conciencia Artificial.

5.7 Aplicación de la Escala

El hecho de establecer un marco para la clasificación y evaluación del nivel de conciencia de un sujeto es de vital importancia en el estudio científico de la conciencia. Como se ha explicado, la propuesta presentada en esta tesis doctoral es un marco para la evaluación de agentes de acuerdo a las características asociadas con la conciencia que estos presentan. Por lo tanto, la escala se puede usar para realizar un estudio comparativo de diferentes implementaciones de conciencia artificial y ver cuáles se parecen más al único ejemplo de conciencia superior que conocemos en la actualidad: el ser humano.

Como se ha explicado anteriormente, los niveles de *ConsScale* están caracterizados por componentes arquitectónicos abstractos y por las capacidades cognitivas que estos generan. Por lo tanto, para poder determinar el nivel de conciencia de una implementación hay que comprobar la presencia tanto de los componentes arquitectónicos como de las capacidades cognitivas.

La detección de los componentes software o hardware se puede hacer mediante la inspección interna del sistema. La presencia de las habilidades cognitivas ha de comprobarse mediante la ejecución de pruebas asociadas. Para poder aplicar la escala a una implementación de conciencia artificial concreta, han de definirse pruebas específicas para cada una de las capacidades cognitivas $CS_{i,j}$. Es decir, mientras que la escala está definida como un marco genérico, la aplicación de la misma es en realidad específica para un dominio, por lo que deben establecerse pruebas específicas de dominio para poder evaluar un agente dado. En otras palabras, ha de usarse una instanciación de la escala para un dominio de problema determinado para poder evaluar un agente de forma práctica.

Al utilizar *ConsScale* para evaluar el nivel de desarrollo cognitivo de un agente se cuenta con las dos herramientas descritas anteriormente: el índice CQS y la

representación gráfica del perfil cognitivo. Ambas herramientas permiten, además de representar el grado de cumplimiento de cada nivel, apreciar el grado de desarrollo cognitivo global del agente. Es decir, combinan los índices L_i o bien en una representación gráfica (RGPC) o bien mediante una fórmula matemática (CQS).

Aunque siempre se usan las herramientas mencionadas anteriormente, se han definido dos métodos diferenciados para realizar la evaluación o caracterización del desarrollo cognitivo de un agente:

- **Proceso Estándar de Evaluación (PEE).** Este proceso está pensado para ser aplicado en implementaciones existentes de agentes (y no en modelos teóricos). El PEE proporciona una medida precisa y confiable del nivel de desarrollo cognitivo de un agente. Sin embargo, se requiere un tiempo y un esfuerzo significativo para realizar la evaluación (ver Figura 60).
- **Proceso Simplificado de Evaluación (PSE)**. Este proceso de evaluación proporciona una aproximación rápida del nivel potencial de desarrollo cognitivo tanto de un agente ya implementado como de un modelo computacional que todavía no se ha implementado. La mayor ventaja de este procedimiento es que la evaluación se puede realizar rápidamente e implica poco esfuerzo (ver Figura 61).

La realización de un PEE requiere disponer del agente que se quiere evaluar y también una definición de pruebas adaptada al dominio de problema correspondiente. Como se ha mencionado anteriormente, la evaluación se basa en los componentes arquitecturales del agente y sus capacidades cognitivas. Los componentes arquitecturales se pueden identificar a través de la inspección interna de la implementación. Las habilidades cognitivas presentes en el agente se pueden evaluar gracias a la definición y ejecución de pruebas de comportamiento específicas adaptadas para el dominio de problema seleccionado.

Las habilidades cognitivas descritas para cada nivel de la escala *ConsScale* se refieren a capacidades genéricas. Por lo tanto, la lista de habilidades cognitivas resumidas en la Tabla 5 no se pueden usar directamente para realizar una evaluación estándar. Se requiere de un proceso de instanciación para poder aplicar la escala a un dominio de aplicación concreto. El proceso de instanciación consiste básicamente en diseñar pruebas conductuales, específicas para el dominio correspondiente, que permitan determinar la presencia de las habilidades cognitivas consideradas en cada nivel de la escala.

Desde el punto de vista de *ConsScale*, un dominio de problema específico se define en términos de:

- La información que los sensores del agente pueden adquirir (**objetos** o perceptos).
- Lo que los actuadores del agente pueden hacer (**acciones** o verbos).

En definitiva, se necesita definir una ontología del dominio sobre el que se quiere realizar la evaluación. Esta ontología se compone de objetos y acciones concretos. De esta forma, las habilidades cognitivas genéricas asociadas a cada nivel de *ConsScale* se pueden traducir en perfiles de comportamiento concretos, los cuales se pueden verificar mediante la observación en tercera persona.

Una vez que se ha determinado la lista de componentes arquitectónicos y de habilidades cognitivas se pueden aplicar las métricas de *ConsScale* para obtener el nivel cualitativo de conciencia, el perfil cognitivo (RGPC) y el índice CQS.

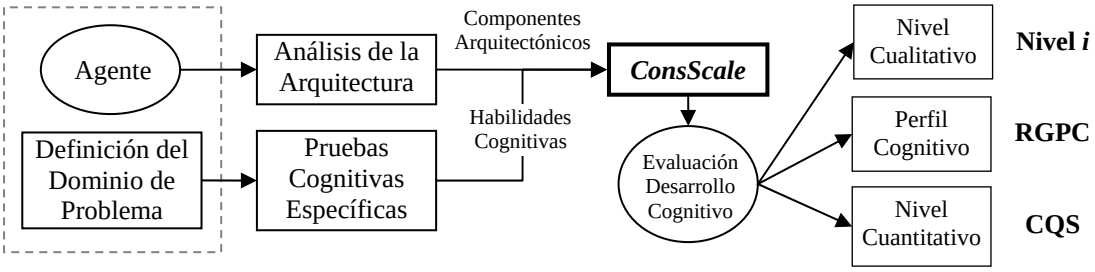


Figura 60. Proceso de Evaluación Estándar (PEE) de *ConsScale*.

Cabe destacar que es necesario desarrollar pruebas cognitivas correspondientes a cada una de las habilidades cognitivas cuya presencia se quiere comprobar. Estas pruebas han de definirse de tal manera que validen la inspiración evolutiva y de integración cognitiva en las que se basa la escala. Es decir, siempre que sea posible las pruebas cognitivas de más alto nivel requerirán la presencia y la integración efectiva de todas (o casi todas) las capacidades cognitivas de niveles inferiores. En caso de que el agente no cumpla estos requisitos, no pasará la prueba y por lo tanto se considerará que no cumple la habilidad cognitiva $CS_{i,j}$ correspondiente. En el Apartado 5.7.1 se describe la aplicación del PEE en el dominio de los agentes autónomos de un videojuego de acción en primera persona.

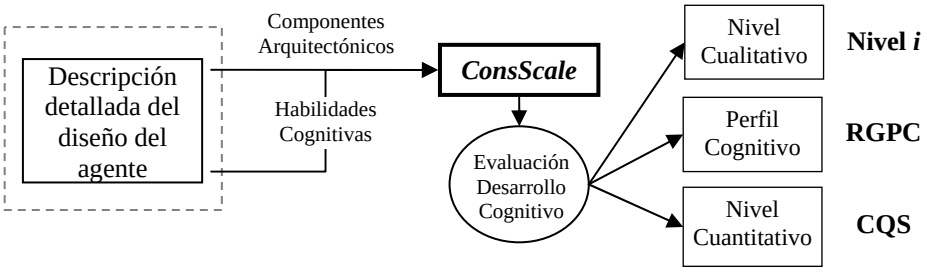


Figura 61. Proceso Simplificado de Evaluación (PSE) de *ConsScale*.

Para la aplicación del PSE se asume la presencia de los componentes arquitecturales y de las habilidades cognitivas simplemente analizando el diseño del sistema. Por lo tanto, no hay necesidad de realizar ninguna prueba ni de realizar una instanciación de la escala en un dominio de problema concreto. Naturalmente, la evaluación obtenida usando este procedimiento no es precisa y sólo puede considerarse como una aproximación (y probablemente muy optimista). Sin embargo, el PSE se puede usar como una forma de evaluar el potencial de un modelo de Conciencia Artificial en las fases tempranas de su diseño y/o implementación.

En el caso de la evaluación de agentes ya implementados, el PEE es el procedimiento que se debería seguir para obtener una medida precisa y realística del

nivel de conciencia del agente. Sin embargo, el PSE constituye una herramienta conceptual potente para evaluar el nivel de desarrollo potencial de una arquitectura cognitiva durante su diseño, incluso antes de que exista ninguna implementación asociada (la Tabla 7 proporciona una comparativa esquemática entre el PEE y el PSE).

Tabla 7. Comparación entre el PEE y el PSE.

	Evaluación Estándar (PEE)	**Evaluación Simplificada (PSE)**
Aplicabilidad	Sólo agentes ya implementados.	Modelos, arquitecturas, diseños, implementaciones.
Precisión	Alta (medida realista).	Baja (medida potencial, optimista).
Coste	Alto (inspección interna de la implementación, diseño y ejecución de pruebas cognitivas).	Bajo (los componentes arquitecturales y las habilidades cognitivas se infieren directamente).
Tiempo de cálculo del índice CQS	Del orden de milisegundos.	Del orden de milisegundos.
Dominio de Problema	Dependiente del dominio.	Independiente del dominio.
Recursos Necesarios	Entorno de pruebas adecuado, procedimientos de prueba y herramientas asociadas, adquisición de datos y herramientas de inspección.	Descripción detallada del sistema.
Salida	Nivel *ConsScale*, perfil cognitivo (RGPD) y medida cuantitativa (CQS).	Nivel *ConsScale*, perfil cognitivo (RGPD) y medida cuantitativa (CQS).

5.7.1 Aplicación de la escala en el dominio de los videojuegos

Con el objetivo de ilustrar el proceso de aplicación de la escala a un agente concreto, en la presente tesis se ha desarrollado un entorno de experimentación basado en el videojuego *Unreal Tournament*, comercializado por la compañía Epic Games Inc.(Epic Games). Concretamente, se estudiará el nivel de desarrollo cognitivo de varios agentes autónomos, o personajes sintéticos, que son capaces de participar en este videojuego de acción en primera persona (este tipo de juegos también son conocidos como FPS o *first-person shooter*). Usando esta plataforma se han evaluado diferentes agentes usando la escala *ConsScale*. Algunos de estos agentes se han desarrollado usando la arquitectura cognitiva también propuesta en esta tesis (ver Capítulo 4). Los resultados obtenidos se describen en el Apartado 7.4.4.

Con el objetivo de tener un dominio de problema bien definido se ha limitado el alcance del entorno de experimentación de la siguiente forma. Se ha considerado que las implementaciones que se van a evaluar están diseñadas para ser personajes sintéticos en un videojuego de tipo FPS (a los personajes sintéticos de los juegos FPS se los suele denominar *bots*). Aunque la experimentación descrita en el Apartado 7.4.4 se basa en el caso particular del juego *Unreal Tournament 2004* (UT2004), la instanciación de *ConsScale* descrita aquí servirá para cualquier otro juego de tipo FPS. Los agentes o bots de UT2004 que se han analizado están diseñados específicamente para competir

entre ellos en una modalidad de juego denominada *deathmatch* (combate a muerte). El objetivo principal de esta modalidad de juego es matar el mayor número posible de oponentes hasta alcanzar o bien un tiempo máximo, o bien un número máximo de muertes de enemigos.

En este dominio de problema se han considerado los siguientes objetos (información que puede ser adquirida a través de los sensores): jugadores, munición y armas. De forma análoga, se han identificado las siguientes acciones básicas: moverse en una dirección, saltar, correr, girar, dañar, disparar y enviar mensajes por "chat".

La aplicación de esta ontología basada en objetos y acciones permite redefinir las habilidades cognitivas de *ConsScale*, adaptándolas a este dominio de problema. Consecuentemente, se pueden asociar perfiles de comportamiento específicos a estas funciones cognitivas, que ahora son dependientes del dominio. Mediante la observación objetiva de estos comportamientos se puede evaluar la presencia de las funciones cognitivas en el agente que se está analizando.

Aunque la ontología definida no coincida con las posibles representaciones internas que los agentes puedan usar, seguirá siendo válida en términos de evaluación del comportamiento. Dado que la evaluación se realiza en base a las observaciones en tercera persona, la forma concreta de las representaciones internas que los agentes usen es irrelevante.

Usando la ontología para videojuegos FPS que se ha esbozado anteriormente, las habilidades cognitivas ($CS_{i,j}$) genéricas descritas en la Tabla 5 se pueden instanciar o traducir en funciones cognitivas específicas de dominio ($DCS_{i,j}$). Adicionalmente, se pueden definir perfiles de comportamiento ($BP_{i,j}$) asociados a casa una de estas funciones cognitivas. En realidad, pueden existir perfiles de comportamiento que se correspondan con más de una función cognitiva. La Tabla 8 contiene la definición de las funciones cognitivas específicas de dominio y sus perfiles de comportamiento asociados.

Tabla 8. Instanciación de *ConsScale* para videojuegos FPS.

Funciones cognitivas específicas de dominio	**Perfiles de comportamiento asociados**
$DCS_{2,1}$: Reflejos.	**$BP_{2,1}$:** Reflejos básicos, como la capacidad de retroceder cuando el bot choca o se bloquea con otro bot o con un obstáculo.
$DCS_{3,1}$: Capacidad de aprender nuevos comportamientos simples adaptados al juego.	**$BP_{3,1}$:** Comportamientos básicos que ayudan al bot alcanzar mejores puntuaciones, como disparar a otros jugadores cuando estos son detectados.
$DCS_{3,2}$: Habilidad para usar el propio estado interno (salud, munición, etc.) para aprender nuevos comportamientos adaptativos.	**$BP_{3,2}$:** Buscar botiquines (*packs de salud*) cuando el nivel de salud es bajo, o buscar munición cuando el bot se está quedando sin ella.
$DCS_{3,3}$: Habilidad para ignorar la entrada sensorial que no es crítica para la tarea actual.	**$BP_{3,3}$:** Ignorar los kits de recarga de munición detectados cuando el bot está envuelto en un combate y no necesita más munición.
$DCS_{3,4}$: Habilidad para descartar acciones que no son adecuadas para la situación actual.	**$BP_{3,4}$:** El bot evita acciones inútiles como disparar a las paredes cuando está huyendo de un enemigo.

Funciones cognitivas específicas de dominio	Perfiles de comportamiento asociados
$DCS_{3,5}$: Habilidad para seleccionar la información que merece la pena recordad (recuperar de la memoria).	**$BP_{3,5}$:** Cuando el bot necesita munición, accede a su memoria para obtener la posición de los kits de recarga de munición vistos anteriormente. Se dirige directamente a recoger el más cercano.
$DCS_{3,6}$: Habilidad para evaluar a otros jugadores como amigos o enemigos. Habilidad para evaluar los beneficios obtenidos gracias a diferentes armas, municiones o packs de salud.	**$BP_{3,6}$:** El bot no ataca a sus amigos. Las tareas de sanación (recuperación del nivel de salud) y rearmamento se realizan rápidamente seleccionando los mejores kits de munición y de salud.
$DCS_{3,7}$: Habilidad para seleccionar qué información debería almacenarse en la memoria.	**$BP_{3,7}$:** El bot almacena en su memoria la posición de los botiquines y de los kits de munición. El bot va directamente a las posiciones recordadas cuando necesita munición o recuperar su salud (ver $BP_{3,5}$).
$DCS_{4,1}$: Habilidad para aprender mediante mecanismos de prueba y error.	**$BP_{4,1}$:** El bot identifica a otros jugadores como amigos o enemigos en base a prueba y error. Si un jugador considerado actualmente como amigo (ver $BP_{3,6}$) comienza a atacar al bot, se pasará a considerarlo como enemigo y se desplegarán los comportamientos adaptativos correspondientes (huir o atacar).
$DCS_{4,2}$: Habilidad para adaptar el comportamiento dirigiéndolo a objetivos específicos.	**$BP_{4,2}$:** El bot muestra comportamientos dirigidos y sostenidos hacia sus enemigos, como perseguirlos, dispararlos continuadamente o huir de ellos.
$DCS_{4,3}$: Habilidad para evaluar el propio rendimiento en combate.	**$BP_{4,3}$:** Las acciones del bot que no contribuyen a las metas perseguidas son descartadas. Por ejemplo, el comportamiento de huida se cambia por otro cuando no contribuye a disminuir el daño recibido.
$DCS_{4,4}$: Habilidad básica para planificar los siguientes movimientos.	**$BP_{4,4}$:** El bot muestra una secuencia coherente de acciones planificadas para alcanzar una cierta meta. Por ejemplo, seguir buscando y recogiendo packs de salud hasta que el nivel de salud del bot es el máximo.
$DCS_{4,5}$: Habilidad para mantener una representación espacial relativa de los objetos del juego.	**$BP_{4,5}$:** El bot es capaz de localizar objetos de forma efectiva y calcular la posición relativa de los mismos independientemente de las posiciones cambiantes de su cuerpo y sus sensores (ver $BP_{4,2}$). El bot muestra una precisión de tiro aceptable.
$DCS_{5,1}$: Habilidad para intercalar la realización de diferentes tareas del juego. **$DCS_{5,2}$:** Habilidad para perseguir la consecución simultánea de varias metas del juego. **$DCS_{5,3}$**: Habilidad para evaluar el propio rendimiento en relación a la consecución de varias metas del juego.	**$BP_{5,1\text{-}3}$:** Algunos comportamientos del bot se interrumpen debido a ciertas circunstancias y luego se retoman. Por ejemplo, un combate se elude porque el bot necesita aumentar su nivel de salud, después de recuperar la salud, el bot continua el ataque. Adicionalmente, el bot estima hasta qué punto las metas se están alcanzando en base a las estrategias empleadas. Los comportamientos más efectivos se repiten con más frecuencia que los comportamientos que suelen conllevar resultados pobres.

Funciones cognitivas específicas de dominio	Perfiles de comportamiento asociados
$DCS_{5,4}$: Habilidad para aprender de las experiencia pasada en el juego.	**$BP_{5,4}$:** El bot utiliza la evaluación realizada según $DCS_{5,3}$ para seleccionar las estrategias más prometedoras (ver $BP_{5,1\text{-}3}$). Por ejemplo, el bot aprende a usar las armas más destructivas cuando las tiene a su disposición.
$DCS_{5,5}$: Habilidad para planificar acciones teniendo en cuenta todas las metas de juego activas.	**$BP_{5,5}$:** El bot intercala sus acciones de forma efectiva tal y como requiere el cumplimiento de las múltiples metas activas. Por ejemplo, el bot modifica ligeramente su trayectoria mientras está persiguiendo y disparando a un enemigo para recoger sobre la marcha los kits de munición disponibles en las inmediaciones.
$DCS_{5,6}$: Capacidad para generar contenido mental específico con un significado real basado en la interacción con el entorno.	**$BP_{5,6}$:** El bot genera de forma autónoma representaciones internas de alto nivel de los sucesos que tienen lugar a su alrededor, por ejemplo, la generación de perceptos que representen zonas del mapa que el bot considera peligrosas para él (sin que esté pre-programado específicamente para realizar esta función).
$DCS_{6,1}$: Habilidad para evaluar el propio estado como actor en el juego. Esta capacidad representa el aspecto funcional de las emociones. **$DCS_{6,2}$:** Habilidad para adaptar los mecanismos de control al estado actual. **$DCS_{6,3}$:** Habilidad para mantener una representación de las emociones tal y como se describen en $DCS_{6,1}$.	**$BP_{6,1\text{-}3}$:** El bot entra en un estado determinado dependiendo de la evaluación que realiza del propio rendimiento. El comportamiento global del bot se ve modulado por el estado emocional en el que se encuentra. Por ejemplo, si el nivel de salud es muy bajo y no se encuentran botiquines, el bot tiende a comportarse como si estuviera asustado, evitando cualquier riesgo.
$DCS_{6,4}$: Habilidad para mantener una representación precisa del cuerpo del jugador.	**$BP_{6,4}$:** El bot controla su posición, gesto y orientación de forma efectiva. Por ejemplo, es capaz de coordinar sus sistemas sensoriomotores para correr en una dirección a la vez que dispara en otra dirección, donde se encuentra un enemigo, gracias a un giro de cintura.
$DCS_{6,5}$: Habilidad para aprender conceptos abstractos relacionados con el juego.	**$BP_{6,5}$:** Las decisiones inteligentes que toma el bot indican que ha aprendido conocimiento específico acerca del juego. Por ejemplo, el bot tiende a atacar enemigos solitarios, mientras que huye de grupos de enemigos.
$DCS_{6,6}$: Capacidad de especificación y representación de un flujo integrado de perceptos que incluya el propio estado.	**$BP_{6,6}$:** El bot genera automáticamente una secuencia de representaciones de alto nivel que resumen la situación actual, por ejemplo, “me están atacando y no consigo librarme de mis enemigos”.

Funciones cognitivas específicas de dominio	Perfiles de comportamiento asociados
$DCS_{7,1}$: Habilidad para mantener un modelo del yo y una representación de segundo orden de la relación entre el yo y la acción percibida en el juego. **$DCS_{7,2}$:** Habilidad para mantener una representación análoga de segundo orden de la relación entre el yo y las acciones del bot. **$DCS_{7,3}$:** Habilidad para mantener una representación de segundo orden de la relación entre los sentimientos y el yo. **$DCS_{7,4}$:** Habilidad para auto-reconocerse como jugador y parte del juego. **$DCS_{7,5}$:** Habilidad para hacer planes incluyendo el modelo del yo como un actor en esos planes.	**$BP_{7,1\text{-}5}$:** El comportamiento del bot indica que está presente un sentido del yo. Las decisiones no se toman simplemente en función del estado del jugador (salud, munición, etc.), sino que se basan en un modelo elaborado del yo que constituye la base de la capacidad de TdM. El bot es capaz de reconocerse a sí mismo y las consecuencias de sus propias acciones. En otras palabras, el agente desarrolla la sensación de estar a cargo. Las posibles pruebas conductuales asociadas incluyen derivados de la prueba del espejo (Haikonen 2007a). Sin embargo, este tipo de pruebas son difíciles de implementar en el entorno de un juego FPS.
$DCS_{7,6}$: Habilidad para imaginar el resultado para el yo de las acciones planificadas.	**$BP_{7,6}$:** El comportamiento del bot está modulado por su habilidad para prever (imaginar) el resultado emocional de una acción planificada. Por lo tanto aparecen nuevos comportamientos como resultado de este mecanismo de planificación avanzado. Por ejemplo, el bot desarrolla nuevas estrategias de ataque que no han sido aprendidas usando aprendizaje por refuerzo, sino que han sido imaginadas.
$DCS_{7,7}$: Habilidad para usar los objetos del juego como herramientas (dado que los juegos FPS proporcionan soporte nativo para el uso de armas y vehículos, su uso no puede considerarse como una capacidad cognitiva del bot).	**$BP_{7,7}$:** El bot es capaz de usar algunos objetos como medio para alcanzar sus objetivos. Por ejemplo, usando un objeto que se pueda mover, como una caja o barril, como un escudo improvisado.
$DCS_{7,8}$: Capacidad de representar y auto-comunicarse (a sí mismo) contenido mental (flujo interior continuo de perceptos).	**$BP_{7,8}$:** El bot reproduce internamente una "película" en forma de secuencia de perceptos que representan lo que pasa alrededor del bot, incluyendo sus propias acciones. Esa película se usa como entrada en los mecanismos de toma de decisiones.
$DCS_{8,1}$: Habilidad para modelar otros jugadores como yos intencionales con subjetividad.	**$BP_{8,1}$:** Dado que el bot puede modelar a otros jugadores como sujetos intencionales, puede predecir sus movimientos. El bot se pone (mentalmente) en el lugar de otros jugadores para poder prever las acciones de sus oponentes. El comportamiento del bot se modula no sólo en función de la información percibida, sino que también se tienen en cuenta los movimientos que previsiblemente realizará el oponente. Por ejemplo, el bot predice el posible camino de huida de un enemigo y hace el movimiento necesario para bloquearlo (incluso antes de que el oponente inicie su movimiento).
$DCS_{8,2}$: Habilidad para aprender de otros jugadores por imitación.	**$BP_{8,2}$:** Como el bot mantiene tanto un modelo del yo como un modelo de los otros, también puede establecer analogías y aprender estrategias observando otros bots. Por ejemplo, el bot puede adquirir nuevas estrategias de ataque llevadas a cabo por jugadores humanos que participan en el mismo juego.

Funciones cognitivas específicas de dominio	Perfiles de comportamiento asociados
$DCS_{8,3}$: Habilidad para colaborar con otros jugadores para obtener mejores puntuaciones. **$DCS_{8,4}$:** Habilidad para hacer planes incluyendo los modelos de los otros jugadores como actores en los planes (intersubjetividad).	**$BP_{8,3\text{-}4}$:** Los bots muestran comportamientos sociales como la formación de grupos que colaboran en los combates.
$DCS_{8,5}$: Habilidad para construir nuevas herramientas que se puedan usar para conseguir metas del juego.	**$BP_{8,5}$:** El bot es capaz de combinar varios objetos del escenario de juego con el objetivo de crear un objeto compuesto que sirva como defensa o ataque. Por ejemplo, el bot construye una barricada improvisada hecha de barriles colocados a lo largo de una línea.
$DCS_{8,6}$: La generación de imágenes mentales incluye contenido relacionado con modelos actualizados de otros individuos.	**$BP_{8,6}$:** El bot mantiene modelos actualizados de otros jugadores y usa esos modelos para construir imágenes mentales más ricas que representen mejor el mundo, por ejemplo, "estoy atacando al jugador XYZ, que normalmente me tiene miedo".
$DCS_{9,1}$: Habilidad para desarrollar estrategias Maquiavélicas como parte del juego. **$DCS_{9,2}$:** Aprendizaje de nuevas estrategias Maquiavélicas.	**$BP_{9,1\text{-}2}$:** El bot es capaz de razonar sobre las capacidades de TdM de los oponentes. Por lo tanto, muestra comportamientos de inteligencia social, como preparar una emboscada.
$DCS_{9,3}$: Habilidad para comunicar el contenido mental.	**$BP_{9,3}$:** El bot usa el sistema de chat del juego para comunicar de forma coherente su estado mental explícito.
$DCS_{9,4}$: Habilidad para desarrollar una cultura.	**$BP_{9,4}$:** El perfil de comportamiento asociado con la cultura requeriría un entorno más complejo. Sin embargo, se podrían observar indicios de rasgos culturales en grupos de bots organizados.
$DCS_{9,5}$: Habilidad para modificar el entorno y adaptarlo a las necesidades propias.	**$BP_{9,5}$:** Como en $BP_{9,4}$ los comportamientos complejos asociados con estas habilidades requieren entornos más elaborados. Sin embargo, un indicio de esta capacidad sería por ejemplo un comportamiento basado en el movimiento y recolocación de objetos móviles del mapa para construir una fortaleza o un muro de defensa.
$DCS_{10,1}$: Habilidad para comunicar de forma precisa el contenido mental. **$DCS_{10,2}$:** Habilidad para superar una versión adaptada para juegos FPS del Test de Turing.	**$BP_{10,1\text{-}2}$:** El bot es capaz de superar un Test de Turing adaptado, como el propuesto en la competición BotPrize (IEEE Symposium on Computational Intelligence and Games 2008). También se podría plantear un Test de Turing clásico usando el chat incorporado en el juego.
$DCS_{10,3}$: Habilidad para desarrollar una civilización y una tecnología.	**$BP_{10,3}$:** Como en el caso de $BP_{9,4}$, los comportamientos complejos asociados con estas habilidades requieren entornos más elaborados.

Funciones cognitivas específicas de dominio	Perfiles de comportamiento asociados
$DCS_{11,1}$: Habilidad para gestionar varios flujos de conciencia concurrentes.	**$BP_{11,1}$:** Por ejemplo, el bot es capaz de mantener conversaciones complejas por el char con humanos (pasando el test de Turing en todas ellas) a la vez que combate a sus enemigos de forma efectiva. Todos los flujos de conciencia compartirían el conocimiento. Es decir, las conversaciones podrían incluir comentarios significativos sobre el desarrollo actual de la partida.

En el contexto de la evaluación de un agente, el orden que las habilidades cognitivas ($CS_{i,j}$) ocupan dentro de un nivel no es significativo (ya que las funciones CS que comparten un mismo nivel no son comparables en términos del orden parcial establecido en el conjunto CCS). Además, como se describe en la Tabla 8 algunas habilidades cognitivas se pueden agrupar y asociarse a un único perfil cognitivo (BP). Para mantener la consistencia en la medición y respetar la jerarquía establecida en el poset (CCS, <), han de diseñarse pruebas cognitivas completas, que sean fieles a la inspiración integradora y basada en el desarrollo que propone la escala *ConsScale*. Es decir, las pruebas cognitivas de más alto nivel deben requerir la presencia y la integración efectiva de las habilidades cognitivas de bajo nivel de acuerdo con las relaciones "<" establecidas entre las mismas (ver Figura 34).

Gracias a los perfiles de comportamiento de dominio específico definidos en la Tabla 8, los bots diseñados para competir en un juego FPS se pueden evaluar en base a observaciones en tercera persona. Los comportamientos correspondientes a los niveles más altos son difíciles de desarrollar e identificar en un entorno de la complejidad de un videojuego. Específicamente, los niveles 9 y 10 requerirían entornos extremadamente complejos, como el mundo real, para poder ser evaluados satisfactoriamente.

Aunque el dominio de problema especificado es relativamente simple, inferir la presencia de los perfiles de comportamiento a partir de observaciones puede ser engañoso en ocasiones. Como el observador humano tiene una gran capacidad de Teoría de la Mente, tenderá a atribuir estados mentales al bot incluso cuando en realidad no existen. Por lo tanto, se requieren protocolos de prueba específicos, como los propuestos en la competición BotPrize (Hingston 2009), que incrementan la probabilidad de obtener evaluaciones precisas.

También existe otro problema de evaluación relacionado con el desarrollo de los propios agentes. Dado que los procesos de aprendizaje tienen lugar a lo largo del tiempo, el mismo agente mostrará diferentes etapas de desarrollo cognitivo a lo largo de su ciclo de vida. Por lo tanto, el proceso de evaluación estándar de *ConsScale* se puede usar también para evaluar la progresión de un agente hacia las capacidades cognitivas humanas.

6 Generación de Qualia Artificiales

6.1 Introducción

¿Pueden las máquinas tener qualia? ¿Pueden los robots construir mundos interiores de experiencia subjetiva? ¿Serán los qualia experimentados por máquinas comparables a la experiencia subjetiva humana? ¿Está preparado un campo tan joven como el de la Conciencia Artificial para dar una contestación satisfactoria a estas preguntas?

En este capítulo, en vez de tratar de dar una respuesta directa a las preguntas planteadas anteriormente, se argumenta que es necesaria una definición formal, o al menos una caracterización funcional, de los qualia artificiales para poder establecer unos principios ingenieriles válidos para la experimentación en Fenomenología Sintética. Además, se propone un modelo concreto para ser usado posteriormente (ver Apartado 7.6) en la investigación realizada sobre los qualia artificiales.

El descubrimiento de las diferencias existentes entre los qualia naturales y artificiales, si existen, es una de las primeras preguntas que es necesario contestar. Es más, si se pudiera establecer una definición menos ambiciosa para los qualia artificiales, el modelo correspondiente se podría implementar y utilizarse para arrojar luz sobre la verdadera naturaleza de la conciencia. En los siguientes apartados se pretende identificar las características clave que pueden contribuir a una caracterización práctica del concepto de qualia en el contexto de la Conciencia Artificial. Específicamente, se analiza la posibilidad de generación de qualia artificiales como medio para proporcionar una nueva herramienta multidisciplinar para la investigación de la cognición natural y artificial.

6.2 Caracterización de los qualia artificiales

Se sabe que la percepción consciente humana no se basa directamente en los datos adquiridos por los sentidos, sino que está fuertemente sesgada debido a factores psicológicos como por ejemplo el contexto cognitivo, la historia del sujeto y las expectativas. A su vez, el contexto y la historia subjetiva se generan condicionados por la forma en que los estímulos se perciben conscientemente. Aunque los humanos tendemos a pensar que percibimos la realidad, los qualia que se generan en nuestros cerebros distan mucho de ser una representación real del mundo. Sin embargo,

generalmente nuestra experiencia consciente del mundo es lo suficientemente confiable para realizar las tareas diarias sin problemas. En resumen, el sujeto interpreta el mundo de una forma que sea ventajosa para sus metas.

La definición intuitiva de la percepción consciente expuesta anteriormente puede considerarse como inspiración para la construcción de un modelo completo de una máquina consciente que posea aspectos fenomenológicos (ver por ejemplo (Noë 2002) para un análisis detallado de la "gran ilusión" de la conciencia y la fenomenología perceptual). Actualmente, parece que no se dispone de un modelo satisfactorio o de una teoría definitiva sobre lo que podrían ser los qualia en una máquina consciente, en un humano o en otros animales. En otras palabras, la dimensión fenomenológica de la conciencia, tanto en criaturas artificiales como en las naturales, sigue siendo esquiva ante la prospección científica. Adicionalmente, como apunta Sloman (Sloman 2007), muchas veces se discuten conceptos erróneos debido a contextos confusos y términos mal definidos. Aunque en la presente tesis se proponen una serie de definiciones nuevas, no se pretende contribuir a la confusión existente en el campo de la Conciencia Artificial, sino ayudar a aclarar conceptos desde el punto de vista de la ingeniería. Tal y como argumenta Sloman (2007), dirigir las preguntas básicas sobre la conciencia a máquinas con diferentes diseños puede ser útil para averiguar qué es realmente lo que necesita ser explicado.

Mientras que muchas implementaciones de IA abarcan algunos aspectos de la cognición, no se ha alcanzado el punto en el que sea posible construir máquinas con un nivel de conciencia equivalente al humano. Uno de los grandes retos que aún queda por alcanzar es el diseño de un modelo computacional de los qualia, es decir, un modelo de los qualia artificiales. Los diseñadores que trabajan en el campo de la Conciencia Artificial necesitan una definición práctica de lo que podría ser una mente consciente artificial. En el presente trabajo se asume que no existe aún una definición completa y científicamente establecida de lo que podrían ser los qualia artificiales. Por lo tanto, se propone soslayar este problema mediante la formulación de definiciones alternativas, temporales y parciales que puedan contribuir al desarrollo del campo de la Conciencia Artificial. Este enfoque no lleva necesariamente a mejores implementaciones en términos de rendimiento, pero abordar este reto podría proporcionar nuevo conocimiento sobre cómo podrían ser las máquinas conscientes, qué estrategias de diseño parecen más prometedoras y como la neurociencia podría beneficiarse de la aplicación de modelos computacionales de los qualia. Como ha sugerido Chrisley (2009), uno de los retos de la Fenomenología Sintética debería ser la caracterización de los estados fenomenológicos que poseen o modelan las implementaciones de Conciencia Artificial.

En los siguientes apartados se pretende proporcionar una descomposición del problema de la fenomenología artificial en pasos o etapas más tratables y reconocibles. Se mantiene la hipótesis de que las definiciones parciales de los qualia artificiales que se presentan aquí pueden constituir herramientas conceptuales útiles en el dominio de la Fenomenología Sintética.

6.3 Abordando el problema de los qualia

Teniendo en cuenta la complejidad del problema descrito en el apartado anterior, no queda otra opción más que tratar con algunos de los puntos controvertidos relacionados

con el estudio científico de la conciencia. En relación con la especificación de estados fenomenológicos, es conveniente distinguir entre diferentes componentes que se pueden identificar en el intrincado concepto de la conciencia. Además, se debe analizar la forma en la que se puede estudiar de forma satisfactoria la fenomenología, particularmente se aborda el problema de las observaciones privadas en primera persona.

6.3.1 Descomposición del concepto complejo de los qualia

Muchos problemas de la ciencia de la conciencia tienen su origen en el hecho de que el término conciencia se puede usar para referirse a múltiples conceptos (Block 2001). En otras palabras, la conciencia puede ser vista como diferentes fenómenos en función de la perspectiva del observador. Si se distingue entre las dimensiones funcional y fenomenológica de la conciencia tal y como sugiere Block (1995), el proceso de percepción se puede ver desde estas dos perspectivas:

- La percepción consciente es el conjunto de experiencias fenomenológicas que constituyen nuestra vida interior (el "cómo es" tener estados mentales que se experimentan subjetivamente [Block 1995]), como por ejemplo la experiencia de la rojez del color rojo.
- La percepción consciente es el conjunto de representaciones funcionales o modelos internos del mundo adaptados a nuestras necesidades, que aparecen disponibles para su uso en el razonamiento y la acción, por ejemplo, la codificación neuronal del color en el cerebro (ver [Lennie 2003] para más detalles).

Estas dos visiones (conciencia fenomenológica y conciencia de acceso respectivamente) no deberían considerarse exclusivas o contradictorias, sino aspectos complementarios del mismo proceso complejo. De hecho, la cognición natural comparte estas propiedades: los contenidos conscientes de nuestras mentes son tanto estados mentales que se experimentan subjetivamente como representaciones funcionales accesibles para el razonamiento y la acción. Como ha sugerido Haikonen (Haikonen 2009), los qualia son los productos directos del proceso de percepción y sin qualia no hay conciencia. Por lo tanto, los qualia no pueden ser ignorados en el estudio de la conciencia, especialmente si se pretende que los modelos computacionales de Conciencia Artificial sean útiles para comprender la cognición humana.

Mientras que los qualia se asocian normalmente con el primer enfoque (conciencia fenomenológica), la mayoría del trabajo realizado en el dominio de los sistemas cognitivos artificiales está relacionado exclusivamente con la segunda visión (conciencia de acceso). Una de las razones de este sesgo es la pobre comprensión de los aspectos fenomenológicos de la conciencia. Otra razón significativa es que aún queda mucho trabajo por hacer en máquinas que aparentemente no necesitan los qualia para realizar las tareas encomendadas de forma efectiva. Un problema relacionado es la siguiente pregunta: ¿por qué algunas máquinas podrían necesitar los qualia?

6.3.2 El problema de la observación en primera persona

El problema de considerar conjuntamente las dos visiones expuestas anteriormente es que la conciencia fenomenológica sólo está disponible al observador en primera persona, es decir, se trata de una propiedad privada (Dennett 1988). La experiencia interior en los congéneres humanos normalmente se infiere a partir de la observación en tercera persona que realizan otros sujetos (también humanos) basándose en la similitud con su propio caso: "si siento dolor cuando me hago una herida, infiero que otros humanos probablemente sientan lo mismo en la misma situación (porque tienen un sistema nervioso como el mío)". Sin embargo, cuando se trata de detectar la presencia de estados fenomenológicos en máquinas ni siquiera se dispone del argumento de la similitud como factor a tener en cuenta en el proceso de inferencia.

Siendo los qualia inherentemente privados, ¿cómo se podría determinar si una máquina experimenta alguna vida interior? Este problema está relacionado con el denominado problema duro de la conciencia (Chalmers 1995), que aparentemente no ha sido resuelto aún. Esencialmente, no existe ninguna explicación convincente para la conciencia fenomenológica (al menos, no existe una teoría que pueda traducirse en un modelo computacional). ¿Significa esto que cualquier intento de crear qualia artificiales será baladí? ¿O se debería, en vez de rendirse ante el reto, tratar de explorar la generación de qualia artificiales como medio para arrojar luz sobre la verdadera naturaleza de la conciencia? ¿Es necesario comprender completamente la naturaleza de la conciencia para poder reproducirla en máquinas? ¿Realmente existe una falta de herramientas científicas que permitan abordar el problema de la conciencia? ¿Se puede desarrollar un modelo de la conciencia fenomenológica basado exclusivamente en observaciones en tercera persona? Quizás estas preguntas no se puedan responder aún, sin embargo los enfoques basados en la observación en tercera persona se pueden usar para progresar en el campo de los sistemas cognitivos artificiales y su aplicación en la biología inspirada en la Inteligencia Artificial.

Típicamente, la aplicación de enfoques basados en la observación en tercera persona consiste en examinar el comportamiento, incluyendo formas de reporte o comunicación precisa (Seth, Baars & Edelman 2005). Sin embargo, los observadores externos también pueden inspeccionar la arquitectura y los mecanismos internos de una criatura. La inspección interna y monitorización de organismos biológicos vivos, incluidos los humanos, son aspectos mucho más problemáticos que la inspección de una implementación de un sistema cognitivo artificial. Por lo tanto, se debe explotar la posibilidad de análisis de las correlaciones entre el comportamiento observado y la inspección interna de las implementaciones de Conciencia Artificial, ya que podría proporcionar información valiosa sobre los modelos que se están evaluando (sin las obvias limitaciones existentes en los experimentos análogos realizados con criaturas biológicas). Siguiendo esta línea de investigación, se puede confeccionar una definición limitada de los qualia artificiales en la que se use exclusivamente la observación en tercera persona. Esta definición parcial podría no explicar los qualia fenomenológicos tal y como aparecen en los humanos, pero se podría usar para crear modelos computacionales y subsecuentes implementaciones, las cuales a su vez podrían usarse para ampliar el conocimiento acerca de la conciencia.

6.3.3 La función de los qualia

Comprender cuál es la función de los qualia y comprender por qué han aparecido como parte de la evolución biológica es una parte esencial del reto del estudio científico de la conciencia. Como viene siendo habitual, la interrelación entre las ciencias naturales y artificiales puede verse desde dos perspectivas: por un lado, una comprensión completa de los qualia, tal y como éstos se manifiestan en las criaturas biológicas, podría hacer posible la construcción de máquinas conscientes; por otro lado, el camino que lleva a una comprensión completa de los qualia en biología podría pasar por la investigación de nuevos modelos computacionales centrados en la fenomenología.

Estas ideas sobre los qualia no están libres de controversia. Mientras que algunos autores argumentan que los qualia son meros epifenómenos (por ejemplo, [Jackson 1982]), en esta tesis se mantiene la hipótesis de que la conciencia fenomenológica apareció como una ventaja evolutiva. Una forma de probar la validez de esta hipótesis consistiría en comparar el rendimiento de máquinas fenomenológicamente conscientes y máquinas inconscientes, ambas enfrentadas a tareas complejas en entornos desestructurados. Dado que tal experimento no es realizable actualmente, la investigación debe centrarse a corto plazo en el mundo biológico y en los modelos computacionales existentes.

Hay muchas características relacionadas con la conciencia que se sabe que son útiles porque contribuyen a la supervivencia (por ejemplo, la teoría de la mente [Vygotsky 1980], por mencionar una). Estas capacidades cognitivas tienen una función y esa es la razón por la que han sido seleccionadas por la evolución. Pero, ¿ocurre lo mismo con los aspectos fenomenológicos de la conciencia? ¿Tienen éstos un papel funcional claro? Los qualia o la experiencia subjetiva no debería verse como un componente adicional de la noción compleja de la conciencia, sino como un proceso que está presente en relación con las capacidades cognitivas. Los qualia son experimentados por una criatura cuando ésta es capaz de inspeccionar internamente algunos de sus procesos perceptuales y usar esa introspección para generar meta-representaciones que a su vez se usan para modular el sistema global. Los qualia son en realidad la salida del proceso de percepción (Haikonen 2009), que en algunos casos se hacen explícitos gracias al acceso transparente al resultado del mecanismo de percepción (la respuesta del sistema sensorial a los estímulos). En resumen, cuando un sujeto es consciente de un objeto rojo en su campo de visión, éste no percibe el color rojo, sino que experimenta el quale de la rojez, que es la reacción de su sistema perceptivo ante el estímulo de color rojo.

El papel de los qualia descrito anteriormente se puede estudiar en los sistemas artificiales. La generación de qualia artificiales según las líneas indicadas anteriormente podría proporcionar nuevo conocimiento sobre la conciencia aplicable a los organismos biológicos. Tal y como argumentan Sloman y Chrisley (2003), una máquina podría incluso desarrollar sus propias ontologías privadas para referirse a sus propios estados y contenidos perceptuales privados. El uso de esta ontología para modular los procesos del sistema es la función de los qualia. Un sistema con qualia es un sistema con capacidades de *meta-gestión* (combinación de introspección y control activo basado en la auto-monitorización). Un esfuerzo serio para el diseño y construcción de tales sistemas contribuiría a la confirmación o refutación de estas hipótesis sobre el papel de los qualia en los organismos biológicos.

6.4 Modelo propuesto para la generación de qualia artificiales

En el dominio de la investigación en Conciencia Artificial los diseñadores tienen que lidiar con el concepto de qualia para poder desarrollar implementaciones a las que posteriormente se puedan calificar como conscientes. Una forma de mitigar la complejidad que conlleva esta tarea es descomponer conceptualmente la noción de qualia diferenciando diversos aspectos que puedan ser analizados independientemente. En la presente tesis se mantiene la hipótesis de que trabajar con estas visiones parciales puede proporcionar nuevo conocimiento que contribuya a explicar científicamente la conciencia. A continuación se especifican unas definiciones parciales, pero complementarias, de los qualia artificiales.

6.4.1 Definiciones parciales de los qualia artificiales

Se propone distinguir entre tres etapas o estadios que caracterizan el desarrollo de los mecanismos que dan lugar a los qualia en las máquinas (La Figura 62 ilustra la relación entre estas etapas):

- ***Etapa 1. Representación Perceptual del Contenido.*** En esta etapa la información adquirida por el sistema de percepción de la máquina se integra e interpreta, generando una representación subjetiva. Este contenido se construye como resultado de la combinación de los subsistemas de detección exteroceptiva y detección propioceptiva, dando así lugar a una representación inherentemente subjetiva (este proceso es el equivalente al descrito en la arquitectura CERA-CRANIUM para la creación de perceptos – ver Apartado 4.5). El proceso que genera este contenido perceptual involucra una comprobación continua de la consistencia de la información. En otras palabras, un número de posibles reconstrucciones parciales del mundo compiten para ser integradas en la representación consistente final. Este conjunto integrado o reconstrucción final del mundo se logra gracias a una integración coherente entre lo que se percibe desde el mundo exterior y lo que está representado actualmente como modelo interno (el Modelo de las Versiones Múltiples de Dennett constituye una descripción más metafórica de este tipo de procesos de creación de contenido en base a mecanismos de competición/colaboración (Dennett 1991)).
- ***Etapa 2. Meta-Representación Perceptual Introspectiva.*** Esta etapa se refiere a la monitorización de los procesos mencionados en la Etapa 1 y también a la creación de meta-representaciones derivadas. El proceso de observación de cómo se crea el propio contenido perceptual de la máquina puede potencialmente dar lugar a una ontología privada (meta-representación) sobre *cómo es* para la máquina experimentar contenidos perceptuales subjetivos.
- ***Etapa 3. Auto-Modulación y Comunicación.*** En el caso de que la máquina sea capaz de desarrollar las etapas 1 y 2, las meta-representaciones de la etapa 2, u ontologías introspectivas (Sloman 2007), se podrían usar para modular la forma en que funciona el sistema de percepción (incluidas las etapas 1 y 2). Este

mecanismo constituye un lazo de auto-regulación que tiene implicaciones funcionales claras (al igual que ocurre con la capa núcleo en CERA-CRANIUM). Es decir, los qualia se definen a este nivel como parte de un proceso causal. Además, las ontologías introspectivas creadas en la Etapa 2 se podrían usar para comunicar a terceros (y también el propio sistema a sí mismo) una descripción de los qualia artificiales que la máquina "experimenta subjetivamente"[11].

Las etapas descritas anteriormente no especifican las cualidades o modalidades sensoriales asociadas a los contenidos específicos de la mente artificial. Tampoco se especifica hasta qué punto estos qualia podrían ser análogos a los producidos en la experiencia consciente de un humano. La presencia de modalidades sensoriales diferentes (telemetría láser, por ejemplo) y diferentes mecanismos para la cognición producirá contenidos conscientes diferentes asociados a las cualidades correspondientes. Además, los mecanismos abstractos de percepción descritos anteriormente podrían no coincidir con el mecanismo real de generación de qualia en los organismos biológicos.

Las etapas propuestas son componentes de un marco conceptual o guía para el diseño de arquitecturas de Conciencia Artificial. Se espera que las implementaciones y experimentos basados en el modelo propuesto (ver Apartado 7.6) clarifiquen algunos aspectos sobre la naturaleza de la experiencia consciente y su impacto en las habilidades cognitivas.

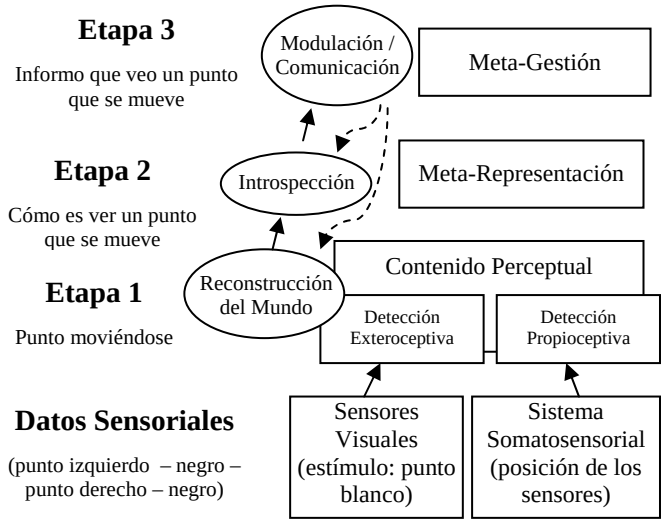


Figura 62. Etapas en el desarrollo de los qualia artificiales.

La definición propuesta anteriormente no abarca la autoconciencia ya que no se considera que ésta sea un requisito para alcanzar la conciencia fenomenológica. Sin embargo, la autoconciencia podría explicarse en el contexto del marco propuesto en base a la inclusión de un modelo del cuerpo en la representación del contenido perceptual. Se esperaría que el concepto de *yo* (o *self*) apareciera como una meta-

[11] Las comillas indican que no se puede demostrar la existencia de estados fenomenológicos asociados al uso de los qualia artificiales en la Etapa 3.

representación en la Etapa 2. Subsecuentemente, se podrían encontrar referencias al *yo* en las comunicaciones generadas en la Etapa 3.

6.4.2 Detectando la presencia de qualia

La definición de los qualia que se ha descrito representa una hipótesis que ha de ser verificada o refutada. Es decir, no se puede tomar como conocimiento válido y establecido. En caso contrario llevaría a la conclusión errónea de que la presencia de qualia se puede comprobar científicamente simplemente verificando la existencia de los mecanismos descritos mediante inspección de un sistema. Por el contrario, lo que se sugiere en el presente trabajo es usar esta aproximación a lo que los qualia podrían ser como hipótesis de trabajo. Este enfoque implica la experimentación con implementaciones de Conciencia Artificial diseñadas para seguir las suposiciones que conforman la hipótesis de trabajo.

Se espera que los resultados de este proceso de evaluación ayuden a discernir si la hipótesis de trabajo planteada era correcta o no. Uno de los beneficios de este tipo de enfoques es que se soslaya el problema de la observación en primera persona. Sin embargo, es necesario realizar esfuerzos adicionales para diseñar experimentos significativos que combinen la salida conductual y la inspección de la arquitectura del sistema. Adicionalmente, se podría poner en el contexto del marco propuesto la identificación de características y fenómenos específicos asociados con los qualia, como por ejemplo la percepción biestable (ambigüedad en la percepción) (Fürstenau 2007).

6.5 Aplicación del Modelo a la Experiencia Visual

Con el objetivo de ilustrar la caracterización propuesta de los qualia artificiales, se ha analizado la percepción consciente del movimiento aparente. Los humanos pueden percibir el movimiento no sólo en objetos que se están moviendo en la realidad, sino que también pueden percibir el movimiento en secuencias de imágenes que contengan estímulos visuales separados espacialmente (Muckli et al. 2002). Los experimentos típicos que se usan para comprobar este efecto consisten en dos estímulos estacionarios que parpadean y que se le presentan al sujeto siguiendo diferentes configuraciones espaciales y temporales (ver Figura 63). Se ha comprobado que hay determinadas frecuencias de sucesión de las imágenes para las que los sujetos informan que perciben conscientemente movimiento (movimiento aparente).

Figura 63. Secuencia de imágenes utilizada para generar la sensación de movimiento.

La secuencia de imágenes representada en la Figura 63 se usa para generar qualia de movimiento aparente en humanos y supuestamente también podrían generar el

mismo efecto en máquinas conscientes. La secuencia de imágenes está compuesta por intervalos inter-estímulo (IIE) en negro que se insertan después de cada estímulo correspondiente a un punto blanco (la secuencia que se repite continuamente es: estímulo de punto blanco en la izquierda – negro – estímulo de punto blanco en la derecha – negro).

Este experimento, que se realiza normalmente con sujetos humanos, también se podría realizar usando una máquina como sujeto. Poniendo este experimento en el contexto de la caracterización propuesta para los qualia, se podría considerar un robot con un sistema de percepción visual inspirado en la corteza visual humana. El contenido básico de cada etapa podría describirse como (la Figura 62 representa los diferentes niveles de contenido en cada una de las etapas que se han definido para el desarrollo de los qualia artificiales):

- Etapa 1: "punto moviéndose".
- Etapa 2: "cómo es ver un punto que se mueve".
- Etapa 3: "informo que veo un punto que se mueve".

En teoría, el proceso de percepción en el robot seguiría estos pasos: en primer lugar, los sensores visuales – un sistema de visión estéreo basado en dos cámaras digitales – adquiere las imágenes usando los correspondientes sensores de detección de luz. Al mismo tiempo, el sistema somatosensorial del robot adquiere la posición relativa de las cámaras, su orientación y su punto de enfoque. La combinación de los datos sensoriales de los sensores exteroceptivos (los mapas de pixeles provenientes de las imágenes) y los sensores propioceptivos (las coordenadas relativas de la posición de las cámaras) se usan para formar perceptos descriptivos siguiendo la filosofía descrita en (Aleksander, Dunmall 2003).

Según se le presenta al robot la secuencia descrita en la Figura 63 se crean perceptos simples que representan la aparición de los puntos y sus posiciones relativas. Subsecuentemente, los detectores de movimiento del robot (estos detectores podrían ser por ejemplo redes de neuronas artificiales), alimentados con la secuencia de perceptos que representan a los puntos, crearían perceptos de movimiento, dependiendo, entre otras cosas, de la duración de los fotogramas IIE. Estos perceptos correspondientes a puntos moviéndose son los contenidos de la Etapa 1 ("punto moviéndose").

La presencia de perceptos de movimiento desencadenaría a su vez una serie de reacciones en el sistema. Por ejemplo, si el robot está diseñado para seguir la trayectoria de ciertos objetos en movimiento o detectar ciertos tipos de trayectorias, los detectores asociados se activarían. También se podrían invocar las evaluaciones afectivas de los perceptos (el robot podría estar diseñado, o podría haber aprendido, que los puntos en movimiento han de ser evaluados de forma positiva y por lo tanto hay que mantener un vínculo con ellos). Si el robot estuviera equipado con un mecanismo para representar estas reacciones generaría meta-representaciones de "*cómo es*" para el robot ver un punto que se mueve. Esto correspondería con la definición de la Etapa 2 del marco propuesto en este capítulo.

Finalmente, si el contenido introspectivo de la Etapa 2 se usa en los procesos de auto-regulación, percepción-acción y también para propósitos de comunicación (bien sea para informar a otros o para informarse a sí mismo), entonces el robot sería capaz de razonar explícitamente acerca de qué significa para él ver un punto que se mueve. Asumiendo que el hipotético robot contara con las suficientes capacidades lingüísticas,

el robot sería capaz de comunicar de forma precisa su contenido mental usando su propia ontología (contenido de la Etapa 3).

6.6 Conclusiones

En este capítulo se ha definido una visión del desarrollo de los qualia basada en trabajo anterior realizado por otros autores. El modelo propuesto constituye un marco conceptual para la creación de nuevos modelos de Conciencia Artificial. La definición de los qualia presentada defiende el papel funcional de la conciencia fenomenológica. Concretamente, se apunta a los mecanismos de auto-modulación e integración de los sistemas perceptuales como funciones clave que se basan en la construcción de meta-representaciones introspectivas.

Se espera que las implementaciones basadas en el marco conceptual propuesto sean capaces de incorporar la dimensión fenomenológica en sus modelos. Asimismo, se confía en que la exploración del espacio de diseño en la dirección apuntada arroje luz sobre el problema de la generación de los qualia en organismos naturales y artificiales.

En el Apartado 7.6 se presentan los experimentos realizados con implementaciones derivadas de CERA-CRANIUM y los resultados obtenidos en experimentos de ilusiones visuales como el usado para ilustrar el presente capítulo. La fenomenología es uno de los campos en los que el avance de la ciencia es más complicado, por lo que se espera que esta disciplina pueda beneficiarse de la aplicación de modelos computacionales como el propuesto en este trabajo.

7 Implementación y Evaluación Experimental

7.1 Introducción

Dado el contexto en el que se enmarcan las líneas de investigación propuestas en la presente tesis, que en el caso del estudio de la conciencia fenomenológica parecen incluso alcanzar los propios límites de la ciencia, cabe hacer énfasis en el seguimiento estricto del método científico. Por lo tanto, la experimentación y evaluación, junto con el diseño de experimentos, son parte crucial del trabajo realizando.

Los métodos de evaluación planteados se basan principalmente en la identificación de ciertas funciones cognitivas asociadas con la conciencia, cómo se pueden implementar en sistemas artificiales, cómo estas funciones pueden interactuar entre ellas y qué beneficios pueden obtenerse de tal implementación e integración. Adicionalmente, se aborda el problema de la fenomenología mediante el estudio de la especificación de los contenidos de la experiencia consciente en un modelo computacional de la conciencia.

En el estudio de la conciencia en sistemas biológicos es imposible realizar este tipo de evaluaciones de forma directa. Los principales problemas asociados a la experimentación con organismos biológicos (incluidos los humanos) es la inexistencia de métodos prácticos (o éticamente aceptables) para añadir, modificar y/o anular determinadas funciones mentales en un sujeto. De hecho, muchos avances en la neurociencia vienen del estudio de pacientes que han sufrido diversos daños cerebrales. Sin embargo, estas limitaciones no impiden que se puedan realizar ciertos estudios comparativos entre humanos e implementaciones de Conciencia Artificial. En este sentido, la evaluación planteada en esta tesis contempla la comparación de los qualia típicamente humanos con la generación de qualia artificiales.

Experimentar con sistemas diferentes también es un problema, ya que la variabilidad existente entre diversos individuos hace más difícil establecer conclusiones unívocas. Por lo tanto, se plantea un proceso de evaluación que emplea un único sistema modular en el que se pueden añadir, modificar y quitar componentes (funciones). De hecho, tal y como se ha descrito en el Capítulo 4, CERA-CRANIUM está concebida como un banco de pruebas con el que se puede evaluar la aportación de diferentes

funciones cognitivas asociadas con la conciencia y la sinergia que potencialmente puede emerger debido a su integración efectiva.

En resumen, el diseño de experimentos y la evaluación realizada durante la presente tesis doctoral giran en torno a las hipótesis de trabajo establecidas inicialmente (ver Apartado 1.1) y los objetivos planteados (ver Capítulo 3). Es decir, se pretende confirmar que:

- Es posible estudiar la conciencia fenomenológica en sistemas artificiales.
- La conciencia tiene una función integradora y adaptativa.
- La conciencia se caracteriza mejor como un proceso y no como una propiedad.
- Para entender la conciencia es necesario descubrir cómo funciona el inconsciente.
- Es posible medir y caracterizar el nivel de conciencia de un sistema artificial.

En definitiva se pretende demostrar que la Conciencia Artificial es un campo de investigación legítimo, que aunque actualmente es inmaduro, puede contribuir significativamente tanto a la comprensión de la mente humana como a la construcción de máquinas más flexibles y robustas. El trabajo realizado en la presente tesis pretende contribuir a este objetivo general mediante el uso de dos herramientas principales: CERA-CRANIUM (ver Capítulo 4) y *ConsScale* (ver Capítulo 5). Ambas herramientas se han aplicado a diferentes dominios de problema con el objetivo de arrojar luz sobre las hipótesis de trabajo resumidas anteriormente.

7.2 Entornos de experimentación

Atendiendo a los requisitos asociados a las hipótesis de trabajo consideradas se distinguen tres contextos diferenciados (pero relacionados) en los que se enmarca la experimentación realizada:

- **Aplicación del modelo de Conciencia Artificial propuesto al control de agentes autónomos**. Se requiere un entorno de experimentación en el que se puedan evaluar diferentes implementaciones parciales del modelo MC^3 (ver Apartado 4.7) utilizando la arquitectura CERA-CRANIUM como banco de pruebas. Concretamente, se ha comprobado el funcionamiento de la arquitectura cognitiva como sistema de control autónomo de diferentes agentes situados (tanto físicos como simulados por ordenador). Este enfoque permite comprobar la capacidad del modelo propuesto para ser aplicado a dominios de problema diferentes.
- **Aplicación del modelo de Conciencia Artificial propuesto a la Fenomenología Sintética.** Se requiere demostrar la capacidad del modelo MC^3 y la arquitectura cognitiva CERA-CRANIUM para servir de plataforma para la experimentación en Fenomenología Sintética. Se ha comparado la especificación de los contenidos de la experiencia consciente que genera una implementación parcial de MC^3 con los contenidos que afirma experimentar un humano cuando observa los mismos estímulos.

- **Evaluación de la escala de medida propuesta**. Se requiere aplicar la medida propuesta en la escala *ConsScale* tanto a las implementaciones que resultan de la experimentación anterior (implementaciones de CERA-CRANIUM) como a otras implementaciones de Conciencia Artificial realizadas por terceros. De esta forma se ha confirmado la capacidad de *ConsScale* para realizar análisis comparativos de diferentes modelos e implementaciones diseñados para diferentes dominios de aplicación.

Concretamente, para la evaluación del modelo CERA-CRANIUM se han considerado los siguientes dominios de aplicación:

- Control autónomo de un robot simulado en tareas de reconocimiento y mapeo de entornos desconocidos utilizando las implementaciones "*CERA-CRANIUM Explorer*" (ver Apartado 7.4.2).
- Control de un robot simulado en tareas de persecución de un objetivo utilizando la implementación "*CERA-CRAMIUM Chaser*" (ver Apartado 7.4.3) .
- Control de un personaje sintético en un videojuego de acción en primera persona utilizando las implementaciones "*CERA-CRANIUM Bot*" (ver Apartado 7.4.4).
- Especificación del contenido de la experiencia visual en CERA-CRANIUM utilizando la implementación "*CERA-CRANIUM Observer*" (ver Apartado 7.6).

Para la validación de la escala *ConsScale* se ha considerado la evaluación de las siguientes implementaciones o modelos (ver Apartado 7.5):

- *CERA-CRANIUM Bot*.
- Agente conversacional Eliza (Weizenbaum 1966).
- Arquitectura mínima para la imaginación funcional en el robot CRONOS (Marques, Holland 2009).
- Modelo LIDA (Franklin et al. 2007a, Baars, Franklin 2009).
- Arquitectura cognitiva de Haikonen (Haikonen 2007b).

7.3 Herramientas y recursos utilizados

A continuación se detallan las principales herramientas y recursos utilizados tanto para la realización de los experimentos descritos en la presente tesis como para el desarrollo y configuración de los propios entornos y agentes implementados.

7.3.1 Robotics Developer Studio

Microsoft Robotics Developer Studio (RDS) es un marco para el desarrollo de aplicaciones de control de robots (Microsoft, Corp. 2006). Como se ha indicado en el Capítulo 4, RDS se ha utilizado como software base para las implementaciones del modelo CERA-CRANIUM. Asimismo, la experimentación y pruebas relacionadas con la arquitectura cognitiva también se han basado en esta misma plataforma, haciendo uso

de los componentes necesarios para la ejecución de aplicaciones robóticas tanto en el simulador (*Visual Simulation Environment*) como con robots reales (ver Figura 64).

Entre otras, las funcionalidades clave de RDS incluyen el soporte en tiempo de ejecución para la concurrencia y la entrada/salida asíncrona (Richter 2006) y el soporte para la ejecución de servicios software distribuidos (Nielsen, Chrysanthakopoulos 2006). Dado que RDS está diseñado para ser una plataforma de desarrollo genérica, puede utilizarse con gran diversidad de hardware de diferentes fabricantes de robots y componentes robóticos.

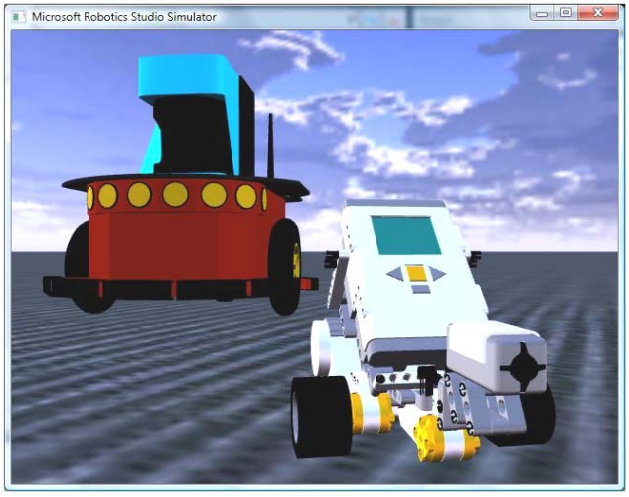


Figura 64. Entorno de Simulación Visual de RDS.

Una aplicación escrita para funcionar con RDS, como es el caso de CERA-CRANIUM, es en esencia una coordinación de diversos servicios distribuidos y asíncronos. En el caso de un ejemplo de aplicación muy sencilla, un sensor se maneja como un servicio que ofrece una entrada de información proveniente del mundo, un actuador tendrá otro servicios asociado que permite el envío de comandos de control y, finalmente, un servicio controlador podría encargarse de interpretar la información obtenida por el sensor y mandar los comandos apropiados al actuador. La coordinación se produce gracias a la comunicación asíncrona entre todos estos servicios. Las aplicaciones se complican cuando el número de sensores y de actuadores aumenta. El funcionamiento de los servicios asociados a los sensores y actuadores sigue siendo análogo al descrito anteriormente, sin embargo, el servicio coordinador (o servicios coordinadores) debe manejar mucha información en tiempo real y aplicar complejas políticas de control (como es el caso de CERA-CRANIUM).

En cuanto al rendimiento, el soporte de ejecución de RDS puede manejar más de 10.000 mensajes por segundo entre servicios cuando todos los servicios están en el mismo nodo u ordenador (dependiendo principalmente de la capacidad computacional del nodo y del tamaño del contenido de los mensajes). Cuando los servicios están distribuidos entre varios nodos y se usa una conexión TCP/IP se pueden alcanzar ratios de unos 1.500 mensajes por segundo. El tiempo medio de latencia está por debajo de 1 milisegundo (Microsoft, Corp. 2006).

7.3.2 Robot Pioneer 3-DX y software asociado

El robot móvil Pioneer 3-DX, fabricado por la compañía MobileRobots Inc., se ha seleccionado como plataforma hardware para evaluar los modelos planteados en un entorno de laboratorio. La integración de la serie Pioneer DX con RDS y su versatilidad a la hora de añadir sensores y actuadores adicionales hacen que sea una buena opción para el desarrollo de los experimentos con robots reales.

El robot Pioneer 3 DX basa su movilidad en dos ruedas motrices que forman una tracción diferencial. En su versión P3-DX8 puede llevar cargas de hasta 23 Kg. de forma robusta y tiene gran autonomía. Aunque el P3-DX ofrece la opción de ordenador empotrado, en la configuración utilizada se ha equipado con un ordenador portátil estándar y sensores adicionales conectados al mismo.

Las principales características hardware de las unidades empleadas son: microprocesador RISC Renesas de 32 bits modelo SH2-7144, telémetro láser Sick LMS-100 orientado frontalmente, cámara PTZ Logitech Sphere MP orientada frontalmente, 3 baterías de ácido intercambiables en caliente (que añaden peso y estabilidad al robot y permiten un consumo de hasta 252 vatios por hora), anillo frontal de 8 sensores Sonar de 15 grados de apertura angular cada uno, parachoques frontales y traseros, dos motores eléctricos que forman un esquema de tracción diferencial, 2 ruedas motrices de 19 centímetros de diámetro y una rueda de giro libre que permiten una velocidad máxima de 1.6 m/s (ver Figura 65).

Figura 65. Robot Pioneer 3-DX.

En los experimentos realizados se han empleado tanto robots reales como robots Pioneer 3-DX simulados usando RDS (ver Figura 66).

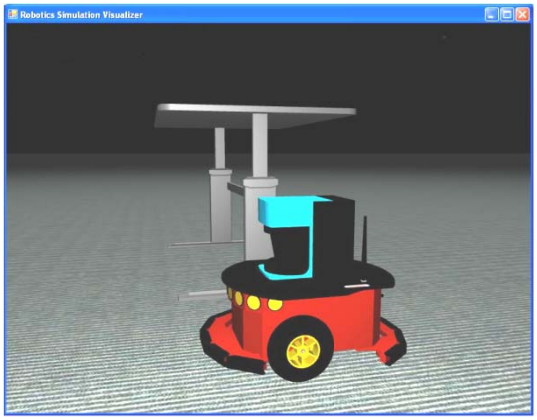


Figura 66. Robot Pioneer 3-DX simulado.

Al igual que otros robots basados en el firmware ARCOS, el robot Pioneer 3-DX8 está basado en un modelo cliente-servidor que proporciona una serie de bibliotecas y utilidades para aplicaciones inteligentes (ARIA). Sin embargo, esta arquitectura software no es de interés en el presente trabajo, por lo tanto en la configuración utilizada en la experimentación se han usando tanto servicios incluidos en la distribución actual de RDS como otros servicios desarrollados ad hoc para controlar los sensores del robot. Concretamente, se han implementado los siguientes servicios para el robot (que corresponden a la capa de servicios sensoriomotores de CERA[12]. Ver Apartado 4.2):

- ***Arcos Sonar Service***. Este servicio accede al estado del robot Pioneer 3 DX y proporciona las lecturas correspondientes a los ocho sensores sonar dispuestos en el anillo frontal del robot (ver Figura 67). El servicio ofrece estos datos mediante un mecanismo de suscripción y notificaciones asíncronas.

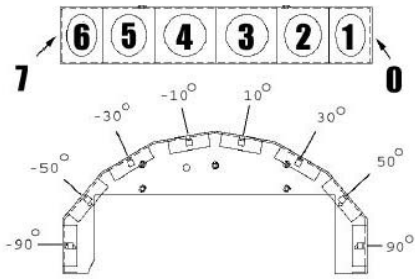


Figura 67. Disposición del arco frontal de sensores sonar en el robot Pioneer 3 DX.

- ***Simulated Sonar Service***. Este servicio es equivalente al anterior, pero está diseñado para ser usado con un robot Pioneer 3 DX simulado. Las lecturas de los ocho sensores sonar simulados se obtienen gracias a la implementación de un algoritmo de trazado de rayos (ver Figura 68). Este algoritmo calcula las distancias que obtendrían los transductores sonar en el entorno tridimensional

[12] Todos estos servicios se han incluido en un paquete denominado CRUBOTS y se distribuyen públicamente bajo una licencia *Creative Commons*. Esto ha permitido a la comunidad de desarrolladores usar este código y proponer mejoras que se han ido implementando durante las sucesivas versiones desarrolladas para la presente tesis.

del mundo simulado en el que se encuentra el robot. Junto con este servicio se han implementado unas entidades simuladas que modelan los transductores sonar reales (modelo SensComp Serie 600) que tiene el robot real. De esta forma las lecturas obtenidas gracias a este servicio se asemejan a las que obtendría el robot real en una situación similar a la simulada.

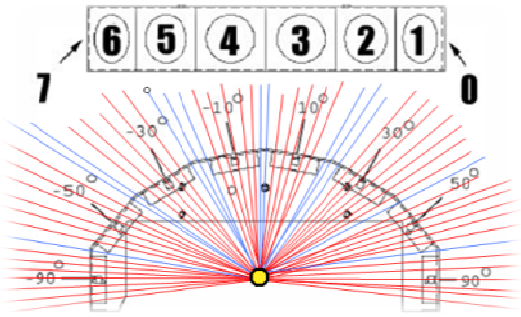


Figura 68. Cálculo de distancias del sonar simulado mediante el trazado de rayos.

- ***Simulated Pioneer 3DX Bumper Service***. Este servicio proporciona notificaciones de contacto correspondientes a los diez sensores de contacto del robot Pioneer 3 DX (dispuestos en dos formaciones, una frontal y otra trasera, de cinco paneles cada una). Análogamente al servicio de sonar simulado, este servicio incluye unas entidades simuladas que modelan la localización relativa y el comportamiento de los sensores de contacto (ver Figura 69). Cada vez que el motor de simulación física detecta un contacto en estas entidades, el servicio procesa esta información y envía a sus suscriptores mensajes de presión o de liberación de los paneles según corresponda (ver Figura 70).

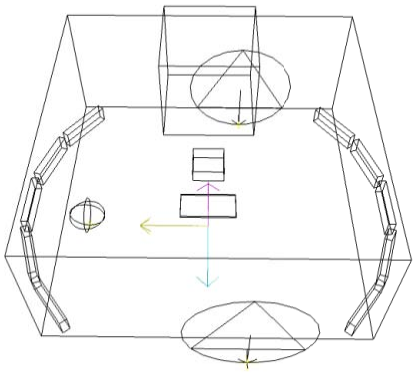


Figura 69. Modelo de los sensores de contacto simulados del robot Pioneer 3 DX.

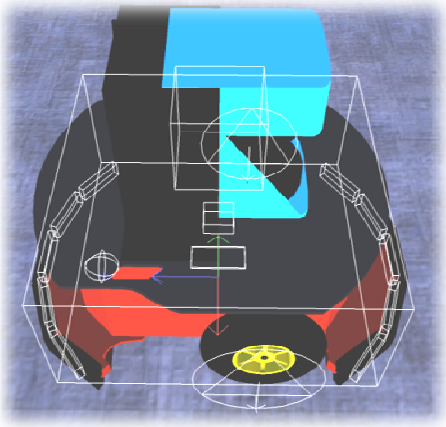


Figura 70. Sensores de contacto en el robot Pioneer 3 DX simulado.

- ***Simulated GPS Service***. Este servicio de GPS simulado proporciona información básica sobre la localización exacta del robot en un entorno simulado. Concretamente, este sensor simulado permite obtener los valores de la orientación y las coordenadas X, Y, Z correspondientes al centro de masas del robot.
- ***Webcam Pan Tilt Service***. Este servicio se comunica con el controlador Logitech para cámaras robotizadas con motores PTZ (Pan-Tilt-Zoom) y permite controlar la inclinación y la orientación de una cámara Logitech Quickcam Sphere MP (ver Figura 71).

Figura 71. Cámara PTZ Logitech Quickcam Sphere.

Otros servicios que se emplean en la capa de servicios sensoriomotores de CERA ya estaban disponibles en la distribución actual de RDS, como por ejemplo el servicio *Arcos Core*, que proporciona acceso al microcontrolador del robot Pioneer 3 DX.

7.3.3 Juego Unreal Tournament 2004, Pogamut y entorno BotPrize

En el ámbito de la aplicación de CERA-CRANIUM al control de agentes autónomos también se han empleado personajes sintéticos en videojuegos (también conocidos como *bots*). A continuación se describen brevemente las herramientas y entornos de experimentación usados.

Con el objetivo de disponer de un videojuego comercial que permitiera la inclusión de bots programables se ha seleccionado el juego Unreal Tournament 2004 (UT2004) desarrollado por Epic Games Inc (ver Figura 72). Asimismo, se ha hecho uso de diversas herramientas que permiten la programación y la depuración de los bots. Se han empleado la biblioteca *GameBots* (Kaminka et al. 2002) y la plataforma Pogamut (Kadlec et al. 2007) para hacer posible la comunicación con el servidor del juego UT2004.

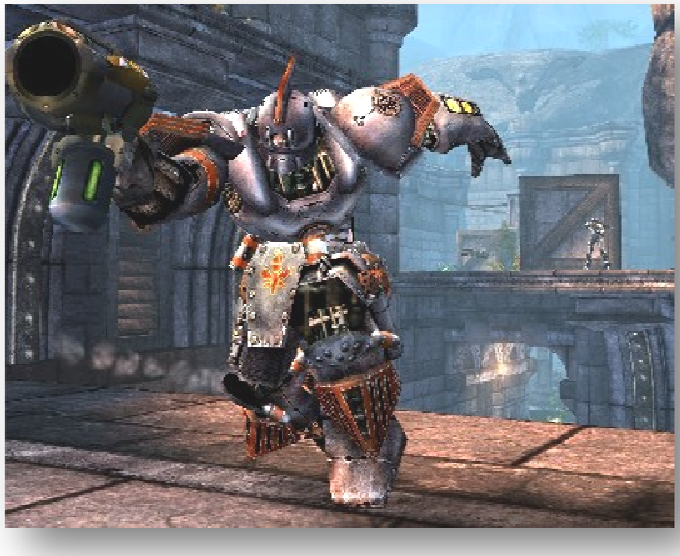


Figura 72. Interfaz del juego Unreal Tournament 2004.

Como CERA-CRANIUM está implementado en .NET Framework y Pogamut está escrito en el lenguaje de programación Java, se ha usado IKVM (Frijters 2009) para generar una versión de Pogamut 2 funcional en el entorno .NET de Microsoft. La Figura 73 representa el entorno de experimentación en el que se han realizado las pruebas con el agente CC-Bot1 (ver Apartado 7.4.4.4).

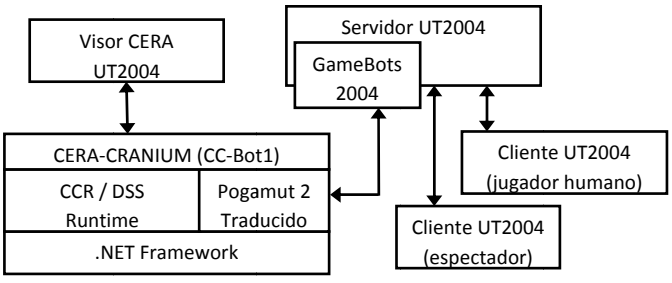


Figura 73. Entorno de experimentación basado en UT2004.

El servidor del juego se ha configurado para iniciar un juego en modo *deathmatch* (“combate a muerte”) en el que dos o más bots pueden competir unos contra otros. El objetivo del modo de juego *deathmatch* es matar a tantos oponentes como sea posible durante la duración de la partida.

Para poder analizar empíricamente el comportamiento de los bots en el juego se ha usado el entorno de la competición 2K BotPrize (Hingston 2009), un Test de Turing para bots. Asimismo, se ha empleado el protocolo de evaluación de los bots introducido en la tercera edición (2010) de la competición (ver Apartado 7.4.4.5).

7.4 Experimentos realizados en el control de agente autónomos

7.4.1 Introducción

Como parte de la evaluación experimental de la presente tesis se han realizado varias implementaciones parciales del modelo MC^3, todas ellas basadas en la arquitectura CERA-CRANIUM. Parte de estas implementaciones están orientadas al control de agentes autónomos. Este es el caso de las implementaciones de las series "CERA-CRANIUM Explorer" y "CERA-CRANIUM Chaser", que están diseñadas para controlar un robot autónomo en tareas de exploración y persecución respectivamente, y las implementaciones de la serie "CERA-CRANIUM bot", que están diseñadas para controlar un personaje sintético en un videojuego de acción en primera persona.

El diseño de los diferentes experimentos realizados en este ámbito con la plataforma CERA-CRANIUM implica la definición de múltiples parámetros de configuración, asociados especialmente con cada implementación parcial concreta del modelo MC^3. En los siguientes apartados se especifican las características principales de las implementaciones realizadas, el dominio de problema donde se han aplicado y los resultados obtenidos.

7.4.2 Aplicación de CERA-CRANIUM a la exploración autónoma

Se ha seleccionado el problema de la exploración automática de entornos desconocidos como dominio inicial para la evaluación de la capacidad de integración de diversas funciones cognitivas en CERA-CRANIUM. Este problema se ha abordado desde la perspectiva de los posibles beneficios que capacidades cognitivas como la atención y el aprendizaje basado en las emociones pueden ofrecer. Las implementaciones realizadas de la serie "CERA-CRANIUM Explorer" (*CC-Explorer*) integran estos conceptos en el marco del control en tiempo real de un agente situado diseñado para la creación de mapas de entornos desconocidos. La experimentación realizada en este contexto permite analizar las relaciones y las sinergias existentes entre algunas de las diferentes funcionalidades cognitivas asociadas con la conciencia.

Es importante remarcar que las implementaciones *CC-Explorer* no pretenden proporcionar una solución al problema clásico de localización y creación de mapas (SLAM). Para ello podrían utilizarse técnicas específicas para localización probabilística como los filtros de Kalman o los métodos de Monte Carlo (Thrun 2000). Dado que el problema de localización no es relevante para los experimentos realizados se ha supuesto que el robot disfruta de una odometría perfecta. Esta situación ideal se ha implementado en el simulador mediante la eliminación del ruido en las medidas de los sensores de distancia y el uso del servicio GPS simulado (descrito en el apartado 7.3.2). En el uso del robot real no es posible eliminar el ruido, ni existe la odometría perfecta. En este caso se ha tratado de minimizar el impacto de la incertidumbre sobre la

localización del agente mediante el uso de velocidades bajas que limitan el impacto de fenómenos inerciales. En cualquier caso, un uso efectivo de *CC-Explorer* en entornos reales requeriría la aplicación de técnicas de localización estadísticas.

Las implementaciones *CC-Explorer* se caracterizan por tener una capa de misión diseñada específicamente para abordar el problema de la exploración de espacios desconocidos y la creación automática de mapas. Los procesadores especializados asociados al ETC de la capa de misión incluyen funciones para la creación de mapas en dos dimensiones y la generación de comportamientos básicos "innatos" para las tareas de exploración. La capa de servicios sensoriomotores contiene los servicios necesarios para controlar el hardware del robot Pioneer 3-DX. Esta capa contiene servicios capaces de comunicarse tanto con sensores reales como con sensores simulados, con lo que el resto de la arquitectura puede funcionar de forma transparente tanto con un robot real como con el simulador de RDS. La capa física está diseñada para crear perceptos simples derivados de la información sensorial específica del robot utilizado. Asimismo, los procesadores especializados de la capa física están orientados a la construcción de perceptos complejos útiles para la posterior construcción de perceptos de misión (actualizaciones de mapa, etc.) La capa núcleo contiene la funcionalidad mínima para habilitar los mecanismos de atención y evaluación global del propio estado.

El comportamiento del robot gobernado por la implementación *CC-Explorer* está determinado por la combinación de las metas activas en las tres capas principales de CERA-CRANIUM. En la capa física, la integridad del hardware del robot se establece como prioridad máxima gracias a un "reflejo" de evasión ante posibles colisiones. Tal y como se esquematiza en la Figura 21, este tipo de comportamiento reflejo aparece como resultado de la interacción en el ETC de la capa física de un procesador reactivo y otros procesadores especializados. En la capa de misión se establecen de forma análoga, mediante procesadores especializados específicos, las metas relacionadas con la exploración de áreas desconocidas. Las metas globales especificadas en la capa núcleo se basan en un modelo de emociones inspirado en el mecanismo de evaluación del propio estado descrito en el modelo MC^3.

Para los experimentos realizados se ha usado inicialmente un entorno virtual que simula el interior de un edificio (ver Figura 74). Los sensores utilizados son los parachoques frontales y traseros (ver Figura 70) y el anillo frontal de ocho transductores sonar (ver Figura 67).

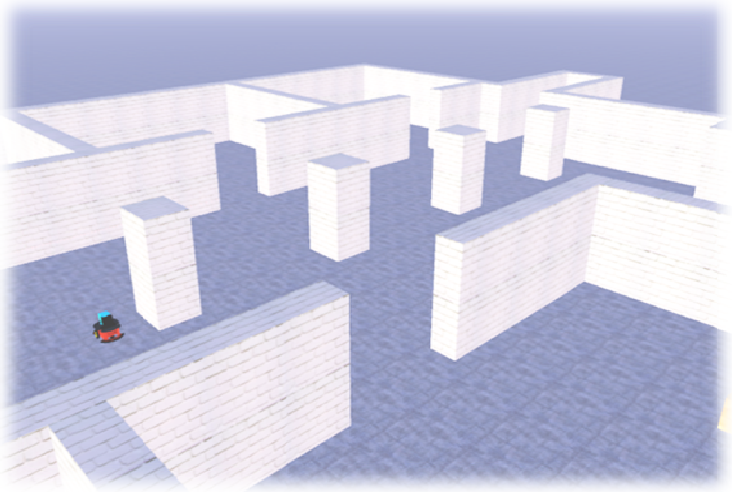


Figura 74. Entorno simulado para tareas de exploración y creación de mapas.

7.4.2.1 Capa física de *CC-Explorer*

La capa física de la implementación *CC-Explorer* se suscribe a los servicios sesoriomotores de la capa inferior (ver Figura 17), de forma que cada vez que un sensor cambia su estado se envía la notificación correspondiente a la capa física. Los procesos de adquisición de los cambios en el estado de los sensores se corresponden con la detección de los sucesos mínimos que el robot puede percibir potencialmente. Siguiendo la notación establecida en el Apartado 4.5, donde *S* representa el espacio sensorial del mundo al que el agente puede acceder, los perceptos mínimos δS_j se corresponden con la información atómica que se obtiene a través de los servicios de los sensores. En *CC-Explorer*, el vector de referencia *j* tiene dos dimensiones espaciales, siendo (x,y) = (0,0) el origen del sistema de referencia del robot (el origen de su punto de vista subjetivo).

Las coordenadas del vector *j* se calculan de forma diferente para cada sensor. Por ejemplo, el conjunto frontal de parachoques del robot Pioneer 3 DX consiste en cinco paneles de contacto. Estos paneles están colocados alrededor del frontal del robot (ver Figura 75), por lo que el vector *j* se calcula dependiendo del panel de contacto presionado.

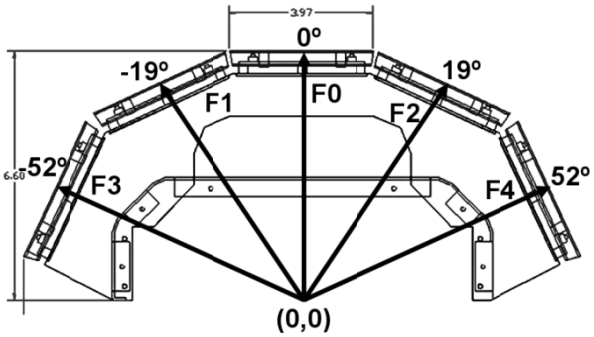


Figura 75. Orientación de los paneles de contacto frontales del robot Pioneer 3 DX.

El conjunto de paneles de contacto frontales del robot Pioneer 3DX están colocados con una orientación de -52º, -19º, 0º, 19º y 52º respectivamente en relación a la bisectriz que divide en robot en dos partes simétricas longitudinalmente. En la capa física hay manejadores para cada tipo de sensor que son capaces de calcular el vector *j* de cada nuevo percepto simple que se crea. El manejador de los parachoques recibe las notificaciones de contacto del servicio Pioneer 3DX Bumper (que se localiza en la capa de servicios sensoriomotores). Con esta información el manejador detecta qué paneles están siendo presionados y asigna los valores correspondientes a los vectores *j* de los perceptos simples – N(δS_J) – que se crean (el *índice-J* puede contener múltiples vectores *j*, así como otros parámetros contextuales adicionales). Seguidamente, estos perceptos simples se envían al ETC de la capa física. Los N(δS_J) correspondientes a contactos en los paragolpes representan la existencia de obstáculos físicos situados en las posiciones relativas indicadas por los vectores *j*.

Dado que los paneles de contacto son partes fijas del robot y su activación sólo se produce por contacto físico, el valor de los parámetros espaciales del vector *j* son

siempre los mismos para cada panel de contacto. Sin embargo, en el caso de otro tipo de sensores (como se explica más adelante) se requiere el cálculo de la posición relativa donde se ha originado el percepto basándose en la propia orientación y posición relativa del sensor. Como ocurre en los sistemas nerviosos biológicos, los manejadores de la capa física de CERA deben ser capaces de estimar la localización física del objeto o suceso que se está percibiendo (Aleksander, Dunmall 2003).

El vector *j* correspondiente a una notificación de contacto de un panel del paragolpes se calcula usando la Ecuación 7, donde *BR* es el radio del panel de contacto, es decir, la distancia desde el centro de masas del robot a la superficie de contacto del panel. *BA* es el ángulo que forma el panel con relación a la bisectriz que separa el robot en dos partes simétricas longitudinalmente y *BH* es la altura a la que están montados los paneles (ver Figura 76).

$$\boldsymbol{j} = (\boldsymbol{X}, \boldsymbol{Y}, \boldsymbol{Z}) = \begin{pmatrix} \boldsymbol{BR} * \boldsymbol{Cos}(\boldsymbol{BA}) \\ \boldsymbol{BH} \\ \boldsymbol{BR} * \boldsymbol{Sen}(\boldsymbol{BA}) \end{pmatrix}$$

Ecuación 7. Cálculo del vector *j* para un panel de contacto.

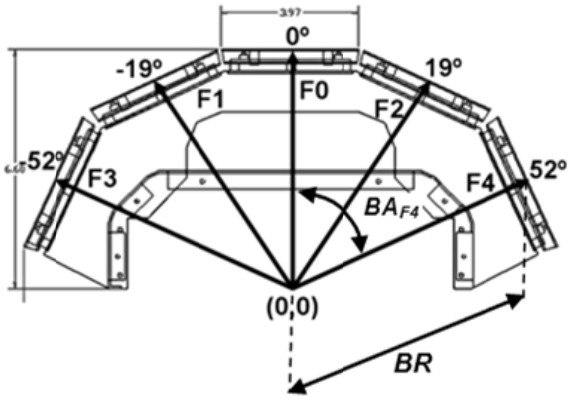


Figura 76. Posiciones relativas de los paneles de contacto del robot Pioneer 3 DX.

Adicionalmente se calculan dos vectores más asociados con el percepto de contacto: el vector de referencia *j* izquierdo (o referente *j* izquierdo) y el referente *j* derecho (ver Figura 77). Estos dos vectores representan la dimensión del percepto (la anchura asignada a la colisión).

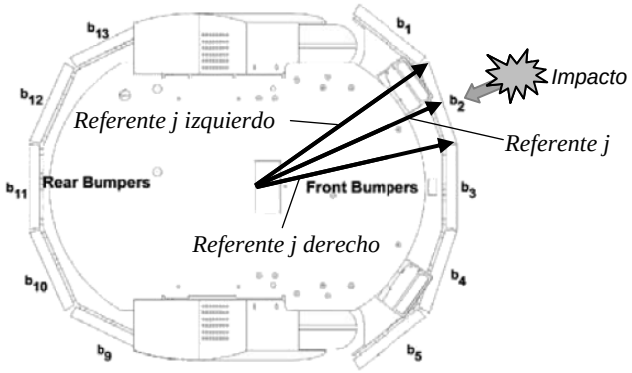


Figura 77. Vectores *j* adicionales para el índice-*J* de un percepto de contacto.

Para calcular el vector referente j asociado a la lectura de un transductor sonar se usa la Ecuación 8, donde *R* es la medida de distancia obtenida por el sonar, *SR* es la distancia desde el centro de masas del robot al transductor y *SA* es el ángulo correspondiente a la orientación del transductor. Los transductores sonar están orientados a -90º, -50º, -30º, -10º, 10º, 30º, 50º y 90º respectivamente con respecto al frente del robot (ver Figura 78).

$$\boldsymbol{j = (X, Y, Z) = \begin{pmatrix} (R + SR) * Cos(SA) \\ SH \\ (R + SR) * Sen(SA) \end{pmatrix}}$$

Ecuación 8. Cálculo del vector *j* para un transductor sonar.

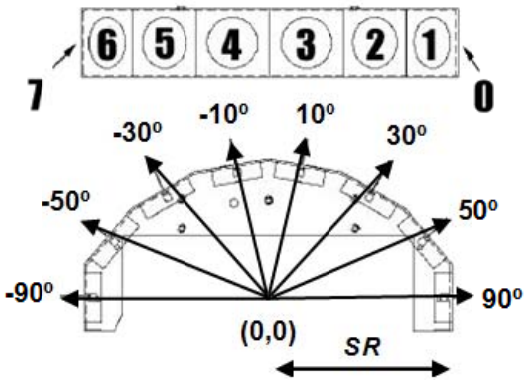


Figura 78. Disposición del sonar frontal del robot Pioneer 3 DX.

Cada transductor sonar es capaz de medir el espacio abierto disponible dentro de un cono tridimensional de 15º de apertura (este rango de apertura corresponde al modelo de transductor SensComp 600). Teniendo en cuenta que los ultrasonidos emitidos por los transductores adoptan la forma un cono tridimensional simétrico, se puede calcular un vector referente *j* para estimar las dimensiones del percepto correspondiente. Es decir, las dimensiones del espacio abierto que se percibe frente al transductor. El principal vector referente *j*, que se calcula usando la Ecuación 8, corresponde con el bisector del cono. Adicionalmente, de forma análoga al proceso descrito para los paragolpes, se calculan dos vectores más: el referente *j* izquierdo y el referente *j* derecho, que representan los límites del percepto en un plano bidimensional (ver Figura 79).

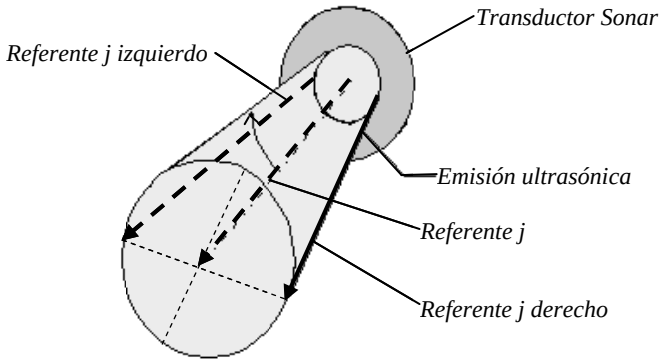


Figura 79. Vectores referentes de un percepto de sonar.

Todos estos vectores se incluyen como parámetros de contextualización espacial en el índice-*J* que se asocia a cada percepto simple. Aunque algunos de los vectores referentes descritos son redundantes son útiles para pre-calcular las regiones del mundo afectadas por un percepto determinado. Asimismo, facilitan la aplicación del mecanismo de contextualización y por ende las funcionalidades cognitivas asociadas, como la atención.

La aplicación de contextos sensoriales da lugar a la formación de perceptos complejos. A continuación se ilustra la formación de perceptos complejos monomodales usando como ejemplo las notificaciones de contacto (en el Apartado 7.4.3.1 se describe la formación de contextos y perceptos multimodales). En el caso de que se notificaran simultáneamente contactos en los paneles b2, b3 y b4, se podrían usar los parámetros contextuales de tiempo y localización relativa para formar un contexto si los tres perceptos simples correspondientes tuvieran marcas de tiempo suficientemente parecidas (ver Figura 80). Por lo tanto, en función del criterio tiempo los tres perceptos simples están asociados en un contexto temporal. Además, como estos paneles de contacto están situados uno al lado del otro, los correspondientes vectores *j* indican proximidad, así que los perceptos simples también se pueden agrupar atendiendo a criterios espaciales.

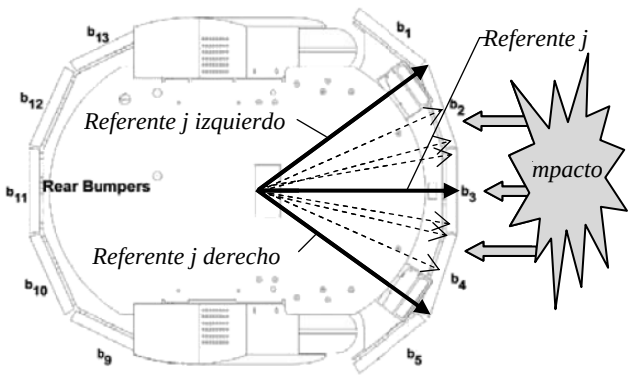


Figura 80. Cálculo de los vectores de referencia de un percepto de contacto complejo.

Gracias a la aplicación de los criterios de contextualización se forma un nuevo percepto complejo que agrupa a los tres perceptos simples creados inicialmente. El nuevo percepto complejo también tiene un *índice-CJ*, que consiste en una combinación de los *índices-J* de los perceptos simples de los que procede. Los vectores *j* del percepto complejo se calculan en función de los vectores *j* de los perceptos simples originales (la Figura 80 representa con líneas sólidas los vectores *j* del nuevo percepto complejo y con líneas discontinuas los vectores *j* de los perceptos simples originales). Un proceso de combinación análogo se realiza con el parámetro *t* (tiempo), asignando una marca de tiempo promedio al perceptos complejo resultante.

La forma en que se calcula el *índice-CJ* de los perceptos complejos varía según la naturaleza de los perceptos simples que entren en juego (forma, dimensiones, etc.). La composición de los *índices-CJ* es trivial cuando todos los perceptos simples pertenecen a la misma modalidad sensorial, como es el caso del ejemplo expuesto anteriormente. Sin embargo, la composición puede ser mucho más compleja cuando se generan perceptos complejos en base a perceptos simples pertenecientes a diferentes modalidades (como es el caso del ejemplo descrito en el Apartado 7.4.3.1). Asimismo, el proceso de creación de perceptos complejos se hace más complicado cuando se tienen en cuenta más parámetros o criterios de contextualización (además de *t* y *j*).

Una vez que los perceptos simples se introducen en el ETC, los preprocesadores de sensor (un tipo concreto de procesadores especializados), los combinan para crear perceptos complejos (ver Figura 81). Estos preprocesadores juegan el papel de los grupos especializados de neuronas que se encargan de detectar (inconscientemente) características o patrones concretos en los datos recibidos de los sentidos. Por ejemplo, el sistema de percepción visual de los mamíferos tiene circuitos neuronales especializados para reconocer simetría vertical, movimiento, profundidad, color o formas (Crick, Koch 1990). De forma análoga, los preprocesadores de sensor de la capa física de CERA proporcionan mecanismos de extracción de características y reconocimiento de patrones apropiados para el entorno del robot. Algunos de los preprocesadores de sensor que se han incluido en *CC-Explorer* son el detector de paredes y el detector de objetos invisibles para el sonar (objetos detectados mediante contactos en los paragolpes pero no detectados por los telémetros sonar).

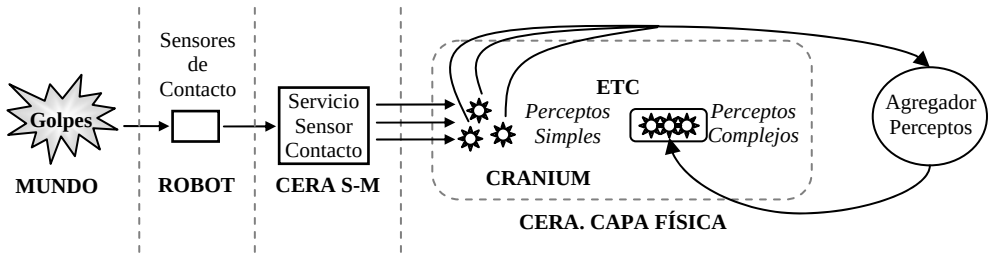


Figura 81. Creación de perceptos complejos en CC-Explorer.

En la capa física de *CC-Explorer* también hay procesadores especializados relacionados con la generación del comportamiento. Como se esquematiza en la Figura 26, los procesadores de la capa física pueden generar comportamientos simples, que a su vez se traducen en comandos que pueden ser enviados a los servicios de control localizados en la capa de servicios sensoriomotores. Inicialmente, se han implementado tres comportamientos simples en *CC-Explorer*: parar, desplazarse hacia el frente y girar.

La operación de desplazamiento hacia el frente toma como parámetro el nivel de potencia de ambas ruedas. La operación de giro usa también un nivel de potencia que ha de aplicarse a cada motor de la tracción diferencial del robot en sentido contrario, provocando así un movimiento de giro. El secuenciador de acciones de la capa física de CERA permite que estas operaciones se ejecuten secuencialmente (ver Figura 26).

Los comportamientos básicos se han definido como N(δB_I), donde *I* es un índice contextual de referencia que principalmente indica la posición relativa que se pretende alcanzar con el movimiento. Por lo tanto, M(*B*) representa el comportamiento global del robot. La secuencia concreta de acciones físicas que se traducen en comandos motores a nivel de la capa física también depende de la generación de comportamientos de alto nivel que se realice en la capa de misión. Asimismo, tanto los comportamientos generados en la capa física como los de la capa de misión están modulados en función de las señales de control emitidas desde la capa núcleo.

Las metas en el nivel físico de *CC-Explorer* pueden verse como "instintos de supervivencia". Tal y como se explica en el Apartado 4.7, las metas del nivel físico se definen como la relación existente entre las percepciones y las acciones. El mecanismo para evitar colisiones puede verse como la implementación de una meta de nivel físico. Sin embargo, hay que tener en cuenta que en el funcionamiento global de CERA-CRANIUM entran en juego diversos lazos de control que operan de forma paralela (ver Figura 22). Esto permite que las capas superiores produzcan una inhibición que evite que se cumplan los objetivos identificados en la capa física.

7.4.2.2 Capa de misión de *CC-Explorer*

La capa de misión usa los perceptos complejos construidos en la capa inferior para generar perceptos de misión (o representaciones del mundo específicas de misión). Esta capa contiene los preprocesadores de misión (un tipo específico de procesadores especializados), que están diseñados para reconocer objetos y sucesos relacionados con la misión usando la información sensorial obtenida de la capa física.

Los segmentos de pared y los obstáculos (perceptos simples y perceptos complejos) detectados por los preprocesadores de sensor y agregadores de perceptos de la capa física se combinan para poder detectar estructuras más abstractas como pasillos o habitaciones. Dado que los perceptos que vienen de la capa física están indizados mediante los *índices-J* o *índices-CJ*, los perceptos de misión se generan como M(S_{CJ}) (ver Ecuación 2). Dado que la meta de misión principal de la implementación *CC-Explorer* es la elaboración de mapas y exploración de entornos desconocidos, se ha usado una representación simple de un mapa bidimensional para la construcción de los perceptos de misión (ver Figura 82).

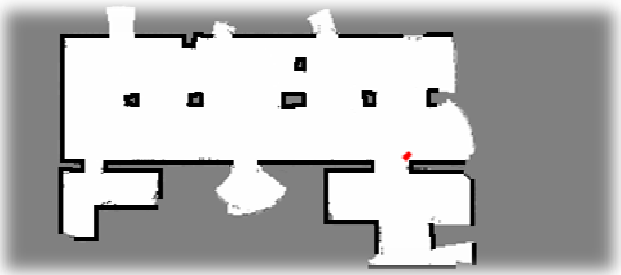


Figura 82. Percepto de misión que representa un mapa bidimensional.

Los comportamientos generados en la capa de misión (comportamientos de misión) se componen de comportamientos básicos N(δB_I) como los generados en la capa física. Los procesadores especializados (planificadores) de la capa de misión generan estos comportamientos en función los perceptos recibidos de la capa física y bajo la influencia de la modulación inducida desde la capa núcleo. En definitiva, diferentes procesadores pueden generar diferentes comportamientos de navegación, siendo la influencia proveniente de la capa núcleo la que en general establece la estrategia que pasa a la capa física mediante el flujo descendente (ver Figura 20) y finalmente se ejecuta.

7.4.2.3 Capa núcleo de *CC-Explorer*

La capa núcleo de CERA se puede ver como el centro de control que orquesta los recursos de procesadores especializados disponibles en las capas inferiores. La parte correspondiente a la implementación parcial del modelo MC^3 realizada en la capa núcleo de *CC-Explorer* es independiente del dominio de aplicación. Todas las representaciones específicas del dominio de aplicación se generan en la capa de misión. En la capa núcleo de CC-Explorer se han implementado una serie de módulos que dan lugar a funcionalidades de propósito general.

El módulo de atención se encarga de dirigir tanto la percepción como la acción. Para tener éxito, el robot tiene que dirigir su atención hacia el cumplimiento de las metas de misión, que se pueden reconocer como soluciones parciales o totales del problema específico para el que se ha diseñado el agente. Sin embargo, el diseño de *CC-Explorer* no sigue esta estrategia de forma directa. En vez de usar las metas de misión directamente para dirigir el foco de la atención, se usan los meta-objetivos definidos en la capa núcleo. Además, el mecanismo de atención implementado es capaz de seleccionar y filtrar información multimodal, ya que se basa en los perceptos complejos y los perceptos de misión (ver Apartado 4.5) para determinar cuáles es la información sensoriomotora más relevante para la situación actual.

El mecanismo de atención se basa en la funcionalidad existente en CERA-CRANIUM para la aplicación de contextos. La selección de estos contextos se usa como medio para seleccionar de forma adaptativa un foco de atención muy limitado dentro de un gran espacio sensoriomotor multimodal. El módulo de atención de la capa núcleo calcula los referentes *I* de interés en función de los perceptos recibidos en la capa núcleo. La información contenida en estos perceptos y que ayuda a establecer qué referentes pueden ser de interés se basa en el mecanismo de evaluación del propio

estado (explicado en el Apartado 7.4.2.4). Con los índices *I* seleccionados se generan comandos de contexto que se envían a las capas inferiores. Para determinar qué N(δB_I) son aplicables se usan los mecanismos de contextualización de CERA-CRANIUM (ver Capítulo 4). El principal criterio utilizado en *CC-Explorer* para la contextualización es la posición relativa. Por lo tanto, los *índices-CJ* de los perceptos representan las localizaciones relativas de esos perceptos. Asimismo, los índices *I* de los comportamientos representan las localizaciones relativas que alcanzaría el robot en caso de ejecutar dichos comportamientos. En general se comparan con los referentes *I* de los posibles comportamientos siguientes N(δB_I) generados en la capa de misión con los comandos de contexto recibidos. Los comportamientos más cercanos a los contextos inducidos tendrán niveles de activación más altos y por tanto más probabilidad de ser seleccionados.

En *CC-Explorer*, los meta-objetivos están relacionados con el estado emocional del robot y permiten tener un mecanismo de atención general capaz de lidiar con múltiples misiones (o metas diferentes de una misma misión). La definición de los meta-objetivos caracteriza la "personalidad" del robot. Inicialmente, se ha considerado un único meta-objetivo general: mantener un estado emocional positivo.

La dimensión emocional mencionada anteriormente es en realidad una caracterización del mecanismo de evaluación del propio estado descrito en el modelo MC^3 (ver Apartado 4.7) implementado en *CC-Explorer*. La evaluación del estado del agente se realiza en función del cumplimiento de las metas establecidas a cada nivel. Es decir, se establecen funciones de evaluación asociadas a las metas implementadas mediante procesadores especializados. A continuación se explica cómo se evalúa el nivel de desempeño del agente en cada capa de CERA.

7.4.2.4 Mecanismo de evaluación del propio estado y emociones en *CC-Explorer*

El envío de comandos motores desde la capa física a la capa de servicios sensoriomotores se realiza en pasos discretos, es decir, el ciclo de ejecución está sincronizado gracias a un reloj interno que marca los *pasos* de ejecución (en los experimentos descritos se ha utilizado un cronómetro interno de 1 milisegundo de resolución y los pasos del ciclo de ejecución son típicamente de 100 milisegundos).

Como se ha mencionado anteriormente, en la capa de misión de *CC-Explorer* se representan los perceptos de misión como mapas bidimensionales del entorno que está explorando el robot. En cada momento, sólo uno de los mapas posibles es seleccionado bajo el foco de atención de la capa de misión (ver Figura 24). Se dice que ha ocurrido una *actualización de mapa* cada vez que el percepto de misión (mapa bidimensional) seleccionado en el foco se actualiza con un nuevo percepto de misión proveniente del ETC de la capa de misión. Si durante el flujo ascendente de percepción, los nuevos perceptos de misión generados no añaden información nueva, el percepto explícito seleccionado no variará. Es decir, no se producirán actualizaciones del estado interno del mundo que mantiene de forma explícita CERA-CRANIUM.

La implementación de un mecanismo de evaluación del propio estado implica la necesidad de monitorizar el propio funcionamiento de la arquitectura. Con este objetivo se han creado procesadores especializados capaces de acceder a parámetros como el

número actual de *pasos* (potenciales movimientos del robot) o el número actual de *actualizaciones de mapa*. Asimismo se cuenta con un mecanismo básico de predicción sensorial. La implementación de este mecanismo consiste en un predictor sensorial que genera un percepto de incongruencia (o *mismatch*) cada vez que el mapa actualizado no coincide con lo esperado. Por ejemplo, cuando se detecta un obstáculo en una zona que antes se había marcado como espacio libre.

Para la evaluación de estado en la capa física se ha desarrollado un procesador especializado de evaluación que calcula el valor de la Ecuación 9. Este procesador genera un percepto de estado que representa el nivel de cumplimiento de los objetivos establecidos en la capa física: "navegar por el entorno de forma segura, sin colisionar con los obstáculos existentes". Los perceptos de estado en *CC-Explorer* consisten básicamente en un número real que indica el nivel de desempeño alcanzado.

$$Eval_{Fisica} = \frac{pasos - colisiones}{pasos}$$

Ecuación 9. Evaluación del rendimiento de la capa física de *CC-Explorer*.

Los perceptos de evaluación de la capa física representan la capacidad que tiene el robot de moverse por su entorno sin colisionar objetos o paredes. Los perceptos de estado de la capa física tienen un valor máximo de 1.0 que indica que no se ha producido ninguna colisión y un valor mínimo de 0.0 que indicaría que el robot está permanentemente golpeando obstáculos.

Al mismo tiempo, en la capa de misión se ejecuta un procesador especializado de evaluación capaz de calcular el rendimiento en cuanto a la consecución de las metas de misión. Este procesador genera dos tipos de percepto de estado. El primero se basa en la Ecuación 10 y refleja el desempeño en cuanto al objetivo de crear un mapa del entorno. El segundo se basa en la Ecuación 11 y refleja el desempeño en cuanto a la creación de un mapa preciso.

$$Eval_{Misión1} = \frac{actualizaciones}{pasos}$$

Ecuación 10. Evaluación de la velocidad en la creación de mapas en *CC-Explorer*.

$$Eval_{Misión2} = \frac{actualizaciones - incongruencias}{pasos}$$

Ecuación 11. Evaluación de la calidad de los mapas creados en *CC-Explorer*.

Finalmente, la evaluación global del propio estado se realiza en la capa núcleo de *CC-Explorer* y se basa en la Ecuación 12, donde *E* representa el conjunto de emociones

consideradas en el modelo MC^3 implementado y *n* es el número de emociones. La función *energía* calcula la intensidad de una emoción dada.

$$Eval_{Núcleo} = \sum_n Energía(E_n)$$

Ecuación 12. Evaluación global del estado en *CC-Explorer*.

El modelo de emociones implementado en *CC-Explorer* se basa en la definición de unas emociones básicas y la asignación de un valor de "energía". En este contexto, las emociones constituyen un mecanismo para sintetizar el rendimiento del agente en cuanto a la consecución de las metas marcadas. Las funciones de evaluación descritas anteriormente se usan para calcular la energía de las emociones. En general, cuando el agente logra progresar en la consecución de una meta se incrementa la energía de emociones positivas como la "alegría". Por el contrario, los fallos repetidos provocan el incremento de la energía de emociones negativas como el miedo o la ira. En *CC-Explorer* se han definido unos operadores emocionales que establecen las relaciones entre determinadas metas y emociones específicas.

En la implementación actual de *CC-Explorer* se han considerado las siguientes emociones definidas en función de su influencia en la modulación del sistema:

- **Curiosidad**: dirige el foco de atención hacia contenidos específicos del espacio sensoriomotor.
- **Miedo**: dirige el foco de atención alejándolo de contenidos específicos del espacio sensoriomotor.

Este modelo básico de las emociones se ha implementado en la capa núcleo de *CC-Explorer* como una forma de establecer reglas adaptativas para el cálculo del *índice-CJ* contextual (ver Figura 83).

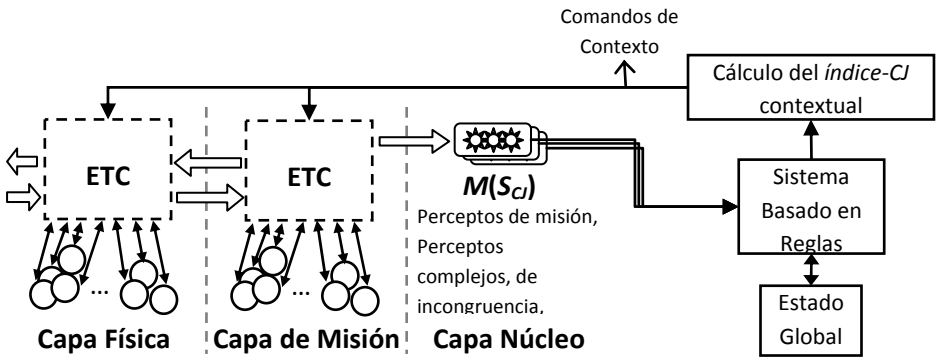


Figura 83. Modulación inducida desde la capa núcleo de *CC-Explorer*.

Al definir la curiosidad como una emoción que dirige la atención hacia contenidos específicos, un *índice-CJ* proveniente de un percepto de novedad desencadenaría la aplicación de una regla asociada con esta emoción. Esto provocaría a su vez que el próximo *índice-CJ* contextual estuviera dirigido en el mismo sentido que el *índice-CJ*

correspondiente a la novedad. Los perceptos de novedad se generan en la capa de misión gracias a procesadores especializados que comparan los mapas existentes con los nuevos perceptos de misión. El ejemplo de la Figura 84 muestra un percepto de misión y también un vector de referencia *j* (con una inclinación de 22.5º y origen en la posición actual del robot) correspondiente a un percepto de novedad generado en la capa de misión.

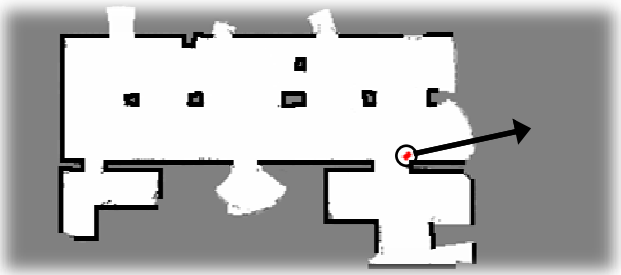


Figura 84. Percepto de novedad asociado a un mapa del entorno.

La aplicación de la regla correspondiente a la emoción de curiosidad producirá un *índice-CJ* contextual que se enviará desde la capa núcleo y a su vez provocará probablemente que se active en la capa de misión un comportamiento orientado en la dirección indicada por *CJ*. Según se mueve el robot se generan más perceptos simples, que se combinan para formar perceptos complejos, que a su vez dan lugar a perceptos de misión (mapas), perceptos de novedad, incongruencia, etc. El modelo implementado en la capa núcleo usa como entrada exclusivamente los perceptos seleccionados en el foco de la capa de misión (contenido explícito). Usando esa información la capa núcleo puede generar un nuevo comando de contexto, cerrando así el lazo de control explícito de la arquitectura CERA-CRANIUM (ver Figura 22).

7.4.2.5 Resultados obtenidos

Se han desarrollado experimentos de navegación en un entorno típico de interior (ver Figura 74) con el objetivo de comprobar la influencia que tiene el mecanismo de atención implementado en *CC-Explorer* en la generación del comportamiento del robot. Para poder analizar efectivamente los efectos del mecanismo de atención en el comportamiento generado se han eliminado de forma artificial los problemas de localización. Concretamente, se ha usado un servicio simulado de GPS (ver Apartado 7.3.2) que proporciona la localización exacta del robot, eliminando así la necesidad de aplicar algoritmos clásicos de SLAM.

El análisis del comportamiento del robot se ha basado en el estudio de la trayectoria seguida y la eficiencia en la creación del mapa del entorno. Para poder observar la influencia concreta del mecanismo de contextualización de CERA-CRANIUM se han realizado diversos experimentos basados en implementaciones de *CC-Explorer* en las que se varía el número y la definición de los contextos activos:

- **Implementación *CC-Explorer-1* (CCE-1)**. Esta implementación se ha creado con el objetivo de generar un comportamiento pobre, que sirva de referencia

para compararlo con el de otras implementaciones de *CC-Explorer*. En este caso, la capa núcleo no envía ninguna consigna contextual y la capa de misión tampoco desempeña ningún papel. Es decir, la implementación funciona únicamente gracias a la operación autónoma de la capa física. Por lo tanto, el comportamiento resultante se basará exclusivamente en la ejecución de los reflejos programados en la capa física. Esta implementación contiene dos procesadores especializados asociados al ETC de la capa física llamados "Detector de objeto más cercano" y "Detector de posible impacto". La Figura 85 el comportamiento típico de CCE-1 en términos de área explorada y trayectoria seguida.

- **Implementación *CC-Explorer-2* (CCE-2)**. Esta implementación incluye un mecanismo de atención que abarca exclusivamente el entorno local del robot. En este caso se establecen un contexto que la capa núcleo activa cuando el robot está fuera de peligro de colisión. Un procesador especializado de la capa física, denominado "Detector de espacio libre", calcula periódicamente un ángulo (relativo al robot) correspondiente a la dirección en la que se detecta que hay más espacio libre para moverse. En la capa de misión se establece la meta de dirigirse hacia las áreas donde hay más espacio libre, que hace que se generen comportamientos en los que el robot se mueve en esas direcciones. Las metas de la capa física sólo se activan cuando los procesadores del nivel físico alcanzan más nivel de activación que los comportamientos que vienen de la capa de misión. Por ejemplo, cuando el robot está demasiado cerca de un obstáculo. El mecanismo de contextualización de CERA-CRANIUM permite que se vete el contexto establecido en la capa de misión a favor del contexto de mayor prioridad presente en la capa física. La Figura 86 describe el comportamiento típico de CCE-2.
- **Implementación *CC-Explorer-3* (CCE-3)**. En esta implementación se ha definido un nuevo contexto con el objetivo de incrementar el rendimiento en la tarea de exploración. Este contexto se activa cuando el robot no está explorando áreas desconocidas. Para esto se ha implementado un nuevo procesador denominado "Mejor dirección de exploración" que calcula la dirección relativa al robot donde hay un área más grande pendiente de explorar. Adicionalmente, en esta implementación se define una meta basada en moverse hacia las áreas inexploradas del mapa. La Figura 87 muestra como la activación de este nuevo contexto hace que el robot tienda a alejarse de áreas ya exploradas previamente.

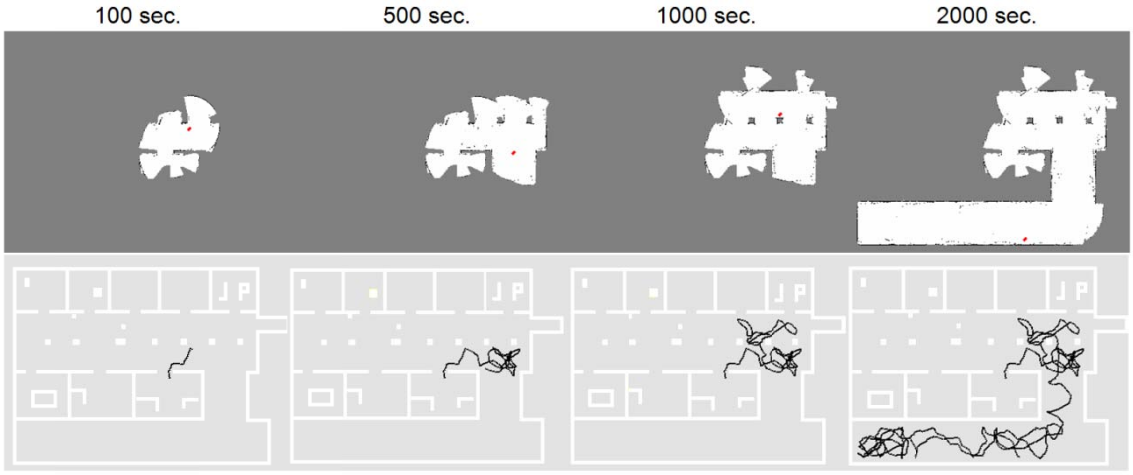


Figura 85. Comportamiento típico de la implementación CCE-1.

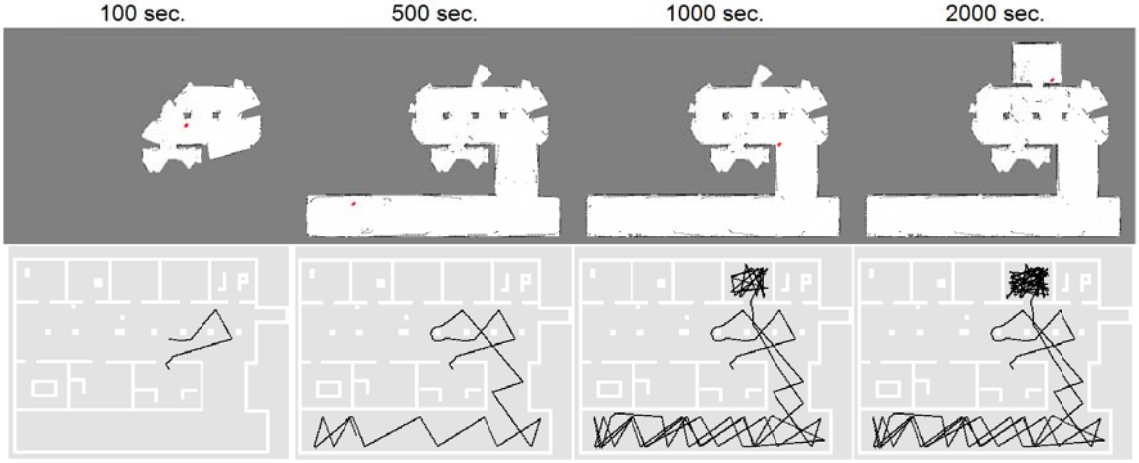


Figura 86. Comportamiento típico de la implementación CCE-2.

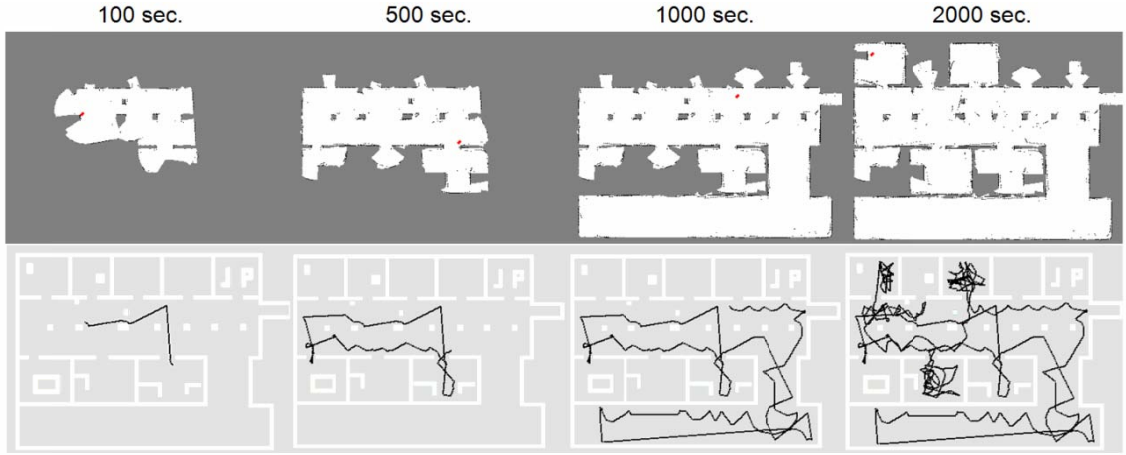


Figura 87. Comportamiento típico de la implementación CCE-3.

La Figura 88 muestra una comparación del rendimiento de las tres implementaciones. La Figura 89 resume el rendimiento comparativo de las implementaciones CCE-1, CCE-2 y CCE-3 en términos de metros cuadrados explorados

por segundo (m^2/s). Como media, CCE-1 es capaz de descubrir 0.19 m^2/s, mientras que CCE-2 y CCE-3 son capaces de construir mapas siguiendo un ratio de 0.29 y 0.36 m^2/s respectivamente. Las áreas del gráfico en las que no se observa un incremento apreciable en el área descubierta corresponden a los momentos en los que el robot queda "atrapado" en una habitación. La falta de un mecanismo de navegación global provoca este tipo de comportamientos. Sin embargo, en cuanto a la navegación local, la aplicación de los contextos proporciona un mecanismo eficiente para integrar estrategias diferentes en función del estado del robot.

La trayectoria que sigue el robot en cada uno de los casos analizados demuestra cómo el mecanismo de atención induce una adaptación mediante el filtrado temporal de información innecesaria. Cuando una meta no se puede cumplir debido a la situación actual del agente, la información sensorial asociada con esa meta se ignora. Por ejemplo, la meta consistente en dirigirse hacia áreas no exploradas no debe aplicarse cuando el robot está maniobrando para evitar un obstáculo.

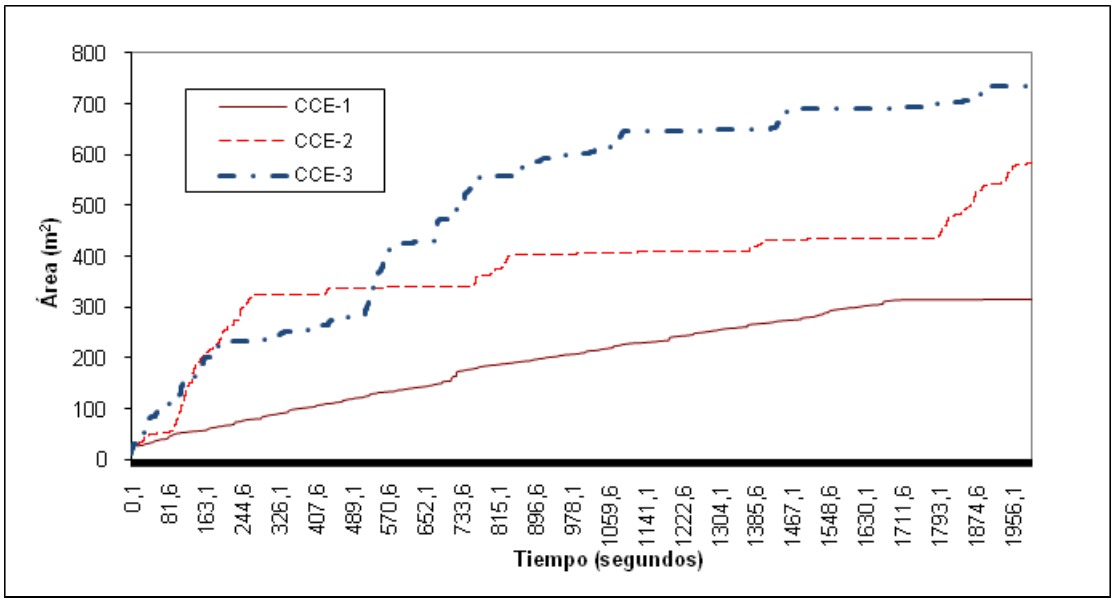


Figura 88. Rendimiento en la exploración realizada por CCE-1, CCE-2 y CCE-3.

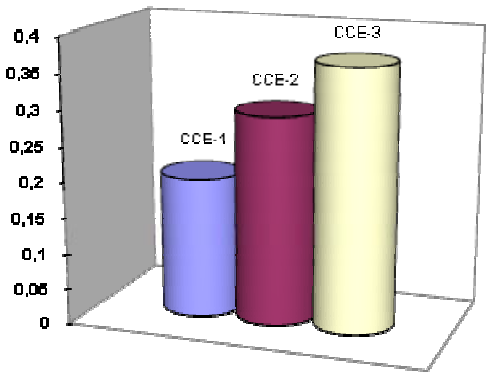


Figura 89. Rendimiento en m^2/s de la exploración realizada por CCE-1, CCE-2 y CCE-3.

7.4.3 Aplicación de CERA-CRANIUM a la tarea de persecución

Para evaluar las habilidades cognitivas desarrolladas en la implementación parcial del modelo MC^3, en este caso se ha seleccionado el problema del reconocimiento de un congénere y su seguimiento o persecución. Principalmente se ha usado un escenario simulado simple y dos robots Pioneer 3 DX simulados. Se ha llamado P3DX-Chaser al robot controlado por la implementación CERA-CRANIUM Chaser (*CC-Chaser*) y P3DX-Target a un robot similar que puede estar controlado por un humano. La misión encargada a P3DX-Chaser es desplazarse de la forma necesaria para mantener una distancia constante y segura entre los dos robots. Para cumplir este objetivo el robot P3DX-Chaser tiene que prestar atención a los perceptos complejos que coincidan con "un objetivo móvil que es un robot Pioneer 3 DX", a la vez que se ignoran otros perceptos que son irrelevantes para la misión actual.

La tarea de persecución definida, al igual que en el caso anterior de la exploración de entornos desconocidos, requiere tomar decisiones acerca de qué entradas sensoriales se procesan y qué repertorio de acciones pueden ejecutarse en un momento dado. Normalmente, la opción de considerar todo el espacio sensoriomotor disponible no es factible, además de innecesaria y muy costosa computacionalmente. El agente debe tener en cuenta su situación actual y la misión asignada para centrar la atención en una selección limitada de perceptos y de posibles comportamientos. Para lidiar con este problema, en la implementación *CC-Chaser* se ha incluido un mecanismo de atención igual al de las implementaciones *CC-Explorer* descritas anteriormente.

En realidad, la principal diferencia entre las implementaciones *CC-Explorer* y *CC-Chaser* reside en la capa de misión y los procesadores especializados asociados a la misma. Dado que el agente controlado es el mismo la capa física es muy similar en ambas implementaciones, aunque la capa física de *CC-Chaser* cuenta con la capacidad de generar perceptos basados en información visual (y la capa de servicios sensoriomotores cuenta con el servicio para adquisición de imágenes de la cámara). En la implementación de *CC-Chaser* se han utilizado las modalidades sensoriales de sonar, visión y contacto, consistiendo los actuadores en el mismo sistema de tracción diferencial usado en *CC-Explorer*. Es decir, en *CC-Chaser*, la capa física usa una modalidad nueva: la visión. El robot P3DX-Chaser cuenta con una cámara de 320x240 pixeles de resolución y un campo de visión de 90º (en su versión simulada).

Dado que el diseño de *CC-Chaser* es muy similar al de *CC-Explorer* detallado anteriormente, se describe a continuación exclusivamente los mecanismos adicionales que se han implementado en el primero, resaltando la técnica de fusión sensorial multimodal que extiende el mecanismo de atención para que pueda ser utilizado en contextos multimodales complejos.

7.4.3.1 Formación de contextos multimodales en *CC-Chaser*

En *CC-Chaser* se aplica la definición de contextos incluso en los primeros pasos en el procesamiento de los datos visuales. De hecho, en vez de realizar una tarea de preprocesamiento de la imagen completa capturada por el sensor de la cámara, cada

fotograma se divide en regiones pequeñas. Sólo una de estas regiones se procesa completamente, imitando así a la fóvea de los ojos biológicos. Los mamíferos usan la fóvea para fijar la vista en un objeto y procesar específicamente esa imagen, mientras que la visión periférica funciona con una resolución muy baja y un procesamiento asociado mucho menor (Wandell 1995).

El procesamiento exclusivo de la región seleccionada (o fóvea simulada) reduce en gran medida los requisitos computacionales del preprocesador de sensor que toma los datos de la cámara. Adicionalmente, tal y como se explica a continuación, esta estrategia permite que el robot centre la atención en regiones específicas también en las siguientes etapas de procesamiento. Sin embargo, antes de la etapa de preprocesamiento, cuando se evalúan los criterios de contexto, todos los segmentos de cada fotograma se procesan de igual forma. Esto es inevitable ya que es necesario calcular los parámetros contextuales de cada percepto antes de poder establecer los contextos activos y seleccionar el foco de atención.

Como se ha visto en el Capítulo 4, los criterios de contexto se usan para establecer el grado de relación entre los perceptos y los contextos indicados en los comandos enviados desde la capa núcleo. El tiempo y la localización relativa son factores básicos que pueden ser tenidos en cuenta independientemente del dominio de aplicación (como por ejemplo en el caso de la exploración y creación de mapas de entornos desconocidos). Sin embargo, se pueden considerar otros factores adicionales dependiendo del dominio de problema y de la riqueza de las representaciones internas (perceptos) que se manejen. En el caso de *CC-Chaser*, se han considerados las propiedades de color y movimiento como criterios adicionales a tener en cuenta. Por lo tanto se han usado cuatro criterios para la formación de contextos.

El criterio del tiempo se refiere al momento exacto en el que se percibe un estímulo. Por lo que el tiempo se convierte en un criterio importante para relacionar un percepto con otro. Dado que diferentes sensores y sus procesadores asociados pueden necesitar diferentes intervalos de tiempo para procesar la información sensorial, se requiere un mecanismo de “alineación del tiempo”. Se ha demostrado que tal mecanismo existen en los organismos biológicos (Senkowski et al. 2007, Giard, Peronnet 1999). Aunque los estímulos visuales y auditivos se procesan a velocidades diferentes, el cerebro es capaz de eliminar el lapso de tiempo entre las diferentes señales procesadas cuyos orígenes fueron adquiridos al mismo tiempo (Spence, Squire 2003). La asignación de marcas de tiempo en CERA-CRANIUM imita la funcionalidad de este mecanismo permitiendo la asociación precisa de perceptos independientemente del tiempo de proceso necesario para crearlos.

La localización relativa es otro criterio fundamental para la formación de contextos dado que la representación de la posición de los objetos en el mundo es un requisito de los agentes situados. Adicionalmente, la localización relativa de un objeto con respecto al cuerpo del agente (o con respecto a cualquier otro sistema de referencia) se requiere para la generación de comportamientos adaptativos. La localización relativa de cualquier elemento del espacio sensorial es necesaria también para la integración de los perceptos complejos en CERA-CRANIUM. Además, permite la selección de una orientación espacial hacia la que dirigir la atención. Se ha demostrado que el cerebro de los mamíferos contiene neuronas que codifican localizaciones espaciales y usan sistemas de referencia, como los centrados en la cabeza o somatotópicos (Avillac et al. 2005, Fogassi et al. 1992).

En un mundo en el que lo patrones de colores se pueden asociar con objetos concretos se debería tener en cuenta esta propiedad. De forma análoga, algunos objetos son móviles mientras que otros permanecen estáticos. Consecuentemente, el movimiento es también una propiedad que se debería considerar como criterio para la formación de contextos. De nuevo, esta decisión de diseño tiene una inspiración biológica basada en la presencia de áreas especializadas en la detección del color y del movimiento en la corteza visual del cerebro humano (Zeki et al. 1991). En el caso particular de *CC-Chaser*, la tarea de reconocimiento de congéneres se ha simplificado enormemente caracterizando a otros robots como "objetos rojos que se mueven".

Siguiendo los principios presentados anteriormente, se han empleado los criterios de contextualización tiempo, localización, color y movimiento para la formación de perceptos y comportamientos. Para poder aplicar estos contextos se requiere que los perceptos incorporen sus parámetros asociados de tiempo estimado, localización relativa, detección de movimiento y color (ver Figura 90). Las propiedades de movimiento se podrían derivar de los parámetros de tiempo y localización. Sin embargo, se ha decidido usar un sistema de detección de movimiento nativo en las que las propiedades de movimiento se obtengan directamente del análisis de la entrada visual.

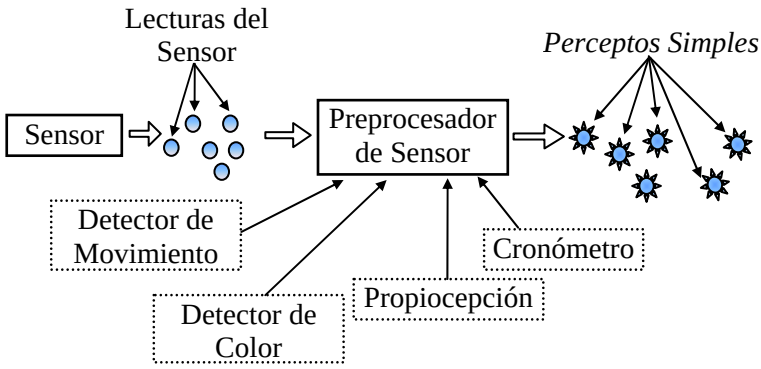


Figura 90. Creación de perceptos simples en *CC-Chaser*.

El módulo de detección de color se encarga de generar un histograma asociado a la entrada visual obtenida por la cámara. De forma similar, el módulo de detección de movimiento calcula continuamente las diferencias entre los datos visuales anteriores y los actuales. Los preprocesadores de sensor usan la salida de estos módulos para generar los perceptos simples que se envían al ETC de la capa física. Los perceptos simples contienen específicamente los siguientes parámetros contextuales:

- **Marcas de tiempo**. Se registran dos marcas de tiempo diferentes en los perceptos simples. La primera marca de tiempo se establece cuando los datos sensoriales se recogen del sensor. Normalmente, esta marca de tiempo la asigna directamente el servicio de sensor de la capa de servicios sensoriomotores de CERA-CRANIUM. La segunda marca de tiempo se asigna cuando el percepto se procesa en el ETC. El intervalo de tiempo entre estas dos marcas puede ser significativo cuando un sensor no para de proporcionar datos y el servicio de sensor no tiene capacidad para procesar todos los datos en tiempo real. De hecho, este intervalo de tiempo se usa para descartar datos sensoriales muy

antiguos que ya no pueden ser significativos para el estado actual del robot. De forma análoga, también se registran dos marcas de tiempo en las acciones simples. La primera se asigna cuando se crea la acción y se encola en el secuenciador de la capa física. La segunda marca se asigna cuando la acción va a ejecutarse. El intervalo de tiempo entre estas dos marcar puede usarse para abortar la ejecución de acciones muy antiguas.

- **Índice *j***. Para la representación de los parámetros de localización tanto de los perceptos como de las acciones se ha decidido usar el centro de masas del robot como origen de un sistema de referencia egocéntrico. El índice *j* está compuesto de vectores que describen la posición y el tamaño estimado del estímulo que ha dado origen al percepto (ver Apartado 4.5). Estos vectores se calculan de forma diferente dependiendo de la naturaleza de cada percepto (el cálculo para el sonar y los parachoques se describe en el Apartado 7.4.2 y el cálculo para los perceptos visuales se describe a continuación).
- **Histograma de color**. Se asigna un histograma de color a cada paquete de datos proporcionado por el servicio de sensor de la cámara (que se corresponde con un segmento de un fotograma). Este histograma representa la frecuencia de los componentes de color que aparecen en el mapa de bits proporcionado. Este parámetro sólo se puede establecer para la información visual. En cualquier caso, cualquier otro tipo de procesamiento más complejo de la información visual no se definiría como parámetro contextual, sino que sería realizado por un procesador especializado y su aplicación estaría limitada probablemente a la región de la fóvea.
- **Índice *m***. Al aplicar la detección de movimiento sobre un paquete de información visual se genera un vector de movimiento llamado índice-m, cuyo valor escalar es cero cuando no se detecta movimiento alguno. Aunque el movimiento se puede detectar usando otras modalidades sensoriales como el sonar, en *CC-Chaser* se ha decidido usar exclusivamente la visión. Aunque este parámetro pueda indicar movimiento en los perceptos simples debido al movimiento relativo del propio robot, este efecto ha de ser corregido a la hora de construir los perceptos complejos.

Las marcas de tiempo se calculan fácilmente usando el cronómetro interno de CERA-CRANIUM. Sin embargo, el cálculo de los vectores *j* requiere mayor elaboración (particularmente en el caso de sensores móviles). El cálculo de los referentes espaciales *j* para los perceptos simples correspondientes a los parachoques y el sonar se describe en el Apartado 7.4.2. El cálculo del vector *j* correspondientes a un segmento de información visual se describe a continuación.

A cada segmento se le asigna un vector referente *j* que corresponde a la posición relativa del centro geométrico de ese segmento con respecto al campo de visión completo. Como la posición de la cámara es fija es fácil estimar la coordenada X relativa (posición izquierda/derecha relativa al robot) conociendo SH, la distancia desde el eje vertical de la cámara (posicionado en el centro del campo de visión) al centro del segmento. La Figura 91 ilustra un ejemplo de la entrada visual segmentada en la que el campo de visión se ha dividido en 64 regiones. La imagen muestra el vector referente *j* para el segmento $S_{2,4}$. La estimación de la distancia a la que se encuentra el estímulo visual correspondiente implica la aplicación de modelos más complejos, como el uso de visión estereoscópica. En la implementación *CC-Chaser* la distancia a los objetos se estima usando exclusivamente los perceptos correspondientes al sonar.

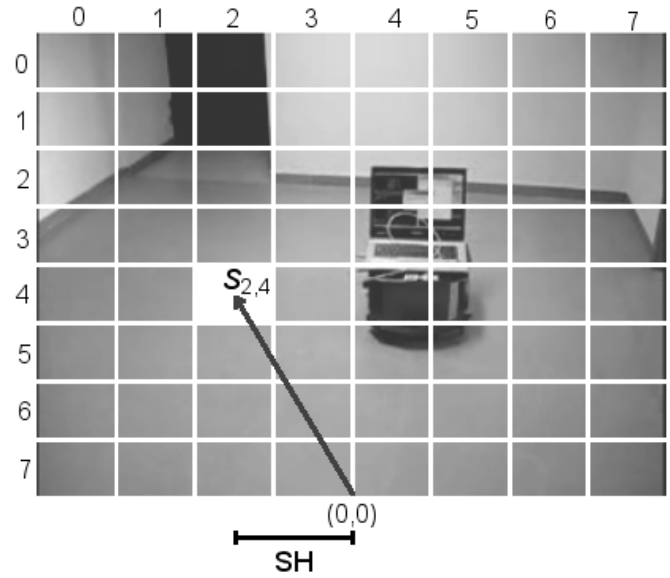


Figura 91. Vector referente *j* de un segmento de datos visuales.

La Figura 92 muestra un ejemplo en el que se indican los vectores referentes *j* sólo para aquellos segmentos en los que se ha detectado algún rasgo destacado. En el caso de *CC-Chaser* el objetivo es seguir a los objetos rojos, por lo que los perceptos indicados tendrán más probabilidades de formar parte de los contextos activos (ya que sus histogramas de color indican una alta frecuencia en el rojo).

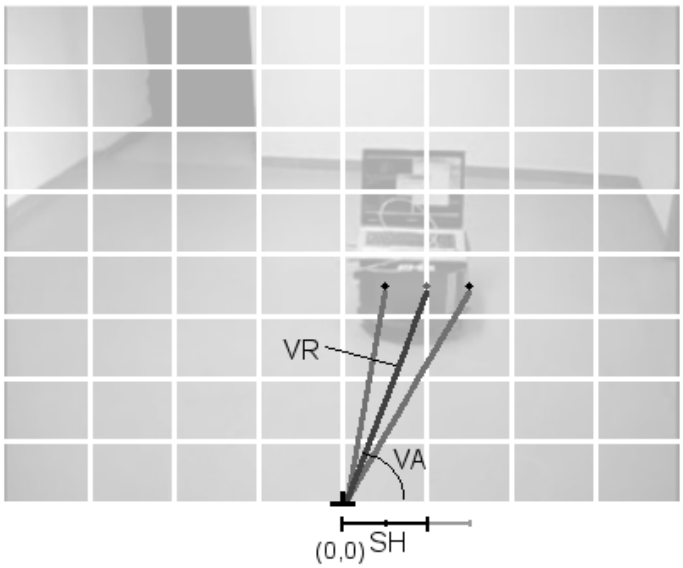


Figura 92. Vectores referentes j asociados a perceptos visuales.

La región correspondiente a la fóvea puede estar formada por más de un segmento visual. Siguiendo la filosofía de funcionamiento de CERA-CRANIUM, inicialmente cada segmento de un fotograma da lugar a un percepto simple, por lo que cada vez que

se obtiene una imagen de la cámara entran 64 perceptos simples al ETC de la capa física. Sin embargo, al aplicar los contextos para la generación de perceptos complejos, el correspondiente procesador de agregación seleccionará sólo aquellos perceptos simples relacionados con el contexto activo. En el ejemplo ilustrado en la Figura 92 se crearía un único percepto complejo cuyo vector de referencia *j* principal sería el denominado *VR*. Este vector se calcula como una combinación de los vectores *j* de los segmentos contiguos que lo forman. Concretamente, se usa la Ecuación 13 para calcular el vector de referencia *j* principal del percepto visual, donde *CH* es la altura a la que está situada la cámara, *VR* es la distancia desde el origen del sistema de referencia óptico al centro del percepto y *VA* es el ángulo que forma el vector con respecto al eje horizontal.

$$j = (X, Y, Z) = \begin{pmatrix} VR * Sen(VA) \\ CH \\ ? \end{pmatrix}$$

Ecuación 13. Calculo del vector de referencia *j* de un percepto visual complejo.

Es importante resaltar que el cálculo de los parámetros de contextualización es rápido y no requiere un procesamiento pesado. Una de las ventajas de tener un mecanismo de atención es el ahorro de la capacidad computacional. Este principio se respeta manteniendo parámetros de contextualización simples. Cuando los parámetros de contextualización son el tiempo y la localización (*t* y *j*) todos los perceptos, independientemente de su modalidad, se pueden comparar unos con otros para establecer contextos. Los contextos que se forman siguiendo este método tienen un significado claro. Por ejemplo, “todos los objetos al alcance del robot” sería un contexto formado por la aplicación del criterio de localización y estimando que la distancia del robot al objeto está por debajo de un umbral; o “todos los sucesos que tuvieron lugar entre hace 5 y 10 minutos” sería un contexto formado por la aplicación del criterio tiempo teniendo en cuenta que las marcas de tiempo de los perceptos coinciden con el intervalo indicado. Se pueden usar más criterios de forma análoga para construir contextos más específicos, que incluso podrían no incluir todas las modalidades sensoriales disponibles. Este es el caso de los criterios de movimiento y color que se han usando en *CC-Chaser*.

El mecanismo de contextualización de CERA-CRANIUM soporta la composición jerárquica, por lo que los perceptos complejos se pueden formar en base a la combinación de:

- Dos o más perceptos simples.
- Dos o más perceptos complejos.
- Cualquier combinación de perceptos simples y complejos.

Para poder ensamblar perceptos coherentes se tiene que establecer una política de prioridades en cuanto a la formación de perceptos complejos. La prioridad más alta en el proceso de contextualización corresponde a la formación de perceptos complejos a partir de perceptos simples de la misma modalidad. La salida de este primer paso es un conjunto de perceptos complejos monomodales. A su vez, estos perceptos complejos multimodales pueden tomar parte en la formación de perceptos complejos multimodales (ver Figura 93).

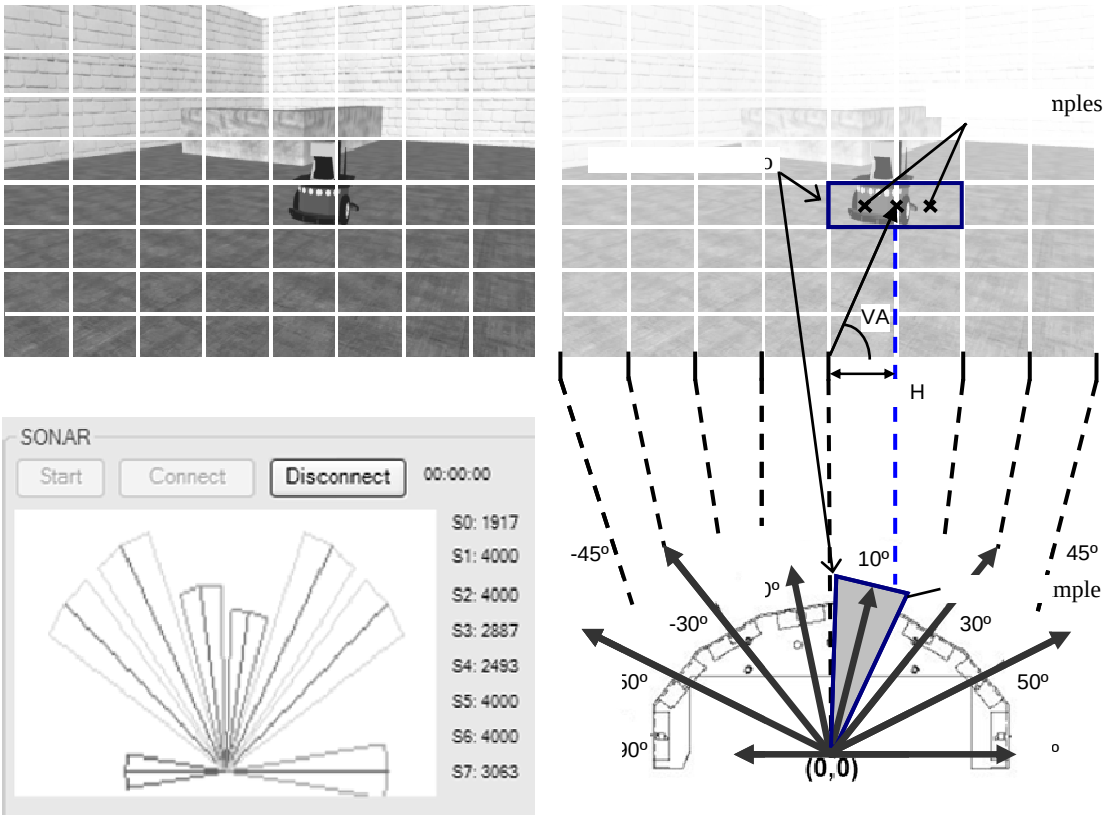


Figura 93. Formación de un percepto complejo multimodal.

Una posible aplicación de la fusión de información sensorial multimodal es la desambiguación o resolución de conflictos en cuanto a datos sensoriales contradictorios. En los dominios de aplicación planteados para el control del robot Pioneer puede obtenerse información contradictoria cuando el sonar no logra detectar un obstáculo (cuando se trata por ejemplo de una esquina que desvía las emisiones de ultrasonidos y las medidas del sonar no son realistas). En este tipo de situaciones el robot puede llegar a colisionar con el obstáculo, creándose los perceptos de contacto correspondientes. Durante el proceso de formación de perceptos complejos se tiene que manejar la información potencialmente contradictoria, como en este ejemplo, en el que las notificaciones de contacto y las medidas de distancia del sonar que no concuerdan. Una estrategia simple pero efectiva es aplicar diferentes niveles de confianza a las modalidades sensoriales. En el caso de la implementación *CC-Chaser* se ha asignado más confianza a las notificaciones de contacto de los paragolpes que a las medidas de distancia del sonar.

Para la tarea de detección y seguimiento se crea un contexto activo en el que los criterios de contextualización de color y movimiento se establecen para activar los perceptos que tienen las propiedades conjuntas de ser rojos y estar en movimiento. Para el control del robot, la capa de misión de *CC-Chaser* se ha diseñado de tal forma que se creen comportamientos cuyo vector de referencia *j* esté orientado hacia los perceptos activos. Las acciones simples generadas subsecuentemente hacen que el robot se mueva siguiendo al objetivo.

El mantenimiento de una distancia constante con el objetivo se consigue gracias a las medidas de distancia obtenidas con el sonar. Como los perceptos simples de visión y sonar comparten los parámetros contextuales de localización (vectores *j*) se puede estimar la distancia al objetivo mediante la contextualización multimodal. De hecho, los perceptos complejos que representan al objetivo están compuestos tanto de perceptos simples visuales como de perceptos simples de sonar. Estos perceptos simples se asociaron gracias a su afinidad en cuanto a localización relativa. Esto significa que los perceptos complejos correspondientes al objetivo incluyen datos de distancia así como una estimación de la localización en el campo de visión.

La Figura 93 muestra un esquema de los datos visuales y de distancia tal y como se representan en la capa física cuando se calculan los parámetros contextuales *t* y *j*. Cuando los preprocesadores de sensor construyen los perceptos simples se incluyen las marcas de tiempo y los vectores de referencia *j*. Todos los perceptos simples generados se envían al ETC de la capa física donde se forman los perceptos complejos gracias a la aplicación de los contextos activos. En el escenario esquematizado en la Figura 93 está activo un contexto para los objetos rojos. La figura muestra un ejemplo de formación de un percepto complejo gracias a la aplicación del mecanismo de contextualización mencionado. Los perceptos simples que corresponden a los segmentos visuales $S_{5,5}$ y $S_{6,5}$ se seleccionan porque son relevantes en términos del criterio de contextualización del color. Dado que sus vectores de referencia *j* son contiguos, se forma un nuevo percepto complejo (visual) monomodal usando estos dos perceptos simples. Como se puede apreciar en la Figura 93 el vector j del nuevo percepto complejo monomodal apunta al centro geométrico del segmento formado como combinación de los segmentos visuales originales. En realidad, se puede apreciar que este vector *j* no apunta al verdadero centro del objetivo. Sin embargo, la aproximación es suficientemente buena para la realización de la tarea de persecución. Una vez formados los perceptos complejos monomodales se aplican los criterios genéricos de contextualización (tiempo y localización) entre los perceptos de diferentes modalidades.

La representación inferior derecha de la Figura 93 corresponde a los vectores *j* de las lecturas del sonar, incluyendo la del percepto simple indicado, que se corresponde con la lectura del transductor sonar colocado en la orientación de +10º. La proyección trazada descendentemente desde el percepto complejo visual hasta el percepto simple de sonar indica que ambos perceptos se asocian y forman así un nuevo percepto complejo multimodal. La asociación temporal es obvia. Sin embargo, la contextualización espacial entre los perceptos visuales y de sonar requiere un alineamiento paramétrico adicional dado que los sensores de diferentes modalidades presentan orientaciones y amplitudes de adquisición de datos específicas. Además, como se ha explicado anteriormente, en los perceptos visuales sólo se considera la coordenada X.

El campo de visión de la cámara utilizada es de 90º, mientras que la cobertura del anillo frontal de transductores sonar abarca 195º (incluyendo los ángulos muertos entre las emisiones cónicas de ultrasonidos). Por lo tanto, sólo los perceptos originados en los 90º centrales de cobertura de sonar se tienen en cuenta para la contextualización multimodal entre visión y sonar. Las líneas discontinuas trazadas en la parte derecha de la Figura 93 representan el alineamiento entre el eje horizontal visual y el intervalo angular central del sonar entre -45º y 45º. En este caso, el valor de SH en el percepto complejo visual corresponde al percepto de sonar originado en el transductor orientado a 10º. La medida representada en este percepto sonar particular (2493 milímetros) constituye directamente la estimación de distancia asignada al percepto complejo

multimodal (dado que el percepto visual por sí mismo no puede proporcionar estimación de distancia alguna).

Los resultados obtenidos utilizando el mecanismo de atención demuestran que la integración multimodal se realiza adecuadamente. Se ha usando un objetivo controlado manualmente y se ha observado que el robot simulado controlado por *CC-Chaser* es capaz de mantener el comportamiento de seguimiento (siempre y cuando el robot perseguido no realice maniobras de evasión intencionadas). En cualquier caso, el objetivo de la experimentación con la implementación *CC-Chaser* no es lograr un comportamiento de seguimiento robusto (para lo cual se necesitarían añadir procesadores especializados capaces de realizar tareas de reconocimiento de patrones mucho más precisas), sino demostrar que se puede realizar una fusión sensorial multimodal efectiva usando CERA-CRANIUM y una implementación parcial del modelo MC^3.

7.4.4 Aplicación de CERA-CRANIUM en los videojuegos de acción

7.4.4.1 Introducción

El desarrollo de personajes sintéticos para videojuegos que sean capaces de producir comportamientos similares a los de los humanos es un problema sin resolver y un gran reto. De hecho, el objetivo último en este campo de investigación es diseñar personajes sintéticos imposibles de distinguir de los humanos. En otras palabras, agentes capaces de pasar el Test de Turing (Turing 1950) (o más específicamente, una versión del Test de Turing adaptada a la evaluación de la credibilidad de los personajes de videojuegos [Livingstone 2006]).

La principal fuente de inspiración para el diseño de personajes sintéticos más creíbles son los modelos psicológicos de la cognición humana. Normalmente, estos modelos y las técnicas asociadas de Inteligencia Artificial se basan en aspectos parciales de los sistemas complejos reales que dan lugar al comportamiento humano. Las emociones, la planificación, el aprendizaje, la capacidad de "ponerse en el lugar del otro" (teoría de la mente), la capacidad de cambio de contexto y los mecanismos de atención son algunos ejemplos destacados de características que se consideran de forma aislada en los modelos de control clásicos basados en Inteligencia Artificial. Las arquitecturas cognitivas artificiales tratan de integrar varios de estos aspectos de forma eficiente en sistemas de control. Sin embargo, el diseño de este tipo de arquitecturas no es simple. En el presente trabajo se argumenta que los esfuerzos de investigación que se están llevando a cabo en el nuevo campo de la Conciencia Artificial podrían contribuir a lidiar con la complejidad inherente a estos modelos y proporcionar un marco útil para el diseño de personajes sintéticos más atractivos para los jugadores de videojuegos. Concretamente, se han realizado dos implementaciones de la arquitectura CERA-CRANIUM adaptadas al control de personajes sintéticos (también conocidos como "*bots*") en videojuegos de acción en primera persona – esta serie de implementaciones se han denominado "*CC-Bot*":

- **CC-Bot1:** Implementación de CERA-CRANIUM basada en RDS capaz de controlar un bot en el juego UT2004 a través de una versión traducida a .NET de Pogamut 2 (ver Figura 73).

- **CC-Bot2:** Implementación Java de CERA-CRANIUM que usa directamente el entorno Pogamut 3 para controlar el bot en el juego UT2004.

Las implementaciones clásicas de personajes de juegos con Inteligencia Artificial, como por ejemplo los fantasmas del Comecocos o las naves alienígenas del juego *Space Invaders*, eran relativamente simples de programar y realmente no se aplicaban técnicas avanzadas de Inteligencia Artificial. Sin embargo, con la evolución de los videojuegos hacia entornos virtuales más complejos y guiones más elaborados, los personajes requieren comportamientos mucho más realistas. Los comportamientos pre-programados son aceptables hasta cierto punto en determinadas situaciones, pero en los juegos realistas de última generación los usuarios esperan encontrarse con oponentes artificiales que se comporten como lo haría un humano. Cuando se compara el comportamiento de estos personajes sintéticos con el comportamiento de jugadores controlados por humanos, el usuario normalmente considera que los bots son decepcionantes.

Los bots actuales diseñados con técnicas de Inteligencia Artificial pueden llegar a ser inteligentes en algún sentido, pero no llegan a alcanzar el comportamiento que desarrolla un jugador humano (como demuestran por ejemplo los resultados de la competición BotPrize (Hingston 2009) – una competición basada en el Test de Turing). Hasta la fecha, jugar con otros humanos es generalmente más realista y cautivador que jugar con bots.

Usar arquitecturas cognitivas inspiradas en modelos de la conciencia podría ser un enfoque efectivo para lidiar con la complejidad de este problema. Adicionalmente, la experimentación en el dominio de los videojuegos puede ser útil para la investigación en Conciencia Artificial.

7.4.4.2 Encarnación y Situacionismo en videojuegos

La encarnación física (embodiness) y el situacionismo en el mundo real (*situatedness*) se consideran factores esenciales en la producción de la conciencia. Algunos investigadores argumentan que la encarnación y el situacionismo son sólo posibles en el caso de los agentes físicos (Prem 1997). Sin embargo, en la presente tesis se mantiene la hipótesis de que los agentes software, como los bots de los videojuegos, también pueden tener un cuerpo y estar situados en el entorno (esta misma posición también se acepta para otros agentes basados en la conciencia, como en IDA (Franklin 2005a).

La encarnación significa tener un cuerpo dotado de sensores y actuadores, los cuales permiten el acoplamiento estructural del agente en su entorno. Por lo tanto, un agente de un videojuego está encarnado en ese sentido (o “encarnado virtualmente” (Goertzel 2008)), ya que obtiene datos sensoriales a través de sus sensores software y también puede realizar acciones usando sus actuadores software (ver Figura 94).

El situacionismo se refiere a la interacción causal con el mundo. En el caso de un personaje de videojuego o bot, la interacción con el mundo simulado altera el entorno de juego, que a su vez influye en el agente. Al mismo tiempo, los jugadores humanos pueden interactuar con el bot, con lo que el mundo real también es afectado causalmente por las decisiones del bot (análogamente, la entrada sensorial del bot viene causalmente determinada por las acciones que realiza en el mundo real el jugador humano).

En este trabajo se ha centrado la investigación en juegos online multijugador en los que se pueden emplear personajes sintéticos. Ejemplos de este tipo de videojuegos son los juegos de acción en primera persona (FPS), los juegos de rol (RPG) y los juegos online multijugador masivos (MMO). Las características comunes de todos estos juegos es que la acción tiene lugar de forma local o remotamente, a través de redes de ordenadores, y que se puede elegir jugar contra otros humanos o contra bots.

El diseño de sistemas de control inteligentes para bots es en esencia equivalente a la tarea de diseñar sistemas de control inteligentes para robots autónomos. Sin embargo, se pueden identificar algunas diferencias importantes:

- Los robots, a diferencia de los bots, poseen un cuerpo físico.
- Los robots, a diferencia de los bots, interactúan directamente con el mundo físico.
- Los robots, a diferencia de los bots, tienen que manejar niveles mucho más altos de ruido e incertidumbre.

En el contexto del presente trabajo se consideran estas características como beneficios en vez de problemas en cuanto a la aplicación de arquitecturas de control basadas en la conciencia. De hecho, las características de los videojuegos y otros entornos de simulación, como el visto en los Apartados 7.4.2 y 7.4.3, los convierte en entornos ideales para centrarse en el control de alto nivel, evitando los típicos problemas asociados a los mecanismos físicos.

7.4.4.3 Arquitecturas cognitivas y videojuegos

Una de las principales razones por las que los bots actuales no son capaces de comportarse de forma análoga a un humano es que sus sistemas de control no poseen la suficiente “potencia cognitiva”. Es decir, no son capaces de integrar el conjunto total de las capacidades cognitivas que normalmente se encuentran en un humano adulto (en el Apartado 5.3.2 hay una lista exhaustiva de funciones cognitivas). En general, hay una falta de combinación efectiva de capacidades como la atención, el aprendizaje, la planificación, etc. Aunque se podrían implementar individualmente modelos computacionales limitados de estas habilidades cognitivas usando técnicas actuales de IA, se necesita un diseño de una arquitectura cognitiva flexible para poder integrar de forma efectiva todas estas funcionalidades.

A continuación se describe la arquitectura software típica de un bot y donde podría localizarse la implementación de las habilidades cognitivas mencionadas anteriormente. La mayoría de las implementaciones de bots se basan en técnicas diferentes como las máquinas de estados finitos, la lógica difusa, las redes de neuronas artificiales, los árboles de decisión, los algoritmos evolutivos, etc. Las arquitecturas software empleadas actualmente en videojuegos se usan como medio para integrar los componentes principales del bot: animación del personaje, movimientos de bajo nivel, coordinación de equipos, habilidades de combate y selección de acciones (Tozour 2002).

El trabajo presentado en esta tesis se centra en el control de alto nivel del bot, dejando de lado los aspectos relacionados con el control de bajo nivel como la animación del cuerpo del bot, la cinemática inversa y la ejecución de acciones básicas

(como saltar o disparar). Normalmente se implementa sobre otros subsistemas de control una capa más alta donde se definen los objetivos del bot y se genera el comportamiento final del mismo (ver Figura 94). Esta capa de control genérica representa el nivel de control más alto. Es en este *Controlador del Comportamiento* donde se determina el comportamiento final del bot. Mientras que el módulo de movimiento determina cómo se puede mover el bot de forma efectiva de un punto a otro, el Controlador del Comportamiento se encarga de decidir hacia donde hay que moverse. Como regla general, los niveles más altos envían comandos abstractos a los niveles más bajos, los cuales son responsables de generar la secuencia de acciones concreta.

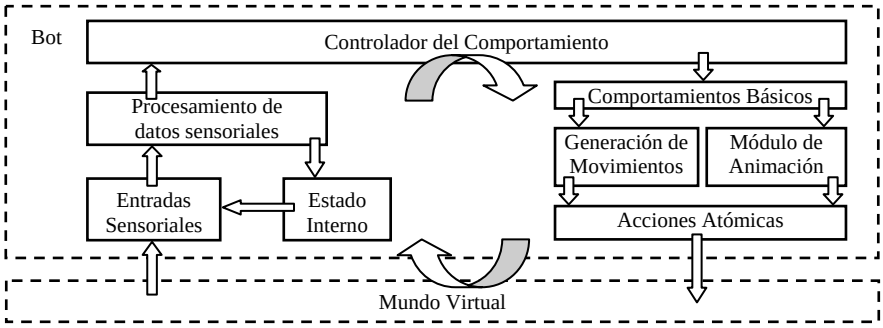


Figura 94. Arquitectura genérica de control de un bot.

Como los bots están diseñados para alcanzar metas específicas de misión, se puede definir un repertorio de comportamientos pre-programados (ver módulo "Comportamientos Básicos" en la Figura 94). En este caso, es responsabilidad del nivel de control superior (el Controlador de Comportamiento en este caso) decidir qué comportamiento básico se activa en cada instante. Se han propuesto diveros enfoques para el diseño de Controladores de Comportamiento. Cuando se adoptan enfoques de inspiración biológica a esta capa de control superior se le suele llamar capa de control cognitivo.

En el contexto del presente trabajo, cuando se hace referencia a la arquitectura cognitiva que controla el bot se apunta al sistema embebido en la arquitectura de control general y que actúa como Controlador de Comportamiento. Por lo tanto, se puede distinguir entre dos niveles de arquitecturas: la arquitectura de control general del bot (tal y como se esquematiza en la Figura 94) y la arquitectura de control cognitiva, que se enmarca en la capa superior de la arquitectura general.

La tecnología actual de videojuegos multijugador proporciona un entorno rico donde humanos y bots pueden interactuar. Además, a diferencia de las tareas del mundo real, donde se usan agentes físicos, los videojuegos proporcionan un entorno donde el ruido y la incertidumbre no representan un reto adicional. Consecuentemente, estos juegos se han convertido en plataformas adecuadas y útiles para la investigación en sistemas cognitivos artificiales. Aunque el problema de la reproducción del comportamiento humano en personajes sintéticos ha sido abordado por múltiples autores (ver [Bauckhage et al. 2007] por ejemplo), la mayoría de los esfuerzos han ignorado la conciencia como posible fuente de inspiración para nuevos desarrollos. Sin

embargo, se han integrado arquitecturas de control complejas como SOAR en sistemas de control de bots (Laird, Newell & Rosenbloom 1987).

7.4.4.4 Diseño de los personajes sintéticos *CC-Bot*

Con el objetivo de explorar las posibilidades de aplicación de la Conciencia Artificial en el dominio de los videojuegos se han desarrollado dos implementaciones de CERA-CRANIUM capaces de controlar un bot (ver Apartado 7.4.4.1). Tanto para la implementación CC-Bot1 como para CC-Bot2 se han desarrollado los siguientes componentes específicos para generar versiones funcionales de CERA-CRANIUM en el entorno descrito anteriormente (todos estos componentes se incluyen en la caja CERA-CRANIUM representada en la Figura 73):

- Capa de servicios sensoriomotores de CERA.
- Procesadores especializados de la capa física de CERA.
- Procesadores especializados de la capa de misión de CERA.
- Modelo de estado interno y metas de la capa núcleo de CERA.

Como se ha decidido usar la representación del agente proporcionada por Pogamut (Pogamut 2 en el caso de CC-Bot1 y Pogamut 3 en el caso de CC-Bot2), ha sido necesario el desarrollo de una capa de servicios sensoriomotores específica para poder tratar las operaciones primitivas sensoriales y motoras disponibles (ver Figura 17). Esto permite integrar sin problemas las primitivas de alto nivel de Pogamut con la capa física de CERA. En otras palabras, las primitivas de Pogamut se traducen en lecturas de sensor de la capa física de CERA, las cuales a su vez se usan para construir los perceptos simples. La capa de servicios sensoriomotores también gestiona la diferencia de representaciones existente entre Pogamut y CERA-CRANIUM. Por ejemplo, algunas primitivas de sensor de Pogamut se invocan periódicamente para obtener el estado del agente y sólo en el caso de detectarse cambios significativos se notifica una lectura a la capa física.

Una vez que las lecturas de los sensores se envían a la capa física se crean los perceptos simples de forma análoga a como se describe en el Apartado 7.4.2.1. En este caso, para los vectores *j* se considera un espacio tridimensional (los parámetros de localización de las primitivas de Pogamut se propagan a la capa de servicios sensoriomotores de CERA para que más tarde, en la capa física, se puedan calcular los *índices-J*. Adicionalmente, también se distingue entre sensores exteroceptivos y propioceptivos. Por ejemplo, el nivel de salud del bot se representa como un percepto propioceptivo que indica el correspondiente estado del agente. Esto ayuda a mantener actualizado el modelo del estado interno que se maneja en la capa de misión.

Los perceptos simples típicos que se generan incluyen perceptos de localización del agente, perceptos de armas vistas en el entorno, perceptos de municiones vistas en el entorno, etc. Gracias al uso de la biblioteca Pogamut, la información representada en los perceptos simples tiene un nivel alto de significado. Por lo tanto, el procesamiento necesario para la construcción de perceptos complejos es mucho menor que en el caso de los escenarios descritos en los apartados 7.4.2 y 7.4.3, en los que se requiere realizar pre-procesamiento, fusión y clasificación de datos antes de poder asignar un significado de alto nivel a los perceptos. Según se van generando los perceptos simples y entran en

el ETC de la capa física, estos se pueden combinar usando contextos espaciotemporales. El principal objetivo de este proceso es obtener perceptos complejos significativos que a su vez se envíen a la capa de misión.

El estado que se mantiene en el ETC de la capa de misión es la representación que se usa para implementar las funciones cognitivas superiores. En este sentido, el ETC de la capa de misión puede considerarse como la memoria de trabajo del bot cognitivo, siendo el foco de este ETC el que almacena los contenidos explícitos de la memoria. Es en este escenario, donde las diferentes visiones del mundo percibidas por el agente compiten por tener un nivel de activación alto. Sólo el percepto ganador en cada caso consigue llegar a la capa núcleo de CERA para participar en el proceso de volición. Es decir, para ser tenido en cuenta en el sistema basado en reglas que hay en la capa núcleo.

En las implementaciones *CC-Bot* se han considerado los siguientes tipos de perceptos de misión: "*Enemigo Acercándose*", "*Enemigo Huyendo*", "*Enemigo Atacando*", "*Estoy en Problemas*", "*Enemigo Destruido*", etc. (ver Tabla 9). Estos perceptos o representaciones parciales del mundo pueden verse como posibles opciones de lo que realmente está pasando en el juego. El mecanismo de contextualización descrito en el Apartado 4.6 (cuyo funcionamiento se ha ilustrado en detalle en el Apartado 7.4.2.1) se aplica en los ETC para construir y seleccionar periódicamente estos perceptos.

Tabla 9. Perceptos implementados en CC-Bot1 y CC-Bot2.

Perceptos	**Descripción**	**Capa**	**CC-Bot1**	**CC-Bot2**
Obstáculo	Representa la posición relativa de un obstáculo.	Física		**X**
Impacto	Representa la posición relativa de un golpe físico.	Física	X	
Daño	Representa una pérdida de salud del bot.	Física	X	**X**
Moviéndose	Representa la dirección y velocidad del movimiento.	Física		**X**
Mirando	Representa la orientación de la mirada.	Física		**X**
Novedad	Representa el tipo y la localización de un suceso inesperado.	Física		**X**
Disparando	Representa el tipo de disparo que el bot está realizando.	Física	X	**X**
Posición Alcanzada	Representa si el bot ha alcanzado una posición determinada.	Física		**X**
Aparición Jugador	Representa la posición de un nuevo jugador detectado en el campo de visión.	Física		**X**
Desaparición Jugador	Representa la desaparición de un jugador que antes estaba visible.	Física		**X**

Perceptos	Descripción	Capa	CC-Bot1	CC-Bot2
Reinicio	Representa un reinicio del juego después de una muerte.	Física		**X**
Atasco	Representa una condición de dificultad de movimiento.	Física		**X**
Arma	Representa las características de un arma.	Misión	X	**X**
Enemigo Acercándose	Representa un jugador enemigo que se acerca.	Misión		**X**
Enemigo Huyendo	Representa un jugador enemigo que se aleja.	Misión		**X**
Enemigo Atacando	Representa un enemigo que está disparando al bot.	Misión	X	**X**
Estoy en Problemas	Representa una situación que puede llevar a la muerte.	Misión		**X**
Enemigo Destruido	Representa la destrucción de un enemigo.	Misión		**X**

Al igual que las áreas especializadas del cerebro humano que son capaces de combinar varias señales aferentes para determinar la presencia de ciertas condiciones, los procesadores especializados de la capa de misión combinan los perceptos simples y los perceptos complejos obtenidos al observar el mundo virtual del videojuego. Por ejemplo, la presencia en el ETC de misión de los siguientes perceptos simples y complejos: "*Estoy Siendo Herido*" y "*Enemigo Acercándose*" podría desencadenar la activación del procesador especializado "*Detector de Ataques Enemigos*". Esto causaría probablemente la generación de un percepto de misión "*Enemigo Atacando*".

La generación de un comportamiento rápido y adaptativo se gestiona gracias a la aplicación de contextos y un sistema basado en reglas inspirado en las emociones. La capa núcleo envía comandos de contexto dirigidos hacia las regiones (*índices-CJ*) donde los perceptos de misión ganadores están situados. Los comportamientos reactivos más básicos consisten en girar para centrar la acción en frente del agente cuando la emoción asociada es neutral, acercarse a los sucesos (perceptos de misión) evaluados como positivos e incrementar la distancia que separa al bot de los sucesos u objetos evaluados como negativos.

La evaluación emocional de los perceptos de misión se basa en un conjunto de metas establecidas en la capa núcleo. En general cualquier percepto que indique que se está realizando un progreso en la consecución de un objetivo se considera positivo. Si se considera que un perceptos supone un inconveniente (como ser atacado), se le asigna un valor emocional negativo.

A continuación se describe la evaluación experimental realizada con las dos implementaciones de CERA-CRANIUM realizadas para el control de bots (CC-Bot1 y CC-Bot2).

7.4.4.5 Evaluación de CC-Bot1

Para evaluar la credibilidad de los bots controlados por la arquitectura CERA-CRANIUM se los ha enfrentado con otros bots diferentes desarrollados para UT2004 (ver Tabla 10). Las pruebas consisten en partidas *deathmatch* contra jugadores humanos, bots basados en reglas sin capacidades de aprendizaje y bots que implementan variantes del algoritmo de aprendizaje Q-Learning (González López 2009).

Tabla 10. Puntuación media de diferentes Bots en UT2004.

Jugador	Puntuación Media
Bot Sistema Basado en Reglas	19.2
Bot Q-Learning	5.2
Bot Híbrido Q-Learning y Sistema Experto	11.3
CC-Bot Adaptativo	18.3

La evaluación de la credibilidad es un problema complejo y muy subjetivo. Evaluar las capacidades de aprendizaje u otros indicadores de rendimiento sería mucho más fácil. En ese caso, se podrían usar medidas objetivas, como la puntuación del juego (ver Tabla 10). En general, estimar hasta qué punto se está alcanzando un comportamiento humano en un agente es un reto. Se han propuesto algunos enfoques interesantes (Sloman 2007, Harnad, Scherzer 2008), pero no son de aplicación práctica en este dominio.

Como solución al problema de la evaluación de la credibilidad en personajes de videojuegos se creó la competición 2K BotPrize, un test de Turing adaptado al juego UT2004 (Hingston 2009). La implementación CC-Bot1 participó en la segunda edición (2009) de esta competición sin logar clasificarse para la final, por lo que no se dispone de datos de la evaluación de los jueces. Sin embargo, en la tercera edición (2010) de la competición 2K BotPrize el bot controlado por la nueva implementación CC-Bot2 logró alcanzar la victoria al ser declarado por los jueces el bot de comportamiento más humano. Los detalles de describen en el siguiente apartado.

7.4.4.6 Evaluación de CC-Bot2

La implementación CC-Bot2 es una evolución de CC-Bot1 que usa una plataforma totalmente distinta (Java) y que solventa problemas técnicos existentes en la implementación original provocados por incompatibilidades en la integración de los diversos componentes, como RDS (basado en .NET) y Pogamut (basado en Java). Asimismo, CC-Bot2 emplea una nueva versión de Pogamut (versión 3) y cuenta con nuevos y mejorados procesadores especializados[13] (ver Tabla 11).

[13] Los detalles de la implementación detallada de los procesadores de CC-Bot2 se encuentran disponibles en http://www.conscious-robots.com/cera-cranium/

Tabla 11. Procesadores especializados de CC-Bot1 y CC-Bot2.

Procesador	**Capa**	**Tarea**	**CC-Bot1**	**CC-Bot2**
AttackDetector	Física	Detectar condiciones compatibles con un ataque enemigo (decremento en el nivel de salud no achacable a otras causas, presencia de fuego enemigo, etc.).	X	X
AvoidObstacle	Física	Generar un comportamiento simple de navegación para esquivar un obstáculo.	X	X
BackupReflex	Física	Generar un movimiento simple de retroceso en respuesta a un impacto inesperado.		X
ChasePlayer	Misión	Generar un comportamiento complejo de persecución de otro jugador.		X
EnemyDetector	Misión	Detectar la presencia de un enemigo en base a determinadas condiciones, como la detección previa de un ataque y la presencia de otros jugadores disparando.	X	X
GazeGenerator	Física	Generar un movimiento simple de orientación de la mirada hacia en punto de atención.		X
JumpObstacle	Física	Generar un movimiento simple de salto para esquivar un obstáculo detectado.		X
KeepEnemiesFar	Misión	Generar un movimiento complejo de huída para maximizar la distancia a los enemigos detectados.		X
LocationReached	Física	Detectar que el bot ha llegado a la posición espacial a la que prentendía llegar.		X
MoveLooking	Física	Generar un movimiento complejo que compagina la locomoción con la orientación de la mirada.		X
MoveToPoint	Física	Generar un movimiento simple de movimiento hacia un punto determinado.	X	X
ObstacleDetector	Física	Detectar la presencia de un obstáculo que impide (o podría impedir el movimiento).		X
ObstacleDetectorNR	Física	Detectar la presencia de un obstáculo sin usar mecanismos de trazado de ratos (como hace el procesador anterior).	X	X
PlayerDissapearDetector	Física	Detectar que un jugador que estaba dentro del campo de visión deja de estar presente.		X
RandomNavigation	Física	Generar un movimiento complejo de navegación aleatoria por el mapa.	X	X
RestartDetector	Física	Detectar que el bot inicia de nuevo la ejecución después de un reinicio (por ejemplo debido a una muerte por fuego enemigo).		X
RunAwayFromPlayer	Misión	Generar un movimiento complejo para huír de determinados jugadores.		X

Procesador	Capa	Tarea	CC-Bot1	CC-Bot2
SelectBestWeapon	Misión	Seleccionar la mejor arma disponible en cada momento.		X
SelectEnemyToShoot	Misión	Seleccionar el enemigo al que conviene disparar en cada momento.		X
StuckDetector	Física	Detectar condiciones en las que el bot no es capaz de avanzar y los comandos de locomoción no surten el efecto esperado.		X

La evaluación de la implementación CC-Bot2 se ha basado en la competición 2K BotPrize[14] (edición 2010). La competición consiste en desarrollar un bot que sea indistinguible de un ser humano. El protocolo de análisis del comportamiento de los bots participantes se define de acuerdo a las siguientes reglas:

- Se usa una versión modificada (específica para la competición) del modo de juego "combate a muerte" en el juego UT2004.
- Los resultados finales de la competición se obtienen al recopilar los datos de tres sesiones de 1 hora de juego cada una realizadas en el servidor oficial de la competición.
- Cada sesión de juego consiste en cuatro combates de 15 minutos cada uno que tienen lugar en entornos simulados diferentes.
- En todos los combates realizados, se conectan al servidor tanto los bots participantes (agentes artificiales) como un número equivalente de jugadores humanos que actúan como jueces.
- Todos los jugadores, tanto los bots como los humanos, aparecen en el juego con nombres asignados aleatoriamente para asegurar el anonimato.
- Algunas funcionalidades normales del juego, como el chat, están deshabilitadas o modificadas para facilitar la realización de una competición justa.
- La principal misión de los bots es parecer lo más humano posible para ganar la competición.
- La principal misión de los humanos es juzgar lo mejor posible al resto de los jugadores.
- Los jugadores humanos cuentan con un tipo de arma especial (*Link Gun*) para realizar sus juicios acerca de los demás jugadores. Utilizando el fuego primario de este arma un juez puede matar a un bot de un solo disparo y obtener 10 puntos al mismo tiempo, mientras que si el disparo alcanza a un jugador humano será el juez el que muera instantáneamente perdiendo 10 puntos. Por otro lado, utilizando el fuego secundario del *Link Gun* un juez puede abatir a un humano instantáneamente y obtener 10 puntos, mientras que si el disparo alcanza a un bot el juez muere y pierde 10 puntos.

[14] http://www.botprize.org

- Los efectos descritos anteriormente sólo se producen la primera vez que se dispara a un jugador particular durante cada combate, es decir, sólo se puede juzgar a cada jugador una vez por combate.
- Los jueces tratarán de usar el *Link Gun* para hacer un juicio sobre otro jugador solamente cuando tengan la suficiente confianza de que han conseguido identificar si dicho jugador es un bot o un humano.
- Los votos recibidos por los bots serán los datos empleados para calcular la puntuación final de los competidores y el bot ganador.
- Se considerará que un bot pasa el test de Turing si es capaz de engañar a una mayoría de los jueces y también alcanza un ratio de humanidad mayor al "menos humano" de los jueces.
- Si ningún bot pasa el test de Turing, el bot ganador será el que alcance la mayor puntuación de votos, es decir, el que más veces ha conseguido hacerse pasar por jugador humano.

En la tercera edición de la competición 2K BotPrize se presentaron 11 equipos procedentes de diversas universidades de Estados Unidos, Japón, Reino Unido, Alemania, Dinamarca, Eslovaquia, Singapur y España. Utilizando las reglas descritas anteriormente como protocolo de evaluación, se obtuvieron los resultados resumidos en la Tabla 12 (ver también Figura 95).

Tabla 12. Resultados finales de la competición 2k BotPrize 2010.

Nombre del Bot	**Miembros del Equipo**	**Afiliación**	**Porcentaje de Humanidad**
Conscious-Robots ***(CC-Bot2)***	Raúl Arrabales Jorge Muñoz	Universidad Carlos III de Madrid. España	**31.81%**
UT^2	Igor Karpov Jacob Schrum Risto Miikulainen	University of Texas, Austin. Estados Unidos.	**27.27%**
ICE-2010	Akihiro Kojima Daichi Hirono Takumi Sato Seiji Murakami Ruck Thawonmas	Ritsumeikan University. Japón.	**23.33%**
Discordia	Casey Rosenthal Clare Bates Congdon	University of Souther Maine. Estados Unidos.	**17.77%**
w00t	Daniel Büscher Matthias Gorzellik Jannis Seyfried Björn Witt	Albert-Ludwigs Universität. Alemania.	**9.30%**

Aunque CC-Bot2 no consiguió pasar el test de Turing, consiguió ganar la competición al ser considerado por los jueces como el bot más humano. Es decir, haber sido capaz de engañar más veces a los jueces de la competición.

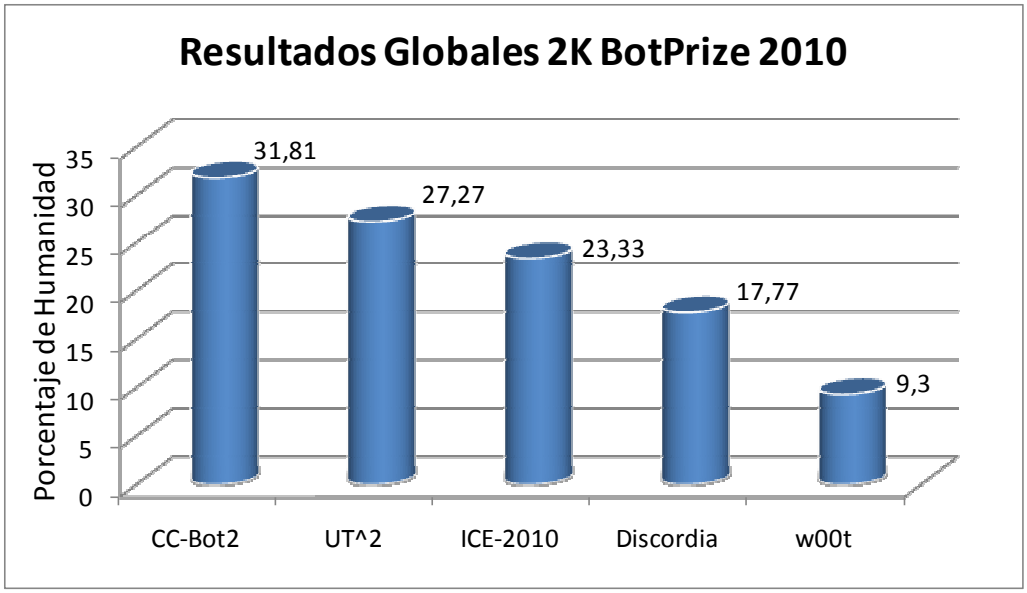
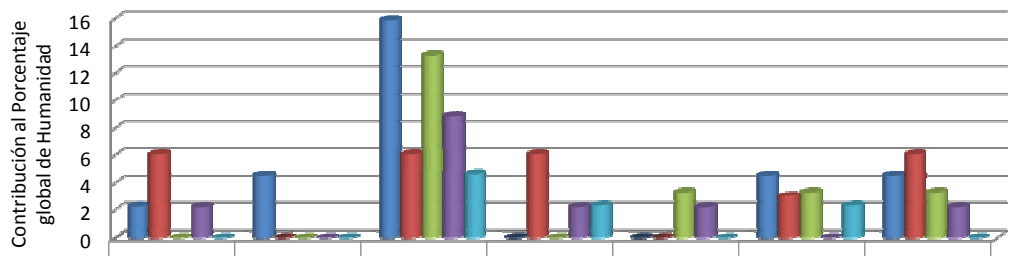


Figura 95. Resultados finales de la competición 2K BotPrize 2010.

CC-Bot2 consiguió la mejor puntuación: un ratio de humanidad del 31,81%, existiendo una diferencia pequeña entre el bot y el "menos humano" de los jugadores humanos (que obtuvo un 35,4% de humanidad[15]). La Figura 96 muestra la contribución detallada de cada juez al ratio de humanidad finalmente obtenido por cada bot. Para obtener estos valores se ha dividido el número de veces que cada juez ha identificado erróneamente a un bot como humano entre el número total de votos emitidos.



	Gordon	Nicola	Weber	Some	Frost	Apache	Swill
CC-Bot2	2,27	4,54	15,9	0	0	4,54	4,54
UT^2	6,06	0	6,06	6,06	0	3,03	6,06
ICE-2010	0	0	13,3	0	3,33	3,33	3,33
Discordia	2,22	0	8,88	2,22	2,22	0	2,22
w00t	0	0	4,65	2,32	0	2,32	0

Figura 96. Votos clasificando a bots como humanos en la competición 2K BotPrize 2010.

Como puede observarse en la Figura 96 tres de los jueces consideraron a CC-Bot2 como el más humano de todos los bots. Asimismo, CC-Bot2 se encuentra entre los bots

[15] Algunos jueces confesaron después de la competición que en determinados momentos intentaron hacerse pasar por bots para perjudicar a otros jueces.

evaluados en más ocasiones (44 evaluaciones recibidas – siendo la media 39 votos por bot), con lo que la evaluación global puede considerarse muy significativa.

7.4.5 Conclusiones

Los experimentos realizados con la arquitectura CERA-CRANIUM abarcan desde tareas simples de exploración o persecución con robots móviles hasta la generación de comportamientos complejos en personajes sintéticos. El objetivo de la experimentación realizada con robots móviles ha sido demostrar la capacidad de la arquitectura para integrar información multimodal en tiempo real y de forma efectiva. Sin embargo, en este contexto se han seleccionado comportamientos simples y entornos simplificados, puesto que configuraciones experimentales más complejas requieren capacidades avanzadas a nivel físico y más recursos hardware. Con el objetivo de estudiar las posibilidades de CERA-CRANIUM en cuanto a la generación de comportamientos más complejos, se ha seleccionado el entornos de los videojuegos de acción. Este ámbito, aunque poco realista desde el punto de vista físico, ha permitido la experimentación con comportamientos muy complejos e incluso la interacción directa con humanos. Asimismo, ha permitido el uso de entornos muy ricos en los que el agente puede demostrar sus capacidades cognitivas superiores, eludiendo los problemas derivados del ruido en los sensores y la incertidumbre en los actuadores.

Los prometedores resultados obtenidos por CC-Bot2 en la competición BotPrize demuestran que las líneas de investigación en Conciencia Artificial también pueden proporcionar ventajas a corto plazo en dominios de aplicación prácticos, como es el caso de los videojuegos. Concretamente, una de las funcionalidades que pueden explotarse en los sistemas cognitivos artificiales basados en la conciencia es la generación de comportamientos típicamente humanos. Esta característica, que ha sido demostrada significativamente en el caso de la implementación CC-Bot2, es de aplicación en otras áreas que también requieren de una efectiva interacción hombre-máquina. Por ejemplo, en el paradigma de la interacción humano-robot en el sector de la robótica doméstica.

Aunque CC-Bot2 no ha logrado pasar el test de Turing, es de esperar que sucesivas versiones de CC-Bot, diseñadas siguiendo la hoja de ruta sugerida en *ConsScale*, puedan superar este reto, al menos en el campo de los videojuegos de acción en primera persona. Como se analiza a continuación, CC-Bot2 es un agente de nivel 2 (*reactivo*), con un perfil cognitivo relativamente limitado en el contexto de *ConsScale*. La mejora de CC-Bot hasta alcanzar un nivel de desarrollo de la conciencia más avanzado, en torno a los niveles 4 y 5, proporcionaría en teoría un agente capaz de superar el test de Turing adaptado a videojuegos. Sin embargo, un test de Turing más exigente, requeriría un nivel de desarrollo cognitivo mucho mayor, equivalente al de un humano (nivel 10).

7.5 Experimentos de evaluación de agentes usando ConsScale

7.5.1 Introducción

El progreso en el campo de la Conciencia Artificial ha de ser valorado en base a las características que demuestran los nuevos modelos e implementaciones que se diseñan. En este apartado se describe la aplicación de la escala *ConsScale* (descrita en detalle en el Capítulo 5) para la evaluación del desarrollo cognitivo de algunos de los modelos más importantes en el campo de la Conciencia Artificial. La escala propuesta establece que el progreso en el desarrollo de la conciencia en agentes artificiales se puede evaluar examinando cómo las habilidades cognitivas asociadas con la conciencia se integran en los diseños existentes.

Los sistemas artificiales creados como parte de los esfuerzos actuales de investigación en Conciencia Artificial normalmente se inspiran en ciertos aspectos de los organismos biológicos. Sin embargo, los modelos específicos que se definen en base a esa inspiración y la forma concreta en la que los mismos se implementan pueden variar significativamente de un sistema a otro. Consecuentemente, no es fácil caracterizar las capacidades cognitivas de una arquitectura artificial de tal forma que se pueda poner en un contexto general, es decir, que se pueda comparar con otras implementaciones basadas en principios distintos. La raíz del problema reside en el hecho de que a menudo se mezclan de forma confusa diversos conceptos y perspectivas bajo la definición de conciencia (Block 1995).

En este trabajo la evaluación del desarrollo de la conciencia se centra en la identificación de las funciones cognitivas más importantes que se asocian con la conciencia. Se pretende responder a la pregunta de cómo estas funciones se pueden integrar de forma efectiva para construir un agente artificial equivalente a un humano en términos cognitivos. La definición de un marco genérico para la evaluación y la caracterización del desarrollo cognitivo de agentes artificiales puede ser beneficioso no sólo para poder realizar estudios comparativos de los modelos existentes, sino para la planificación y el establecimiento de nuevas líneas de investigación para la creación de futuras implementaciones. *ConsScale* es una propuesta pensada para definir este marco en base a criterios arquitecturales y del comportamiento observado.

Mientras que la mayoría de las propuestas actuales para medir la conciencia están basadas en medidas de integración de la información (ver Apartado 2.3), *ConsScale* se basa en las capacidades funcionales de alto nivel que presenta el sistema estudiado. Es importante remarcar que la propuesta presentada en esta tesis no ignora la importancia de la integración de la información como una propiedad clave de los organismos conscientes. De hecho se pretende caracterizar cómo la integración efectiva de la información y las sinergias entre funciones pueden contribuir a la generación de comportamientos normalmente atribuibles a seres conscientes. En resumen, mientras que medidas como Φ examinan exclusivamente las capacidades de integración de información del sistema (Tononi 2004, 2008), *ConsScale* pretende especificar – a nivel funcional – lo bien que esta integración se traduce en comportamiento adaptativo. La integración de la información y las medidas basadas en el comportamiento han de combinarse para obtener un método de evaluación completo para máquinas potencialmente conscientes.

La principal herramienta conceptual que se usa para la caracterización del nivel de desarrollo cognitivo de la conciencia es la definición de un conjunto parcialmente ordenado de habilidades cognitivas. Esta taxonomía, basada en el desarrollo de la conciencia, se usa para analizar, clasificar y comparar el perfil cognitivo tanto de modelos computacionales de la conciencia no implementados aún como de implementaciones existentes.

7.5.2 Evaluación de agentes artificiales usando *ConsScale*

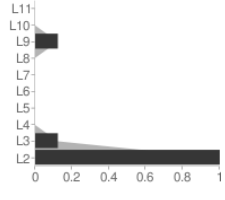
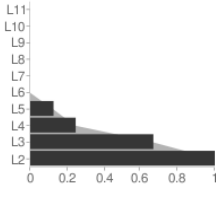
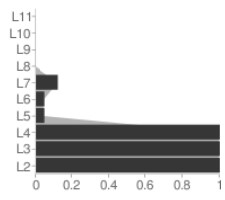
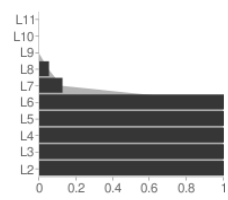
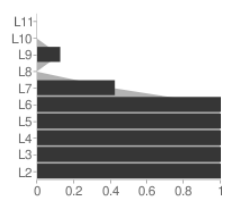
Con el objetivo de proporcionar una visión aproximada, pero ilustrativa, del estado del arte en Conciencia Artificial se han analizado los siguientes modelos e implementaciones usando el PSE (ver Apartado 5.7):

- **Eliza**: uno de los primeros programas para la interacción verbal escrita (*chatterbot*) (Weizenbaum 1966).
- ***CC-Bot2***: personaje sintético (bot) autónomo para Unreal Tournament 2004 implementado usando la arquitectura cognitiva CERA-CRANIUM (ver Apartado 7.4.4).
- **Arquitectura Mínima de Imaginación Funcional en CRONOS/SIMNOS (AMIF)**: implementación de un mecanismo de imaginación funcional que permite que un agente encarnado simule sus propias acciones y sus consecuencias sensoriales de forma interna, consiguiendo beneficios en la generación de su comportamiento gracias a este mecanismo (Marques, Holland 2009, Marques 2009) (ver Apartado 2.2.5).
- **Modelo LIDA**: LIDA es un modelo computacional de la cognición (que no se ha implementado completamente todavía) que está basado principalmente en la teoría del ETG (Ramamurthy et al. 2006, Franklin et al. 2007a) (ver Apartado 2.2.4.1).
- **Arquitectura Cognitiva de Haikonen**: arquitectura cognitiva basada en representaciones de señales distribuidas y Neuronas Asociativas de Haikonen (Haikonen 2007b) (ver Apartado 2.2.4.2).

La Tabla 13 resume los resultados obtenidos al aplicar el proceso simplificado de evaluación (PSE). La primera fila indica el nombre de los modelos analizados, la segunda fila contiene la lista de habilidades que cada modelo abarca, más abajo se indica el valor del CQS, luego se representan los perfiles cognitivos de cada modelo y finalmente se muestra el nivel conceptual de *ConsScale* que alcanza cada modelo de acuerdo al PSE.

Es conveniente resaltar que el PSE simplemente proporciona una aproximación de lo que podría ser el nivel real *ConsScale* de una implementación. La evaluación que se obtiene para modelos que no han sido implementados totalmente tendrá que ser confirmada en el futuro mediante la aplicación del PEE a las implementaciones correspondientes. Para las implementaciones o modelos que consideran un periodo de desarrollo cognitivo, la evaluación realizada considera el nivel potencial que pueden alcanzar al termino de su periodo de desarrollo.

Tabla 13. Resumen de los resultados de la aplicación del PSE.

ELIZA	***CC-Bot2***	**AMIF**	**LIDA**	**Haikonen**
$CS_{2,1}$; $CS_{3,3}$; $CS_{3,4}$; $CS_{3,5}$; $CS_{9,3}$.	$CS_{2,1}$; $CS_{3,1}$; $CS_{3,2}$; $CS_{3,3}$; $CS_{3,4}$; $CS_{3,5}$; $CS_{3,6}$; $CS_{4,1}$; $CS_{4,5}$; $CS_{5,2}$; $CS_{5,4}$.	$CS_{2,1}$; $CS_{3,1}$; $CS_{3,2}$; $CS_{3,3}$; $CS_{3,4}$; $CS_{3,5}$; $CS_{3,6}$; $CS_{3,7}$; $CS_{4,1}$; $CS_{4,2}$; $CS_{4,3}$; $CS_{4,4}$; $CS_{4,5}$; $CS_{5,4}$; $CS_{6,4}$; $CS_{7,1}$; $CS_{7,2}$; $CS_{7,5}$; $CS_{7,6}$.	$CS_{2,1}$; $CS_{3,1}$; $CS_{3,2}$; $CS_{3,3}$; $CS_{3,4}$; $CS_{3,5}$; $CS_{3,6}$; $CS_{3,7}$; $CS_{4,1}$; $CS_{4,2}$; $CS_{4,3}$; $CS_{4,4}$; $CS_{4,5}$; $CS_{5,1}$; $CS_{5,2}$; $CS_{5,3}$; $CS_{5,4}$; $CS_{5,5}$; $CS_{5,6}$; $CS_{6,1}$; $CS_{6,2}$; $CS_{6,3}$; $CS_{6,4}$; $CS_{6,5}$; $CS_{6,6}$; $CS_{7,1}$; $CS_{7,2}$; $CS_{7,3}$; $CS_{7,6}$; $CS_{8,1}$.	$CS_{2,1}$; $CS_{3,1}$; $CS_{3,2}$; $CS_{3,3}$; $CS_{3,4}$; $CS_{3,5}$; $CS_{3,6}$; $CS_{3,7}$; $CS_{4,1}$; $CS_{4,2}$; $CS_{4,3}$; $CS_{4,4}$; $CS_{4,5}$; $CS_{5,1}$;$CS_{5,2}$; $CS_{5,3}$; $CS_{5,4}$; $CS_{5,5}$; $CS_{5,6}$; $CS_{6,1}$; $CS_{6,2}$; $CS_{6,3}$; $CS_{6,4}$; $CS_{6,5}$; $CS_{6,6}$; $CS_{7,1}$; $CS_{7,2}$; $CS_{7,3}$; $CS_{7,4}$; $CS_{7,8}$; $CS_{9,3}$.
CQS: **0.19**	CQS: **0.51**	CQS: **12.37**	CQS: **102.27**	CQS: **114.39**
				
2 (*reactivo*)	2 (*reactivo*)	4 (*atencional*)	6 (*emocional*)	6 (*emocional*)

Observando los perfiles cognitivos obtenidos en la Tabla 13 se puede comprobar fácilmente que todos los modelos de Conciencia Artificial considerados siguen en gran medida el camino incremental indicado en los niveles jerárquicos de *ConsScale*. Este es de hecho el resultado esperado debido a las dependencias existentes entre las funciones que están situadas en diferentes niveles. Sin embargo, los modelos de Conciencia Artificial, así como los organismos biológicos, podrían presentar perfiles cognitivos "atípicos". Por ejemplo, el perfil asociado a una persona autista o a un agente artificial específicamente pre-programado para reconocer su propia imagen especular – $CS_{7,4}$ (sin cumplir con otras funcionalidades cognitivas de más bajo nivel). Normalmente, estos perfiles cognitivos atípicos aparecen en la naturaleza debido a una lesión cerebral o una enfermedad genética. Sin embargo, en el caso de los agentes artificiales la presencia de este tipo de perfiles podría indicar un diseño muy orientado a una tarea concreta o incluso la presencia de comportamientos pre-programados diseñados específicamente para engañar a otros métodos de evaluación cognitiva clásicos. Con el objetivo de evaluar de forma apropiada este tipo de sistemas y obtener una medida justa de la capacidad cognitiva global del agente, el índice CQS representa la dependencia cognitiva jerárquica y aplica una función de peso basada en la sinergia entre las diferentes funciones definidas ($CS_{i,j}$).

Eliza es básicamente un agente reactivo diseñado para detectar y seleccionar palabras clave en la entrada y, usando conjuntamente un script y ciertas técnicas de reconocimiento de patrones, generar una respuesta en forma de comunicación verbal precisa ($CS_{9,3}$). Aunque supuestamente este agente posee una de las habilidades cognitivas de más alto nivel, su índice CQS es bajo porque *ConsScale* premia la integración incremental de las habilidades cognitivas. En este caso particular, no importa lo bueno que sea el agente produciendo comunicaciones lingüísticas bien formadas. Si el contenido mental que se comunica no está creado por una combinación adecuada de habilidades cognitivas de más bajo nivel, la escala no puede considerar que el agente sea "cognitivamente avanzado".

CC-Bot cumple con algunas características de los niveles 3, 4 y 5. Sin embargo, se clasifica como un agente de nivel 2 porque *ConsScale* requiere el cumplimiento completo de los niveles inferiores para poder ser calificado como perteneciente a un

determinado nivel *i*. El índice CQS de un agente reactivo puro es 0.18. Sin embargo, la puntuación de *CC-Bot* (0.51) indica que el agente presenta capacidades cognitivas adicionales (como se puede observar en el perfil cognitivo asociado que aparece en la Tabla 13). Asimismo, *CC-Bot* está lejos de alcanzar el nivel 4, para el cual el CQS asociado sería de 12.21 o superior.

La arquitectura AMIF se ha clasificado como nivel 4. Siendo una arquitectura mínima esta propuesta es prometedora en cuanto a la posibilidad de alcanzar puntuaciones más altas en *ConsScale*. De hecho, la Arquitectura de Múltiples Pasos con Memora desarrollada por el mismo autor mejora este modelo incluyendo la función $CS_{5,1}$ (Marques 2009).

LIDA y la arquitectura de Haikonen presentan una escasa diferencia por lo que al PSE de *ConsScale* se refiere. No obstante, se requeriría una evaluación extensiva de las implementaciones completas (que a fecha de hoy no existen) para comprobar si dichos modelos pueden promocionar hasta el nivel 7 (*auto-consciente*). En el caso de la arquitectura de Haikonen se podría argumentar que el concepto de yo podría emerger como parte de la ontología privada que desarrolla el agente (Doan 2009b). Sin embargo, en el caso de LIDA, se tendrían que implementar los módulos específicos para implementar diversas facetas del yo autoconsciente (Franklin, *comunicación personal*).

7.5.3 Conclusiones

El análisis de los modelos de Conciencia Artificial seleccionados indica que los perfiles de *ConsScale* asociados a las correspondientes implementaciones – después de aplicar el PEE – tendrían buenas puntuaciones exclusivamente en la parte baja de los diagramas. En cualquier caso, después de haber aplicado el PSE se pueden extraer las siguientes conclusiones:

- Aunque las relaciones de dependencia entre los CS de niveles adyacentes (que se ilustran en la Figura 34) pueden ser objeto de controversia y requerir refinamientos ulteriores, queda claro que las funciones localizadas en los niveles más altos de la escala requieren la realización efectiva y la integración de las funciones asignadas a niveles inferiores. Por lo tanto, al menos desde una perspectiva de alto nivel, la jerarquía cognitiva propuesta en *ConsScale* coincide con las restricciones ingenieriles encontradas en las principales implementaciones actuales de modelos de Conciencia Artificial (así como con las dependencias equivalentes observadas en la filogenia biológica).
- De foma similar, el análisis de los sistemas seleccionados confirma que las habilidades de más alto nivel no constituyen requisitos para conseguir la realización de las funciones de más bajo nivel, justificando así las relaciones de orden ascendente definidas en *ConsScale*.
- Aunque aún es necesario realizar mucho trabajo para construir implementaciones reales capaces de cumplir con éxito las expectativas correspondientes a los niveles más bajos de *ConsScale*, el reto actual del campo de la Conciencia Artificial es crear nuevas criaturas artificiales cuyos perfiles cognitivos tiendan a rellenar la mitad superior del diagrama (mientras que también se mantienen puntuaciones altas en la mitad inferior).

Como se ha demostrado en el apartado anterior, los métodos de evaluación propuestos (PEE y PSE) son aplicables a implementaciones muy diferenciadas, permitiendo así la realización de análisis comparativos incluyendo cualquier modelo nuevo que pudiera aparecer en el campo de la Conciencia Artificial. Sin embargo, estos métodos tienen algunas desventajas: mientras que el PEE permite un análisis preciso de una implementación concreta, es por necesidad dependiente del dominio, por lo tanto sólo puede realizarse un análisis comparativo preciso si todos los sistemas involucrados están diseñados para funcionar en el mismo contexto.

Para poder comparar sistemas concebidos para ser usados en diferentes dominios – como es el caso del análisis descrito en el apartado anterior – es necesario aplicar el PEE. Desafortunadamente, este método sólo puede proporcionar una evaluación aproximada, la cual podría ser muy sensible a las interpretaciones arbitrarias de los CS en el contexto de cada uno de los sistemas analizados. Por ejemplo, se dice que un agente satisface $CS_{3,4}$ si es capaz de "*seleccionar adaptativamente información motora relevante*". Esto podría significar cosas muy diferentes dependiendo del contexto de aplicación concreto, involucrando mucho más esfuerzo de desarrollo en unos contextos con respecto a otros. Para el agente *CC-Bot*, $CS_{3,4}$ se traduce como "*la capacidad del bot de descartar acciones que no son adecuadas para su situación actual"*, como por ejemplo disparar a las paredes cuando se está huyendo de un enemigo. Para la arquitectura de imaginación funcional, $CS_{3,4}$ se podría traducir como la "*capacidad de pre-seleccionar acciones motoras dirigidas hacia la meta planteada*" (Marques, Holland 2009), como mover los brazos en la dirección del objeto que el robot tiene que derribar. Mientras que la implementación y evaluación de estos dos comportamientos puede implicar diseños, técnicas y esfuerzos muy distintos, su relevancia cognitiva es equivalente desde el punto de vista de *ConsScale*. En otras palabras, el PSE no tiene en cuenta la complejidad del dominio de aplicación de cada sistema. Por lo tanto, las métricas que se obtienen no son sensibles a la versatilidad de los agentes en cuanto a su aplicación a diferentes dominios. Como se ha mencionado anteriormente, para poder obtener medidas precisas y justas en ese sentido ha de usarse el PEE, restringiendo la evaluación a un único dominio de aplicación.

Otro problema relacionado con la evaluación particular de cada CS es que el cumplimiento de una determinada habilidad cognitiva se considera una propiedad binaria. Sin embargo, las implementaciones reales a menudo muestran una frontera difusa entre los comportamientos que podrían considerarse que satisfacen un determinado CS y los que no. Por ejemplo, en el caso de $CS_{7,4}$, se podría usar la prueba del espejo para evaluar a los agentes. Un resultado típico de esta prueba es que el robot es capaz de reconocerse a sí mismo en el espejo el 70% de las veces (Takeno, Inaba & Suzuki 2005). Traducir arbitrariamente este tipo de resultados a una propiedad binaria induce claramente ambigüedad en la medida. Este efecto negativo se podría disminuir considerando el cumplimiento parcial de los CS y/o la aplicación de lógica difusa en el cálculo de los parámetros L_i.

Aunque la medida propuesta no abarca directamente el problema de la conciencia fenomenológica, se podría argumentar la existencia de una posible correlación entre la aparición de estados fenomenológicos y la sinergia funcional existente en el sistema. Aunque la sinergia funcional podría no ser necesaria para la aparición de estados fenomenológicos, sí que es un requisito plausible para la formación de qualia, es decir, para la generación del contenido integrado de la experiencia consciente.

Teniendo en cuenta el planteamiento de Haikonen según el cual los qualia son la forma en la que la información sensorial se manifiesta en la mente (Haikonen 2009), es necesario considerar la capacidad de producción de qualia artificiales cuando se determina el nivel de conciencia de una máquina. En este sentido, la definición de *ConsScale* incluye un orden parcial para el desarrollo y la generación de los qualia:

$$CS_{4,5} < CS_{5,6} < CS_{6,6} < CS_{7,8} < CS_{8,6} < CS_{9,3} < CS_{10,1}.$$

Considerando este orden parcial perteneciente al poset (CCS, <) y los modelos analizados en el apartado anterior, se puede decir que el camino que *ConsScale* sugiere en cuanto a la generación de qualia encaja con la filosofía de los diseños actuales de Conciencia Artificial.

7.6 Experimentos realizados en el dominio de la Fenomenología Sintética

7.6.1 Introducción

Tal y como se ha descrito en el Capítulo 6, la naturaleza de los qualia y su posible generación en máquinas es un tema muy controvertido. Incluso es habitual que la propia existencia del concepto *quale* se ignore en los trabajos de Conciencia Artificial. En la experimentación realizada se ha adoptado un enfoque pragmático ante este problema usando la perspectiva planteada por la Fenomenología Sintética. En concreto, se ha explorado la generación de qualia visuales usando la arquitectura cognitiva CERA-CRANIUM (implementación *CC-Observer*). Se espera que los resultados preliminares obtenidos como parte de esta línea de investigación permitan una mejor caracterización e identificación de los qualia artificiales como los productos directos de la percepción consciente en máquinas. Adicionalmente, se ha utilizado el modelo MC^3 para caracterizar los procesos de percepción implícitos y explícitos.

Se ha usado el efecto de movimiento aparente descrito en el Apartado 6.5 para realizar un estudio práctico de la generación de experiencia visual sintética. Gracias a un subsistema de inspección interna, se han analizado los perceptos implícitos y explícitos generados por *CC-Observer* cuando se le presentan diferentes estímulos visuales. La inspección de los estados internos generados dentro de la arquitectura cognitiva permite analizar las posibles analogías con los procesos cognitivos humanos.

Aunque los resultados empíricos obtenidos usando máquinas no son directamente aplicables a los humanos, la construcción de modelos o simulaciones de los estados de la experiencia consciente puede potencialmente proporcionar nuevo conocimiento sobre el funcionamiento de la cognición humana. En este caso particular, como se ha usado una arquitectura inspirada en la Teoría del Espacio de Trabajo Global (ETG), se ha logrado proporcionar una explicación acerca de cómo la experiencia consciente se podría generar a partir de un ETG. Por lo tanto, además de la retroalimentación que se pueda obtener en términos de modelado de la cognición humana, este enfoque de Fenomenología Sintética también puede contribuir al diseño de máquinas funcionalmente conscientes.

Dado que el enfoque adoptado en este trabajo tiene un carácter marcadamente centrado en la funcionalidad, no se presentan hipótesis concretas respecto a la existencia

de estados fenomenológicos en la máquina. Se ha considerado que en principio no se requiere ningún sustrato especial para generar este tipo de estados. De hecho, la propuesta presentada se basa en un modelo computacional, que se puede implementar como una máquina virtual (en el sentido descrito en (Sloman, Chrisley 2003)) en un ordenador convencional basado en la arquitectura de Von Neumann. Consecuentemente, la hipótesis de trabajo planteada se basa en el modelo propuesto en el Capítulo 6, estableciendo que los qualia son los únicos contenidos de la experiencia consciente y tienen una funcionalidad clara. La posibilidad de que los procesos propuestos para la generación de qualia impliquen la generación de estados fenomenológicos asociados es un tema de debate abierto que excede el alcance de la presente tesis doctoral.

7.6.2 Implementación *CC-Observer*

CC-Observer es una implementación derivada de la arquitectura cognitiva CERA-CRANIUM en la que sólo se utiliza una pequeña parte de CERA. En *CC-Observer* se ha analizado en detalle el funcionamiento de CRANIUM (ver Figura 97) utilizando un visor para inspeccionar su estado interno. Con respecto a las modalidades sensoriales soportadas, *CC-Observer* sólo tiene capacidad para el procesamiento visual. Asimismo, no se usa ninguna característica de generación de comportamiento. Es decir, *CC-Observer* es un mero observador.

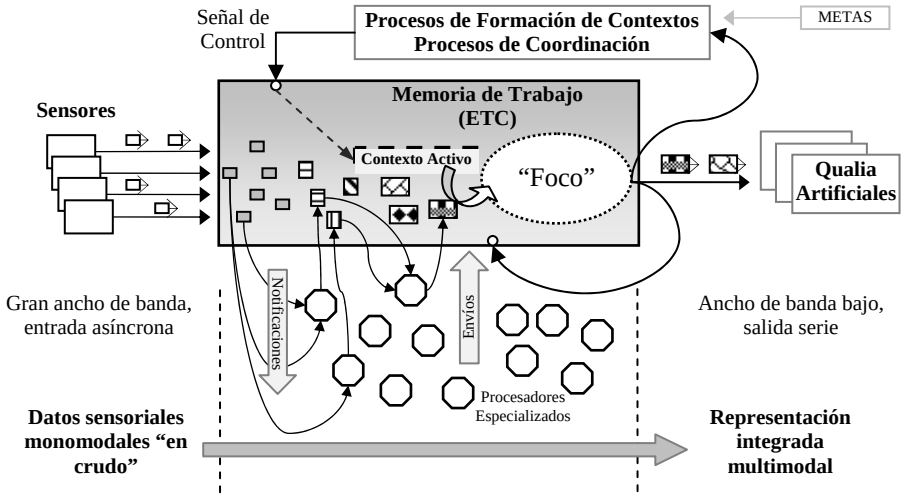


Figura 97. CRANIUM aplicado a la generación de qualia.

Desde el punto de vista de la generación de qualia, el ETC se puede ver como un sistema de procesamiento de la información que toma como entrada una gran cantidad de datos sensoriales "en crudo" y genera un resumen o información integrada, que se caracteriza por su aparición en serie y su tamaño mucho menor. El filtrado e integración de la información se logra gracias al procesamiento distribuido que realizan los procesadores especializados. Como se describe en el Capítulo 4, la forma en que estos procesadores tienen acceso a la información se regula mediante la aplicación de contextos ad-hoc ascendentes y contextos descendentes determinados en la capa núcleo.

Para los experimentos preliminares de generación de qualia artificiales se ha usado una configuración mínima de la arquitectura cognitiva (ver Figura 98). A diferencia de las implementaciones de CERA-CRANIUM descritas en los apartados anteriores, *CC-Observer* tiene un solo ETC, el de la capa física, no habiéndose implementado una capa de misión completa. Tampoco se han usado actuadores y la capa núcleo consiste en una implementación mínima para la definición de contextos. Asimismo, sólo se cuenta con una cámara digital como único sensor y el flujo de información perceptual está limitado a la modalidad visual.

Los mapas de bits con las imágenes obtenidas en la cámara se adquieren periódicamente gracias al servicio de sensor de la cámara localizado en la capa de servicios sensoriomotores de CERA (ver Figura 98). Los datos de sensores propioceptivos se adquieren gracias a servicios específicos localizados en la capa de servicios sensoriomotores. En el caso de la visión, se dispone de la posición relativa y orientación de la cámara. Como se ha descrito en la implementación *CC-Explorer* (Apartado 7.4.2), en este caso también se crean los perceptos simples combinando datos exteroceptivos y propioceptivos. Por ejemplo, un percepto simple de *CC-Observer* contiene un mapa de bits (imagen) más la posición y orientación de la cámara que se registro cuando se capturó dicha imagen. Para construir representación descriptivas se usan *índices-CJ* al igual que en *CC-Explorer* y *CC-Chaser*.

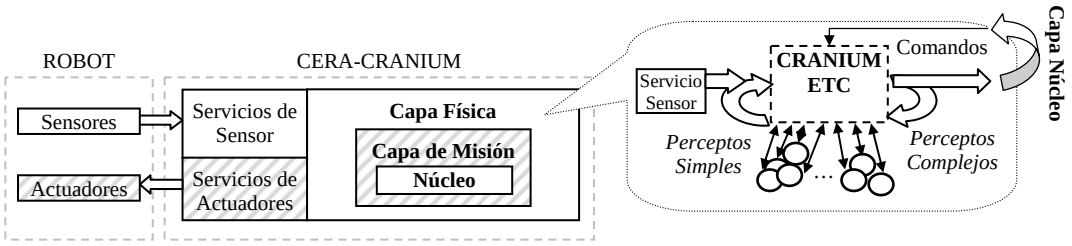


Figura 98. Configuración mínima de *CC-Observer*.

7.6.3 Configuración y metodología experimental

Para los experimentos relacionados con Fenomenología Sintética se ha usado la implementación *CC-Observer* descrita anteriormente. Aunque el sistema está diseñado para trabajar tanto con robots reales como simulados, se ha utilizado inicialmente un entorno simulado. El robot simulado es una versión modificada del Pioneer 3 DX en el que se han deshabilitado los actuadores y el único sensor disponible es una cámara simulada orientada al frente (configurada para ofrecer una resolución de 320x200 píxeles y un campo de visión de 72º). Se ha escrito un servicio de sensor específico para la cámara que permite la inserción de una entrada visual sintética específica (en vez de tomar las imágenes directamente del simulador o del entorno real).

El único meta-objetivo que se ha definido en la capa núcleo se basa en la detección de objetos relevantes. Por lo tanto, el funcionamiento del ETC se modula en función de contextos relativos que apuntan a las novedades detectadas. El tipo de novedades que se detectan depende del proceso de percepción, es decir, de los procesadores especializados específicos que se usen. En la versión actual de *CC-Observer* se han

usado dos tipos de procesadores: detectores de movimiento y detectores de Regiones de Interés (RdI).

Para los experimentos se han usado los siguientes tipos de estímulos visuales:

- ***S1***. Objeto estático blanco sobre fondo oscuro.
- ***S2***. Objeto blanco moviéndose a lo largo de una trayectoria rectilínea.
- ***S3***. Dos puntos blancos estacionarios que parpadean alternativamente (ver Figura 63).

S1 y *S2* se generan usando el simulador de RDS, mientras que *S3* se genera usando el mecanismo de inserción de imágenes mencionado anteriormente. *S3* consiste en una secuencia diseñada para inducir la sensación de movimiento (efecto de movimiento aparente) tal y como se describe en el Apartado 6.5.

La capacidad de informar a un tercero sobre los perceptos que percibe el sistema – específicamente aquellos con un significado fundado en la realidad que rodea al agente – es una de las características que indica la presencia de un mecanismo de conciencia (Haikonen 2007b). Por lo tanto, para estudiar los qualia artificiales en *CC-Observer* es necesario analizar qué tipo de contenido integrado es el que es capaz de generar el agente. Dado que el sistema implementado no cuenta con un mecanismo de comunicación similar al que tiene un humano, se ha optado por usar un mecanismo de inspección del estado interno: el Visor CERA o CERA *Viewer* (ver Figura 99).

Figura 99. Proceso de especificación de los qualia.

Aunque la salida del visor no se puede comparar directamente con la comunicación verbal precisa que generan los sujetos humanos, se pueden adoptar estrategias alternativas para comparar los contenidos de la percepción consciente en los humanos con la especificación del contenido de la percepción explícita de *CC-Observer*. Por ejemplo, el mismo observador humano puede confirmar si el contenido de su experiencia visual coincide con los perceptos integrados que se representan en el visor CERA.

El esquema de comparación propuesto constituye un paso inicial hacia enfoques de Fenomenología Sintética más completos. De las tres etapas que se han definido en la caracterización propuesta de los qualia artificiales (Apartado 6.4), *CC-Observer* abarca la Etapa 1 (Representación Perceptual del Contenido) y hasta cierto punto la Etapa 3 (Auto-Modulación y Comunicación). Una implementación completa que abarque completamente las tres etapas del modelo propuesto está fuera del alcance de la presente tesis.

Los experimentos se han realizaron de la siguiente forma: se expuso a los estímulos visuales *S1*, *S2* y *S3* tanto al sujeto humano sin previo conocimiento del dominio (*H*)

como a la implementación mínima de CERA-CRANIUM (*CC-Observer*). Se le pidió a *H* que prestara atención a los objetos blancos y que comunicara verbalmente las acciones que percibía al mirar la pantalla del ordenador. A *CC-Observer* se le configuró para perseguir un único meta-objetivo: prestar atención a los objetos relevantes. Es decir, la capa núcleo estaba programada para generar contextos espaciales apuntando a las RdI o los objetos en movimiento.

El visor CERA se programó para generar una representación de los perceptos complejos superpuesta sobre el campo de visión de la cámara. Sólo se activaron dos procesadores especializados en el ETC de la capa física: un detector de RdI específico para objetos blancos y un detector de movimientos basado en los cambios de los píxeles. Consecuentemente, *CC-Observer* sólo es capaz de especificar (usando píxeles de color rojo) la localización de los perceptos complejos que corresponden a RdI integradas y la dirección del movimiento (usando una marca de color negro que indica la dirección del movimiento).

El intervalo de la memoria de trabajo (tiempo máximo que los perceptos permanecen almacenados en el ETC y por lo tanto disponibles para los procesadores especializados) se configuró en 500 milisegundos. La duración de los fotogramas con el punto blanco en el estímulo *S3* fue de 100 milisegundos y el intervalo entre estímulos (IIE) 50 milisengundos.

7.6.4 Resultados

Cuando se le mostró el estímulo *S1*, *H* comunicó que "*había un objeto blanco sobre el suelo, localizado cerca del centro de la pantalla*". La salida del visor CERA cuando se le mostró el mismo estímulo visual coincidió con parte de la descripción de *H* (ver Figura 100a). Dadas las limitaciones de la implementación *CC-Observer*, todo el significado que el visor CERA puede representar es exclusivamente acerca de objetos blancos y objetos en movimiento. Por lo tanto, no puede aparecer ninguna representación relacionada con el concepto "suelo" en el visor CERA.

Cuando se le mostró el estímulo *S2*, *H* informó acerca de "*un objeto redondo moviéndose uniformemente de derecha a izquierda*". Las representaciones del visor CERA coincidieron de nuevo con parte de la especificación proporcionada por *H* (ver Figura 100b). Dado que el procesador detector de movimiento no proporciona ninguna medida relacionada con la velocidad *CC-Observer* no pudo percibir la velocidad constante del objeto en movimiento.

Tal y como se esperaba, cuando a *H* se le mostró el estímulo *S3* informó acerca de "*una pelota moviéndose constantemente de un lado a otro*". Sin embargo, *CC-Observer* no produjo una representación equivalente (ver Figura 100c). El visor CERA mostró marcas de movimiento hacia la izquierda y hacia la derecha como era de esperar (este comportamiento cesa si el intervalo de la memoria de trabajo se más corto que el IIE), pero no se representó una continuidad del movimiento durante los fotogramas en negro (IIE) después de cada punto blanco. En resumen, la representación generada por *CC-Observer* no correspondió con la experiencia visual continua integrada ("*moviéndose constantemente*") que expresó *H*.

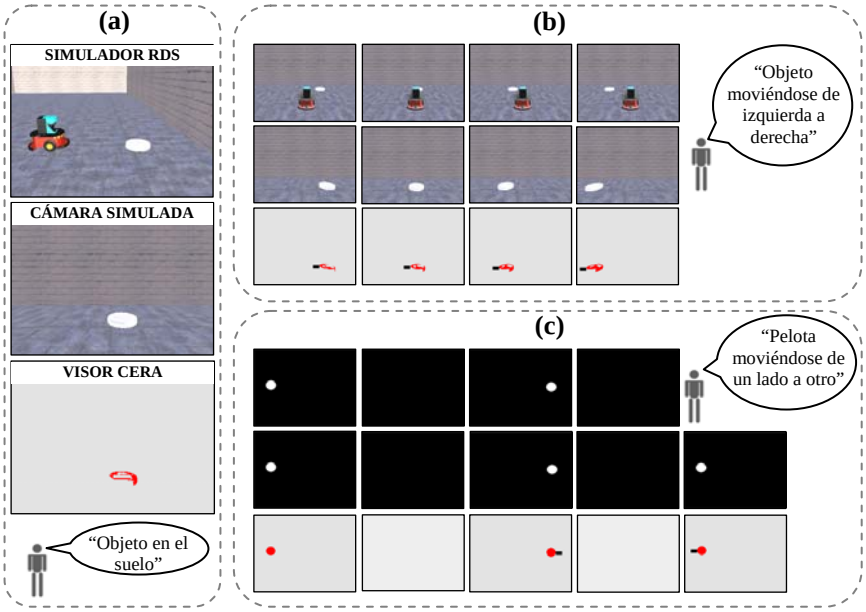


Figura 100. Estímulos visuales *S1*, *S2* y *S3* y especificaciones de *H* y *CC-Observer*.

7.6.5 Conclusiones

La línea de investigación sobre Fenomenología Sintética iniciada con este trabajo parece prometedora. Sin embargo, queda mucho por hacer para conseguir especificaciones del contenido de la experiencia equiparables a las proporcionadas por humanos. Los resultados preliminares indican que la implementación propuesta ha de ser mejorada para poder especificar de forma más precisa la experiencia visual humana.

Se sabe que la percepción humana está condicionada en gran medida por las expectativas. Por lo tanto, el siguiente paso en la mejora de la arquitectura propuesta es incluir la generación de perceptos basados en las expectativas. Es de esperar que el uso de expectativas contribuya además al diseño de un sistema más robusto tal y como demanda la experimentación con imágenes reales.

8 Conclusions and Future Work

8.1 Conclusions

This thesis has approached the problem of Machine Consciousness from a cognitive perspective. However, the study of the phenomenal dimension of consciousness in the realm of artificial systems has not been neglected. In fact, two approaches typically unrelated such as phenomenology and cognitive modeling have been integrated into the same research work. Therefore, one of the major contributions of this work lies in the discovery of new possibilities for the study of consciousness and the use of computational models both in machine and human consciousness research.

In this work, consciousness has been characterized as an integrative "super-function" that can be gradually developed in artificial agents. On one hand, a novel scale (*ConsScale*) has been proposed for the practical measurement of the level of the development of this super-function; on the other hand, a cognitive architecture (CERA-CRANIUM) based on a computational model of consciousness has been implemented and used for experimentation in the following areas:

- The identification of key cognitive functions associated with consciousness.
- The study of the interaction and possible effective integration of some of these cognitive functions.
- The modeling and specification of the contents of the conscious experience (artificial qualia).

ConsScale has become a pragmatic framework for addressing the problem of measuring consciousness in artificial systems. Although this approach is, of necessity, incomplete, the proposed scale is a practical tool in terms of applicability and effective assessment. The parallel development of the scale and the cognitive architecture has permitted a reciprocal feedback during design phases, thus enriching both lines of research.

Doing research in consciousness implies dealing with controversial and open problems, such as the application of the scientific method to the subjective experience or the problem of finding an appropriate definition for cognition. In that regard, the present work has been based on the preliminary adoption of a set of working hypotheses (see Chapter 1), which have been put to test during the development of the thesis obtaining the following conclusions:

- **Hypothesis**: *phenomenal consciousness can be studied using artificial systems*.
    - **Conclusion**: the definition and the existence of phenomenal states remains an open problem both in philosophy and science. The work

carried out in this thesis has demonstrated that it is possible to use computational models to at least generate specifications of conscious experience. Studying how different artificial architectures generate mental content specifications can be useful to analyze and compare how human brain produces qualia.

- **Hypothesis:** *the functional role of consciousness is integration and adaptation*.
    - **Conclusion:** the identification of a set of essential cognitive functions associated with consciousness and the structure of their interrelations in the proposed models (*ConsScale* and MC^3) has contributed to a plausible characterization of the processes of perception and behaviour generation. To a greater or lesser extent conscious beings are characterized by the effective integration of cognitive capabilities. Specific dependency and synergy structures between these functional components have been described and put to test in implementations like CC-Bot (most human bot in BotPrize 2010 Turing test competition). Obtained results indicate that consciousness could be considered the consequence of the self-organization and self-regulation of this type of complex systems.
- **Hypothesis:** *consciousness is a process, not a property of matter*.
    - **Conclusion**: although the present thesis cannot offer any convincing objection against panpsychism, the artificial qualia generation model as described in Chapter 6, allows for the speculation about the possible association of phenomenal states and mental content generation processes. More conclusive results about this hypothesis could be obtained performing experiments with machines rated as *ConsScale* level 7 or above, which as discussed in this thesis are not available yet.
- **Hypothesis:** *conscious processing is just the tip of the iceberg*.
    - **Conclusion:** most theories of consciousness claim that the vast majority of cognitive systems processes take place unconsciously. The models proposed in this thesis are based on these theories and none of the aspects analyzed during the research work indicate any different position. Essentially, the problem about this hypothesis lies in the identification of the specific processes or states that can be regarded as conscious. In other words, all the processes performed by a machine are in principle considered unconscious; the problem is how to prove that certain selection of processes can be considered conscious. In this thesis conscious experience has been characterized in terms of explicit contents (artificial qualia) generated by the proposed computational model. Therefore, the tip of the iceberg is characterized by the sequential thread of selected explicit content. This provides a content-centric explanation, but possible associated phenomenal states remain elusive for clear identification (or existence).
- **Hypothesis:** *consciousness can be scientifically studied*.
    - **Conclusion**: although the scientific community generally agrees that rigorous scientific study of functional aspects of cognition is possible, the scientific study of phenomenal consciousness remains uncertain. As mentioned above, the work carried out in this thesis can be used as the

base for new speculations about the study of phenomenal experience. However, no reliable proof can be offered about that hypothesis using the present work. Given that the hard problem of consciousness seems to resist direct research efforts towards its resolution, an alternative approach is to address the easy problems of consciousness using *ConsScale* as guideline.

- **Hypothesis:** *the level of consciousness of an artificial system can be measured*.
    - **Conclusion**: this thesis has focused on the measurement of functional aspects of consciousness. Therefore, no rigorous conclusion can be drawn as for the measurement of phenomenal states. However, the characterization of artificial qualia described in Chapter 6 provides a plausible approach to the problem of Synthetic Phenomenology. This approach, based on the specification of the contents of conscious experience, is also considered in *ConsScale*. Regarding the cognitive functionality associated with consciousness, the present thesis has demonstrated, through the application of the proposed scale *ConsScale*, that it is possible to measure and characterize the development of functional consciousness in artificial systems.

In addition to the main conclusions presented above, other practical conclusions have been drawn regarding Machine Consciousness research. Thanks to the definition of a measure both qualitative and quantitative, a comparative analysis of different Machine Consciousness models and implementations has been possible. Moreover, the proposed scale can be considered a roadmap for the design of more advanced cognitive agents expected to develop human-like behaviour. First steps in this roadmap have been taken using the CERA-CRANIUM cognitive architecture.

The potential of *ConsScale* and CERA-CRANIUM has been illustrated in different problem domain applications: videogame synthetic characters, mobile robotics, and Synthetic Phenomenology. Regarding the latter, it is interesting to remark that the proposed framework permits the study of consciousness from a different perspective, as the inner working of the artificial brain can be inspected during the experiment. Subject introspection is not as private as in humans, as the designer have access to the "mental" processes of the system. Obviously, the experimentation with artificial systems cannot completely substitute the experimentation with biological conscious organisms. However, in the area of Artificial Intelligence, 50 years of interplay between computation and biological sciences have provided useful feedback to both domains. The rigorous research in Machine Consciousness calls for the same interplay and reciprocal feedback.

From the point of view of practical applications of Machine Consciousness in the short-term the present thesis also provides new perspectives, like the improved believability in computer game bots. The promising results obtained in the 2K BotPrize 2010 competition, where the CERA-CRANIUM bot was considered the most human-like character, indicate that current research direction might allow the development of new agents able to pass the videogame version of the Turing test.

Taking into account possible future advancement of the Machine Consciousness field, the work carried out in this thesis – *ConsScale* specifically – can be useful as a suitable framework for the analysis of legal and moral status of autonomous robots. Legal regulations for level 7 (self-conscious) machines would be clearly necessary in a hypothetical scenario in which service robots were extensively integrated in the society.

From an engineering point of view, the present thesis has contributed with the design of a multi-modal attention mechanism, which seamlessly allows the addition of new sensory modalities. Additionally, an effective integration of bottom-up and top-down attentional flows has been developed, having several cognitive functions integrated in a system that self-regulates.

The bottom-up perception flow implemented in CERA-CRANIUM can also be considered as the base for an effective combination of symbolic and subsymbolic approaches. The proposed architecture provides a practical integration of the duality established by the implicit processing in the specialized processors (numerical and subsymbolic) and the explicit processing (symbolic) in the core layer. While the input in the physical layer is parallel, subsymbolic, and high-bandwidth, the core layer works with a low-bandwidth sequential flow of symbolic information.

Ultimately, considering the contribution of this work from the perspective of new challenges, this thesis might be considered a step forward towards the challenge of self-conscious machines.

8.2 Future work

Both the definition of *ConsScale* and the MC^3 model are obviously subject of improvement. Although most important cognitive aspects have been considered, the proposed models could be enhanced using a more complete repertoire of cognitive functions. Additionally, many functions defined in the MC^3 model have not been implemented in CERA-CRANIUM. For instance, long-term memory mechanisms or learning mechanisms have not been implemented.

A crucial aspect of human consciousness is explicit learning. This thesis has not addressed this topic, therefore a number of research projects can be identified in that direction, aiming at implementing different learning techniques within the framework of the proposed models. Another key aspect that, as remarked in Section 7.6.5, will be helpful in the Synthetic Phenomenology research, is the design of a mechanism for the generation of more complex expectations (predictions) in CERA-CRANIUM.

Additionally, many of the mechanisms currently present in CERA-CRANIUM could be significantly improved. For instance, context formation could have been made more flexible using fuzzy logic techniques or new specialized processors could be added to perform a more sophisticated visual processing. Regarding the application in different problem domains, depending on the complexity of the agent, CERA-CRANIUM would require new sensorimotor services, communications services, etc.

Taking into account current state of the art in Machine Consciousness, it is clear that one of the aspects that requires more effort is the development of self-consciousness in machines. In that regard, CERA-CRANIUM could be enhanced with episodic memory mechanisms, a model of the self, second order models of the relation between the self and the environment, etc. These sorts of improvements, applied to implementations like CC-Bot, might potentially give place to a new generation of agents able to pass constrained versions of the Turing test.

The *ConsScale* implicit roadmap is actually a list of items for future work. Abilities associated with level 6 and above (self-modeling, theory of mind, inner speech, etc.) are considered as possible areas for future work. The very definition of *ConsScale* and the

associated methodology for assessment could also be improved in the future. Fuzzy logic techniques could be used for the evaluation of the level of fulfillment of cognitive skills.

Regarding the possible areas for practical application, it is important to remark that the Machine Consciousness field is still very young and significant practical applications are not expected in the short-term. In fact, the only short-term practical application of the work carried out in this thesis is the design of believable computer game bots. Nevertheless, the progress in the research topics mentioned above could potentially generate useful contributions in practical domains like human-robot social interaction.

8.3 Publications

This section lists the papers where the results of this doctoral thesis have been published:

- JCR Journals:
  - Arrabales, R. Ledezma, A. and Sanchis, A. "**ConsScale: A Pragmatic Scale for Measuring the Level of Consciousness in Artificial Agents**". *Journal of Consciousness Studies*. Vol. 17. No. 3-4. March-April 2010. pp. 131-164(34).
  - Arrabales, R. and Sanchis de Miguel, A. "**Applying Machine Consciousness Models in Autonomous Situated Agents**". *Pattern Recognition Letters*. Special Issue on Pattern Recognition in Multidisciplinary Perception and Intelligence. Volume 29. Issue 8. Pages 1033-1038. June 2008.
- International Peer-Review Journals:
  - Arrabales, R. Ledezma, A. and Sanchis, A. "**The Cognitive Development of Machine Consciousness Implementations**". *International Journal of Machine Consciousness*. Vol 2. Issue 2. December 2010. pp. 213-225.
  - Arrabales, R. Ledezma, A. and Sanchis, A. "**Strategies for Measuring Machine Consciousness**". *International Journal of Machine Consciousness*. Vol 1. Issue 2. December 2009. pp. 193-201.
  - Arrabales, R. Ledezma, A. and Sanchis, A. "**A Cognitive Approach to Multimodal Attention**". *Journal of Physical Agents*. Volume 3. Issue 1. Pages 53-64. January 2009.
  - Arrabales, R. Ledezma, A. and Sanchis, A. "**Integrated Attention for Cognitive Robotics**". *Communications of the Systemics and Informatics World Network* (SIWN). Volume 5. August 2008. Pages 1-5.
- International Conferences:
  - Arrabales, R. Ledezma, A. and Sanchis, A. "**Towards the Generation of Visual Qualia in Artificial Cognitive Architectures**". Accepted for publication at BICS 2010 (Brain Inspired Cognitive Systems).
  - Arrabales, R. Ledezma, A. and Sanchis, A. "**On the Practical Nature of Artificial Qualia**". Symposium on AI-Inspired Biology. The 2010 Annual Convention of the Society for the Study of Artificial Intelligence and Simulation of Behaviour (AISB 2010). Leicester, UK. March 2010.
  - Arrabales, R. Ledezma, A. and Sanchis, A. "**Assessing and Characterizing the Cognitive Power of Machine Consciousness Implementations**". AAAI 2009 Fall Symposium on Biologically Inspired Cognitive Architectures (BICA-2009). Technical Report FS-09-01. pp. 16-21. AAAI Press. ISBN 978-1-57735-435-2.

- Arrabales, R. Ledezma, A. and Sanchis, A. "**Towards Conscious-like Behavior in Computer Game Characters**", in Proceedings of the IEEE Symposium on Computational Intelligence and Games 2009 (CIG-2009) pp. 217-224. ISBN 978-1-4244-4815-9.
- Arrabales, R. Ledezma, A. and Sanchis, A. "**Establishing a Roadmap and Metrics for Conscious Machines Development**". In Proceedings of the 8th IEEE International Conference on Cognitive Informatics. G. Baciu, Y. Wang, Y.Y. Yao, W. Kinsner, K. Chan & L.A. Zadeh (Eds.). Hong Kong. June 2009. Pages 94-101. ISBN 978-1-4244-4642-1.
- Arrabales, R. Ledezma, A. and Sanchis, A. "**Designing Human-like Video Game Synthetic Characters through Machine Consciousness**". Towards a Science of Consciousness 2009. Hong Kong. June 2009. Pages 74.
- Arrabales, R. Ledezma Espino, A. and Sanchis de Miguel, A. "**Modelling Consciousness for Autonomous Robot Exploration**". In 2nd International Work-Conference on the Interplay between Natural and Artificial Computation, IWINAC 2007. Lecture Notes in Computer Science Series, Vol. 4527. pp. 51-60.

- National Conferences:
  - Arrabales, R. and Sanchis de Miguel, A., "**La Aplicación de Modelos de Consciencia Artificial en los Sistemas Multiagente**". Actas del Campus Multidisciplinar en Percepción e Inteligencia, CMPI-2006, Vol. 1 pp. 401-412. July 2006. ISBN 84-689-9560-6.

- Workshops:
  - Arrabales, R. Ledezma, A. and Sanchis, A. "**CERA-CRANIUM: A Test Bed for Machine Consciousness Research**". International Workshop on Machine Consciousness 2009. Hong Kong. June 2009. Pages 105.
  - Arrabales, R. Ledezma, A. and Sanchis, A. "**ConsScale: A Plausible Test for Machine Consciousness?**". Proceedings of the Nokia Workshop on Machine Consciousness - 13th Finnish Artificial Intelligence Conference (STeP 2008). Helsinki. Finland. Pages 49-57.
  - Arrabales, R. Ledezma, A. and Sanchis, A. "**A Multimodal Attention Mechanism for Autonomous Mobile Robotics**". IX Workshop on Physical Agents 2008. Joaquín Lopez and Matías García (Eds.) Vigo. Spain. September 2008. Pages 121-128.
  - Arrabales, R. Ledezma, A. and Sanchis, A. "**Criteria for Consciousness in Artificial Intelligent Agents**". ALAMAS+ALAg 2008 - Adaptive Learning Agents and Multi-Agent Systems - Proceedings of the ALAMAS+ALAg Workshop at AAMAS 2008. Estoril. Portugal. Pages 57-64. 2008.
  - Arrabales, R. and Sanchis de Miguel, A. "**A Machine Consciousness Approach to Autonomous Mobile Robotics**". In the 5th International Cognitive Robotics Workshop. AAAI-06. Boston, MA. July 2006.

- Posters:
  - Arrabales, R. Ledezma, A. and Sanchis, A. "**ConsScale: A Cognitive Scale Inspired on Consciousness**". 4th International Conference on Cognitive Systems (CogSys 2010). Zürich. Switzerland. January 2010.
  - Arrabales, R. Ledezma, A. and Sanchis, A. "**CRANIUM-CERA Cognitive Architecture**". 1st International Seminar on New Issues of Artificial Intelligence. Universidad Carlos III de Madrid. February 2008.

Glosario

API	Interfaz de Programa de Aplicaciones (en inglés: Application Program Interface).
ARCOS	Software avanzado de control y operación del robot Pioneer (en ingles: *Advanced Robot Control & Operations Software*).
ARIA	Interfaz Avanzado de Aplicaciones Robóticas (en inglés: *ActivMedia Robotics Interface for Application*).
CCR	Soporte en tiempo de ejecución para Concurrencia y Coordinación (en inglés: *Concurrency and Coordination Runtime*).
CERA	Arquitectura de Razonamiento Consciente y Emocional (en inglés: *Conscious and Emotional Reasoning Architecture*).
CLR	Lenguaje Común en Tiempo de Ejecución de la máquina virtual de .NET Framework (en inglés: Common Language Runtime).
CLS	Índice acumulado por niveles (en inglés: Cummulative Level Score).
CQS	Índice cuantitativo de *ConsScale* (En inglés: *ConsScale* Quantitative Score).
CRANIUM	Gestor Subyacente para una Arquitectura de Robótica de Inspiración Neurológica (Cognitive Robotics Architecture Neurologically Inspired Underlying Manager).
CSS	Conjunto de habilidades cognitivas (en inglés: Cognitive Skills Set).
CVAAR	Método de Cuantificación de Vectores por Asignación Adaptativa de Recursos (en inglés: ARAVQ – Adaptive Resource-Allocating Vector Quantizer).
DLL	Biblioteca de Enlace Dinámico (en inglés: Dynamic Linking Library).
DSS	Servicios Software Descentralizados (en inglés: Decentralized Software Services).
DSSP	Protocolo Descentralizado de Servicios Software (en inglés: *Decentralized Software Services Protocol*).
EEG	Electroencefalograma.
EPO	Escala de Probabilidad Ordinal (en inglés: OPS – Ordinal Probability Scale).
ETC	Espacio de Trabajo Compartido.
ETG	Espacio de Trabajo Global (en inglés: GW – Global Workspace).
FPS	Videojuego de acción en primera persona (*First-Person Shooter*).
GPS	Sistema de Posicionamiento Global (en inglés: Global Positioning System).
GWT	Teoría del Espacio de Trabajo Global (en inglés: GWT – Teoría del Espacio de Trabajo Global).
HTTP	Protocolo de Transferencia de Hipertexto (en inglés: Hypertext Transfer Protocol).
IA	Inteligencia Artificial

IAG	Inteligencia Artificial General (en inglés: AGI – Artificial General Intelligence).
ICC	Centro de Curvatura Instantáneo (en inglés: Instantaneous center of curvatura).
LASER	Amplificación de Luz por Emisión Estimulada de Radiación (en inglés: Light Amplification by Stimulated Emission of Radiation).
MC^3	MCCC – Modelo de la Cognición en CERA-CRANIUM.
MCCA	Modelos Cognitivos de Conciencia Artificial.
MMO	Juego Online Multijugador Masivo (Massivelly Multiplayer Online Game).
NASA	Administración Nacional de Aeronáutica y del Espacio (en inglés: *Nacional Aeronautics and Space Administration*).
PEE	Proceso Estándar de Evaluación.
PSE	Proceso Simplificado de Evaluación.
PTZ	Pan-Tilt-Zoom (Panorámica-Inclinación-Aumento).
RCD	Regiones de Distancia Constante (en inglés: Regions of Constant Distance).
RDS	Robotics Developer Studio.
RGPC	Representación Gráfica del Perfil Cognitivo.
RISC	Reduced Instruction Set Computer (Ordenador con Conjunto Reducido de Instrucciones).
RPG	Juego de Rol (*Role-Playing Game*).
SLAM	Localización y construcción de mapas de forma simultánea (en inglés: Simultaneous Localization And Mapping).
SOAP	Protocolo de Acceso a Objetos Simples (en inglés: Simple Object Access Protocol).
SONAR	Navegación y alcance por sonido (en inglés: SOund NAvigation and Ranging).
TCP/IP	Transmission Control Protocol / Internet Protocol (Protocolo de Control de Transmisión / Protocolo de Internet).
UT2004	Unreal Tournamente 2004
VPL	Lenguaje de Programación Visual (en inglés: Visual Programming Language).

Referencias

Aberg, M. & Rantala, A. 2008, "Neuron Microchips", *Proceedings of the Nokia Workshop on Machine Consciousness 2008*, pp. 46-48.

Aleksander, I. & Morton, H. 2006, "On architectures for synthetic phenomenology", *Proceedings of the AISB06 symposium on integrative approaches to machine consciousness*, eds. R. Chrisley, R. Clowes & S. Torrance, pp. 16-28.

Aleksander, I. & Morton, H. 2007, "Why Axiomatic Models of Being Conscious?", *Journal of Consciousness Studies*, vol. 14, no. 7, pp. 15-27.

Aleksander, I. 2005, "Machine consciousness" in *Progress in Brain Research*, ed. Steven Laureys, Elsevier, pp. 99-108.

Aleksander, I. & Dunmall, B. 2003, "Axioms and Tests for the Presence of Minimal Consciousness in Agents", *Journal of Consciousness Studies*, vol. 10, no. 4-5, pp. 7-18.

Anderson, J.R. 1993, *Rules of the Mind*, Lawrence Erlbaum, Hillsdale, NJ.

Anderson, J.R., Matessa, M. & Lebiere, C. 1997, "ACT-R: A theory of higher level cognition and its relation to visual attention", *Human Computer Interaction*, vol. 14, no. 4, pp. 439-462.

Andrade, J. 1996, "Investigations of Hypesthesia: Using Anesthetics to Explore Relationships between Consciousness, Learning, and Memory", *Consciousness and Cognition*, vol. 5, no. 4, pp. 562-580.

Arrabales, R., Ledezma, A. & Sanchis, A. 2010, "On the Practical Nature of Artificial Qualia", *Proceedings of the 2010 Annual Convention of the Society for the Study of Artificial Intelligence and Simulation of Behaviour (AISB 2010)*.

Arrabales, R., Ledezma, A. & Sanchis, A. 2009a, "CERA-CRANIUM: A Test Bed for Machine Consciousness Research", *Proceedings of the International Workshop on Machine Consciousness 2009*.

Arrabales, R., Ledezma, A. & Sanchis, A. 2009b, "Establishing a Roadmap and Metrics for Conscious Machines Development", *Proceedings of the 8th IEEE International Conference on Cognitive Informatics*, eds. G. Baciu, Y. Wang, Y.Y. Yao, W. Kinsner, K. Chan & L.A. Zadeh.

Arrabales, R., Ledezma, A. & Sanchis, A. 2009c, "A Cognitive Approach to Multimodal Attention", *Journal of Physical Agents*, vol. 3, no. 1, pp. 53-64.

Arrabales, R., Ledezma, A. & Sanchis, A. 2008, "A Multimodal Attention Mechanism for Autonomous Mobile Robotics", *Proceedings of the IX Workshop on Physical Agents*, pp. 121-128.

Arrabales, R., Ledezma, A. & Sanchis, A. 2010a, "ConsScale: A Pragmatic Scale for Measuring the Level of Consciousness in Artificial Agents", *Journal of Consciousness Studies*, vol. 17, no. 3-4, pp. 131-164.

Arrabales, R., Ledezma, A. & Sanchis, A. 2010b, "Towards the Generation of Visual Qualia in Artificial Cognitive Architectures", *Proceedings of the Brain Inspired Cognitive Systems 2010 conference*, eds. C. Hernández, J. Gómez & R. Sanz, Madrid, Spain, pp. 34-46.

Arrabales, R., Ledezma, A. & Sanchis, A. 2009a, "Designing Human-like Video Game Synthetic Characters through Machine Consciousness", *Towards a Science of Consciousness 2009.* Hong Kong, pp. 74.

Arrabales, R., Ledezma, A. & Sanchis, A. 2009b, "Towards Conscious-like Behavior in Computer Game Characters", *IEEE Symposium on Computational Intelligence and Games*Milano, Italy, pp. 217-224.

Arrabales, R., Ledezma, A. & Sanchis, A. 2007, "Modeling Consciousness for Autonomous Robot Exploration", *2nd International Work-Conference on the Interplay between Natural and Artificial Computation (IWINAC 2007)*, pp. 51-60.

Atkinson, A.P., Thomas, M.S.C. & Cleeremans, A. 2000, "Consciousness: Mapping the Theoretical Landscape", *Trends in Cognitive Sciences*, vol. 4, no. 10, pp. 372-382.

Avillac, M., Deneve, S., Olivier, E., Pouget, A. & Duhamel, J. 2005, "Reference frames for representing visual and tactile locations in parietal cortex", *Nature neuroscience*, vol. 8, no. 7, pp. 941-949.
Baars, B., Ramsoy, T. & Laureys, S. 2003, "Brain, conscious experience and the observing self", *Trends in neurosciences*, vol. 26, no. 12, pp. 671-675.
Baars, B.J. 2002, "The conscious access hypothesis: Origins and recent evidence", *Trends in Cognitive Science*, no. 6, pp. 47-52.
Baars, B.J. 1997, "In the Theatre of Consciousness: Global Workspace Theory, A Rigorous Scientific Theory of Consciousness", *Journal of Consciousness Studies*, no. 4, pp. 292-309.
Baars, B.J. & Franklin, S. 2009, "Consciousness is Computational: The LIDA Model of Global Workspace Theory", *International Journal of Machine Consciousness*, vol. 1, no. 1, pp. 23-32.
Baars, B.J. 1988, *A Cognitive Theory of Consciousness*, Cambridge University Press, Cambridge.
Balduzzi, D.A.T.,Giulio 2009, "Qualia: The Geometry of Integrated Information", *PLoS Comput Biol*, vol. 5, no. 8, pp. e1000462.
Balkenius, C. & Morén, J. 2003, "From isolated components to cognitive systems", *ERCIM News*, no. 16.
Basar, E., Basar-Eroglu, C., Karakas, S. & Schurmann, M. 1999, "Oscillatory brain theory: a new trend in neuroscience", *Engineering in Medicine and Biology Magazine*, vol. 18, no. 3, pp. 56-66.
Bauckhage, C., Gorman, B., Thurau, C. & Humphrys, M. 2007, "Learning Human Behavior from Analyzing Activities in Virtual Environments", *MMI-Interaktiv*, vol. 12, pp. 3-17.
Beetz, M., Rajan, K., Thielscher, M. & Rusu, R.B. 2006, *Report on the 5th International Cognitive Robotics Workshop (The AAAI-06 Workshop on Cognitive Robotics)*, Boston, Massachusetts.
Behrman, E.C., Steck, J.E. & Skinner, S.R. 1999, "A spatial quantum neural computer", *International Joint Conference on Neural Networks*, IEEE, pp. 874-877.
Blackburn, S. 2004, "New Scientist's selection on consciousness", *New Scientist*, no. 2464.
Block, N. 2001, Oct. 20, 2001-last update, *Some Concepts of Consciousness*. Available: http://www.nyu.edu/gsas/dept/philo/faculty/block/papers/Abridged%20BBS.htm [2009, May 20] .
Block, N. 1980, "Introduction: What Is Functionalism?" in *Readings in Philosophy of Psychology*, ed. N. Block, Harvard University Press, Cambridge, MA.
Block, N. 1995, "On a Confusion about a Function of Consciousness", *Behavioral and Brain Sciences*, no. 18, pp. 227-287.
Brooks, R.A., Breazeal, C., Marjanovic, M., Scassellati, B. & Williamson, M.M. 1998, "The Cog project: Building a humanoid robot", *Computation for Metaphors, Analogy and Agents*, ed. C. Nehaniv, Springer-Verlag.
Brooks, R. 1986, *Achieving Artificial Intelligence through Building Robots*, Massachusetts Institute of Technology, Cambridge, MA, USA.
Brown, R.A. 1997, "Consciousness in a Self-Learning, Memory-Controlled, Compound Machine", *Neural Networks*, vol. 10, no. 7, pp. 1333-1343.
Bunge, M. 2002, *El problema mente-cerebro. Un enfoque psicobiologico.* Tecnos, España.
Buttazzo, G. 2001, "Artificial consciousness: Utopia or real possibility?", *Computer*, vol. 34, no. 7, pp. 24-30.
Byrne, R.W. & Whiten, A. 1988, *Machiavellian Intelligence: Social Expertise and the Evolution of Intellect in Monkeys, Apes and Humans*, Clarendon Press, New York.
Calverley, D.J. 2005, "Towards a method for determining the legal status of a conscious machine", *Proceedings of the AISB05 symposium on next generation approaches to machine consciousness*, eds. R. Chrisley, R. Clowes & S. Torrance.
Carruthers, P. 2002, "The cognitive functions of language", *Behavioral and Brain Sciences*, vol. 25, no. 6, pp. 657-674.
Carruthers, P. 2000, *Phenomenal Consciousness: A Naturalistic Theory*, Cambridge University Press, Cambridge.
Carruthers, P. 2009, "Higher-Order Theories of Consciousness" in *The Stanford Encyclopedia of Philosophy*, ed. Edward N. Zalta, Fall 2009.
Chalmers, D. 2004, "How can we construct a science of consciousness?" in *The Cognitive Neurosciences III* MIT Press, Cambridge, MA.
Chalmers, D. 1995, "Facing Up to the Problem of Consciousness", *Journal of Consciousness Studies*, vol. 2, no. 3, pp. 200-219.
Chella, A. 2009, "Editorial", *International Journal of Machine Consciousness*, vol. 1, no. 1, pp. 1-5.
Chella, A., Frixione, M. & Gaglio, S. 2008, "A cognitive architecture for robot self-consciousness", *Artificial Intelligence in Medicine*, vol. 44, no. 2, pp. 147-154.
Chella, A. & Gaglio, S. 2009, "In Search of Computational Correlates of Artificial Qualia", *The Second Conference on Artificial General Inteligence (AGI-09)*.

Chella, A. & Macaluso, I. 2009, "The perception loop in CiceRobot, a museum guide robot", *Neurocomputing*, vol. 72, no. 4-6, pp. 760-766.
Chrisley, R. 2009, "Synthetic Phenomenology", *International Journal of Machine Consciousness*, vol. 1, no. 1, pp. 53-70.
Chrisley, R.L. 1995, "Taking embodiment seriously: nonconceptual content and robotics" in *Android Epistemology*, eds. K.M. Ford, C. Glymour & P. Hayes, MIT Press, Cambridge, MA, USA, pp. 141-166.
Ciompi, L. 2003, "Reflections on the role of emotions in consciousness and subjectivity, from the perspective of affect-logic", *Consciousness \& Emotion*, vol. 4, no. 2, pp. 181-196.
Cleeremans, A. 2005, "Computational correlates of consciousness ", *Progress in brain research*, vol. 150, pp. 81-98.
Cotterill, R. 2003, "CyberChild. A Simulation Test-Bed for Consciousness Studies", *Journal of Consciousness Studies*, vol. 10, no. 4-5, pp. 31-45.
Crick, F. & Koch, C. 2003, "A framework for consciousness", *Nature neuroscience*, vol. 6, pp. 119-126.
Crick, F. 1994, *The Astonishing Hypothesis: The Scientific Search for the Soul*, Charles Scribner's Sons, New York.
Crick, F. & Koch, C. 1990, "Towards a Neurobiological Theory of Consciousness", *Semin Neurosci*, no. 2, pp. 263–275.
Cytowic, R.E. 2002, *Synesthesia. A Union of the Senses*, Bradford Books.
Damasio, A.R. 1995, *Descartes' Error: Emotion, Reason, and the Human Brain*, Harper Perennial, New York, NY.
Damasio, A.R., Everitt, B.J. & Bishop, D. 1996, "The Somatic Marker Hypothesis and the Possible Functions of the Prefrontal Cortex", *Philosophical Transactions: Biological Sciences*, vol. 351, no. 1346, pp. 1413-1420.
Damasio, A., Damasio, H., Tranel, D. & Brandt, J. 1990, "Neural regionalization of knowledge access: preliminary evidence", *Symp Quant Biol.*, vol. 55, pp. 1039-1047.
Damasio, A.R. 1999, *The Feeling of What Happens: Body and Emotion in the Making of Consciousness*, Heinemann, London.
De Veer, M.W., Gallup, G.G., Theall, L.A., van den Bos, R. & Povinelli, D.J. 2003, "An 8-year longitudinal study of mirror self-recognition in chimpanzees (Pan troglodytes)", *Neuropsychologia*, vol. 41, no. 2, pp. 229-234.
Dehaene, S., Sergent, C. & Changeux, J. 2003, "A neuronal network model linking subjective reports and objective physiological data during conscious perception", *Proceedings of the National Academy of Sciences of the United States of America*, vol. 100, no. 14, pp. 8520-8525.
Dennett, D.C. 2003, "Whos On First? Heterophenomenology Explained", *Journal of Consciousness Studies*, vol. 10, pp. 19-30.
Dennett, D.C. 1997a, "Consciousness in Human and Robot Minds" in *Cognition, Computation and Consciousness*, eds. M. Ito, Y. Miyashita & Edmund T. Rolls, Oxford University Press, .
Dennett, D.C. 1997b, *Kinds Of Minds: Toward An Understanding Of Consciousness*, Basic Books, New York.
Dennett, D.C. 1996, "Facing Backwards on the Problem of Consciousness", *Journal of Consciousness Studies*, vol. 3, no. 1, pp. 4-6.
Dennett, D.C. 1991, *Consciousness Explained*, Little, Brown and Co, Boston.
Dennett, D.C. 1988, *Quining Qualia*, Oxford University Press.
Doan, T. 2009a, *Damasio's Consciousness Model*. Available: http://www.conscious-robots.com [2009, May. 13].
Doan, T. 2009b, *Pentti Haikonen's Architecture for Conscious Machines*. Available: http://www.conscious-robots.com [2009, Dec. 16].
Dobbyn, C. & Stuart, S. 2003, "The Self as an Embedded Agent", *Minds and Machines*, vol. 13, no. 2, pp. 187-201.
Doesburg, S.M., Kitajo, K. & Ward, L.M. 2005, "Increased gamma-band synchrony precedes switching of conscious perceptual objects in binocular rivalry", *Neuroreport*, vol. 16, no. 11.
Edelman, G.M. 1989, *The remembered present*, Basic Books, New York.
Edelman, G.M. 1987, *Neural Darwinism: The Theory of Neuronal Group Selection*, Basic Books, New York.
Edelman, D.B., Baars, B.J. & Seth, A.K. 2005, "Identifying hallmarks of consciousness in non-mammalian species", *Consciousness and Cognition*, vol. 14, no. 1, pp. 169-187.
Edelman, G.M. 2003, "Naturalizing consciousness: A theoretical framework", *Proceedings of the National Academy of Sciences of the United States of America*, vol. 100, no. 9, pp. 5520-5524.
Edelman, G.M. 1992, *Bright air, Brilliant fire (on the matter of the mind)*, Basic books, 1992.

Edelman, G.M. & Tononi, G. 2000, *A Universe Of Consciousness: How Matter Becomes Imagination*, Basic Books, NY.

Epic Games, I. , *Unreal Tournament 2004*. Available: http://www.unrealtournament.com/ [2009, May 25] .

Fenigstein, A., Scheier, M.F. & Buss, A.H. 1975, "Private and public self-consciousness: Assessment anf theory", *Journal of Consulting and Clinical Psychology*, pp. 522-527.

Fenwick, P.B.C., Kostopoulos, G.K., Liu, L.C. & Ioannides, A.A. 2003, "Consciousness and its correlates in awake condition, in different sleep stages and in epilepsy", *Proceedings of the International Joint Conference on Neural Networks*IEEE, , pp. 276-281.

Finger, S. 1995, "Descartes and the pineal gland in animals: A frequent misinterpretation", *Journal of the history of the neurosciences*, vol. 4, pp. 166-182.

Fogassi, L., Gallese, V., Pellegrino, G., Fadiga, L., Gentilucci, M., Luppino, G., Matelli, M., Pedotti, A. & Rizzolatti, G. 1992, "Space coding by premotor cortex", *Experimental Brain Research*, vol. 89, no. 3, pp. 686-690.

Fong, T., Nourbakhsh, I. & Dautenhahn, K. 2003, "A survey of socially interactive robots", *Robotics and Autonomous Systems*, vol. 42, no. 3-4, pp. 143-166.

Forlizzi, J. & DiSalvo, C. 2006, "Service robots in the domestic environment: a study of the roomba vacuum in the home", *HRI '06: Proceedings of the 1st ACM SIGCHI/SIGART conference on Human-robot interaction*ACM, New York, NY, USA, pp. 258.

Franklin, S., Ramamurthy, U., DiMello, S.K., McCauley, L., Negatu, A., Silva, R.L. & Datla, V. 2007a, "LIDA: A Computational Model of Global Workspace Theory and Developmental Learning", *AAAI Fall Symposium on AI and Consciousness: Theoretical Foundations and Current Approaches*AAAI Press, Menlo Park, California, pp. 61-66.

Franklin, S., Ramamurthy, U., DMello, S.K., McCauley, L., Negatu, A., Silva, R. & Datla, V. 2007b, "LIDA: A computational model of global workspace theory and developmental learning", *AAAI Fall Symposium on AI and Consciousness: Theoretical Foundations and Current Approaches*.

Franklin, S. 2005a, "A Consciousness Based Architecture for a Functioning Mind" in *Visions of Mind*, eds. D. Davis & P.A. Hershey, IDEA Group, Inc.

Franklin, S. 2005b, "Evolutionary pressures and a stable world for animals and robots: A commentary on Merker", *Consciousness and cognition*, vol. 14, pp. 115-118.

Franklin, S., Kelemen, A. & McCauley, L. 1998, "IDA: A Cognitive Agent Architecture", *IEEE Conference on Systems, Man and Cybernetics*, vol. 14.

Frijters, J. 2009, , *IKVM, an implementation of Java for Mono and the .NET Framework.* Available: http://www.ikvm.net [2009, May 27] .

Frintrop, S. 2006, *VOCUS: A Visual Attention System for Object Detection and Goal-directed Search*, .

Fürstenau, N. 2007, "A Computational Model of Bistable Perception- Attention Dynamics with Long Range Correlations" in *KI 2007: Advances in Artificial Intelligence*, pp. 251-263.

Gallup, G.G. 1977, "Self-recognition in primates: A comparative approach to the bidirectional properties of consciousness.", *American Psychologist*, vol. 32, no. 5, pp. 329-337.

Gamez, D. 2009, "The Simulation of Spiking Neural Networks" in *Handbook of Research on Discrete Event Simulation Environments: Technologies and Applications*, eds. E.M.O. Abu-Taieh & A.A. El Sheikh, IGI Global, Hershey, Pennsylvania, pp. 337-358.

Gamez, D. 2006, "The XML approach to synthetic phenomenology", *Proceedings of the AISB06 symposium on integrative approaches to machine consciousness*, eds. R. Chrisley, R. Clowes & S. Torrance, , pp. 128-135.

Gamez, D. 2005, "An Ordinal Probability Scale for Synthetic Phenomenology", *Proceedings of the AISB05 Symposium on Next Generation approaches to Machine Consciousness*, eds. R. Chrisley, R. Clowes & S. Torrance, pp. 85-94.

Gamez, D. 2008, "Progress in machine consciousness", *Consciousness and cognition*, vol. 17, no. 3, pp. 887-910.

Gavrilets, S. & Vose, A. 2006, "The dynamics of Machiavellian intelligence", *PNAS*, vol. 103, no. 45, pp. 16823-16828.

Giacino, J., Kalmar, K. & Whyte, J. 2004, "The JFK Coma Recovery Scale-Revised: measurement characteristics and diagnostic utility", *Arch Phys Med Rehabil*, vol. 85, no. 12, pp. 2020-2029.

Giacino, J.T. 2004, "The vegetative and minimally conscious states: consensus-based criteria for establishing diagnosis and prognosis", *Neuro Rehabilitation*, vol. 19, no. 4, pp. 293-298.

Giard, M.H. & Peronnet, F. 1999, "Auditory-Visual Integration during Multimodal Object Recognition in Humans: A Behavioral and Electrophysiological Study", *The Journal of Cognitive Neuroscience*, vol. 11, no. 5, pp. 473-490.

Goertzel, B. 2008, "Achieving Advanced Machine Consciousness through Integrative, Virtually Embodied Artificial General Intelligence", *Proceeding of the Nokia Workshop on Machine Consciousness*, ed. P.O.A. Haikonen, Helsinki, pp. 19-21.
Goertzel, B. 2009, "OpenCogPrime: A cognitive synergy based architecture for artificial general intelligence", *IEEE International Conference on Cognitive Informatics*, pp. 60-68.
Goertzel, B., Pennachin, C., Geisweiller, N., Looks, M., Senna, A., Silva, W., Heljakka, A. & Lopes, C. 2008, "An Integrative Methodology for Teaching Embodied Non-Linguistic Agents, Applied to Virtual Animals in Second Life", *Artificial General Intelligence 2008, Proceedings of the First AGI Conference*, pp. 161-175.
González López, F.J. 2009, *Diseño e Implementación de un Personaje Sintético Inteligente para un Videojuego de Acción en Primera Persona*, Universidad Carlos III de Madrid, Leganés.
Gray, J. 2003, "How are qualia coupled to functions?", *Trends in Cognitive Sciences*, vol. 7, no. 5, pp. 192-194.
Grossberg, S. 2003, "The Brain's Cognitive Dynamics: The Link between Learning, Attention, Recognition, and Consciousness", *Knowledge-Based Intelligent Information and Engineering Systems; LNCS*, eds. V. Palade, R.J. Howlett & L.C. Jain, Springer, pp. 5-12.
Grossberg, S. 1987, "Competitive learning: From interactive activation to adaptive resonance", *Cognitive Science*, vol. 11, no. 1, pp. 23-63.
Haikonen, P.O.A. 2009, "Qualia and Conscious Machines", *International Journal of Machine Consciousness*, vol. 1, no. 2, pp. 225-234.
Haikonen, P.O.A. 2007a, "Reflections of Consciousness: The Mirror Test", *Proceedings of the 2007 AAAI Fall Symposium on Consciousness and Artificial Intelligence*, pp. 67-71.
Haikonen, P.O.A. 2007b, *Robot Brains. Circuits and Systems for Conscious Machines*, John Wiley & Sons, UK.
Haikonen, P. "Quasi-Quantum Computing in the Brain?", *Cognitive Computation*, vol. 2, no. 2, pp. 63-67.
Hameroff, S.R. & Penrose, R. 1996, "Orchestrated reduction of quantum coherence in brain microtubules: A model for consciousness", *Toward a Science of Consciousness*, eds. S. Hameroff, A. Kaszniak & A. Scott , MIT Press.
Harnad, S. 1994, "Levels of Functional Equivalence in Reverse Bioengineering: The Darwinian Turing Test for Artificial Life", *Artificial Life*, vol. 1, no. 3, pp. 293-301.
Harnad, S. 1990, "The symbol grounding problem", *Physica D: Nonlinear Phenomena*, vol. 42, no. 1-3, pp. 335-346.
Harnad, S. & Scherzer, P. 2008, "First, scale up to the robotic Turing test, then worry about feeling", *Artificial Intelligence in Medicine*, vol. 44, no. 2, pp. 83.
Hernández, C., López, I. & Sanz, R. 2009, "The Operative Mind: A Functional, Computational And Modeling Approach To Machine Consciousness", *International Journal of Machine Consciousness*, vol. 1, no. 1, pp. 83-98.
Hesslow, G. 2002, "Conscious thought as simulation of behaviour and perception", *Trends in cognitive sciences*, vol. 6, no. 6, pp. 242-247.
Hingston, P. 2009, "A Turing Test for Computer Game Bots", *IEEE Transactions on Computational Intelligence and AI in Games*, pp. 169 - 186.
Hofstadter, D. 1995, "The Copycat Project: A model of mental fluidity and analogy-making" in *Fluid Concepts and Creative Analogies* Basic Books, New York, NY.
Holland, O. & Goodman, R. 2003, "Robots with internal models" in , ed. O. Hollland, Imprint Academic, Exeter, UK.
Holland, O. & Knight, R. 2006, "The anthropomimetic principle", *Proceedings of the AISB06 symposium on biologically inspired robotics*, eds. J. Burn & M. Wilson.
Holland, O. 2007, "A Strongly Embodied Approach to Machine Consciousness", *Journal of Consciousness Studies*, vol. 14, pp. 97-110(14).
Holmes, N. 2002, "Would a digital brain have a mind?", *Computer*, vol. 35, no. 5, pp. 112-111.
Humphrey, N. 1999, *A history of the mind: evolution and the birth of consciousness*, Springer, New York.
Hunt, E. & Lansman, M. 1986, "Unified Model of Attention and Problem Solving", *Psychological review*, vol. 93, no. 4, pp. 446-461.
Huxley, T.H. 1898, "On the Hypothesis that Animals are Automata, and its History" in *Method and Results: Essays by Thomas H. Huxley* D. Appleton and Company, New York, pp. 555-580.
IEEE Symposium on Computational Intelligence and Games 2008, 15th December-last update, *2K Bot Prize Competition*. Available: http://www.botprize.org/ [2008, Dec. 15].
International Federation of Robotics 2009, *World Service Robots 2009*, IFR Statistical Department.
Itti, L., Rees, G. & Tsotsos, J.K. 2005, *Neurobiology of Attention*, Elsevier, San Diego, CA.

Jackson, F. 1982, "Epiphenomenal Qualia", *The Philosophical Quarterly*, , no. 32, pp. 127-136.
Jennett, B. 2002, "The Glasgow Coma Scale: History and current practice", *Trauma*, vol. 4, no. 2, pp. 91-103.
Kadlec, R., Gemrot, J., Burkert, O., Bída, M., Havlíček, J. & Brom, C. 2007, "POGAMUT 2 - A Platform for Fast Development of Virtual Agents' Behavior", *11th International Conference on Computer Games: AI, Animation, Mobile, Educational & Serious Games*.
Kaminka, G.A., Veloso, M.M., Schaffer, S., Sollitto, C., Adobbati, R., Marshall, A.N., Scholer, A. & Tejada, S. 2002, "GameBots: a flexible test bed for multiagent team research", *Commun.ACM*, vol. 45, no. 1, pp. 43-45.
Kieras, D.E. & Meyer, D.E. 1997, "An overview of the EPIC architecture for cognition and performance with application to human-computer interaction", *Hum.-Comput.Interact.*, vol. 12, no. 4, pp. 391-438.
Kim, L. 1997, "A proposed model of human consciousness system with applications in pattern recognition", *First International Conference on Knowledge Based Intelligent Electronic Systems*, ed. L.C. Jain, IEEE, Adelaide. Australia., pp. 159-166.
Kitamura, T., Otsuka, Y. & Nakao, T. 1995, "Imitation of Animal Behavior with Use of a Model of Consciousness-Behavior Relation for a Small Robot", *4º IEEE International Workshop on Robot and Human Communication*Tokyo, pp. 313-316.
Koch, K. & Tononi, G. 2008, "Can Machines Be Conscious?", *IEEE Spectrum. Special Report: The Singularity.*, [Online], pp. 22 Aug 2009. Available from: http://www.spectrum.ieee.org/biomedical/imaging/can-machines-be-conscious.
Koenigs, M., Young, L., Adolphs, R., Tranel, D., Cushman, F., Hauser, M. & Damasio, A. 2007, "Damage to the prefrontal cortex increases utilitarian moral judgements", *Nature*, , no. 21.
Koivisto, M. & Revonsuo, A. 2003, "An ERP study of change detection, change blindness, and visual awareness", *Psychophysiology*, vol. 40, no. 3, pp. 423-429.
Kozma, R. 1997, "On the Conscious and Subconscious Components of Knowledge Representation in Neural Networks", *IEEE International Conference on Neural Networks (IJCNN'97)*, IEEE, Houston, TX, pp. 2519-2523.
Kubota, N., Kojima, F. & Fukuda, T. 2001, "Self-consciousness and emotion for a pet robot with structured intelligence", *Joint 9th IFSA World Congress and 20th NAFIPS International Conference*IEEE, pp. 2786 - 2791.
Kurzweil, R. 2000, *The age of spiritual machines*, Penguin Putnam, London, UK.
LaBar, K.S. & Cabeza, R. 2006, "Cognitive neuroscience of emotional memory", *Nature reviews. Neuroscience*, vol. 7, no. 1, pp. 54-64.
Laird, A., Newell, J.E. & Rosenbloom, P.S. 1987, "SOAR: Architecture for General Intelligence", *Artificial Intelligence*, vol. 33, pp. 1-64.
Langley, P. & Choi, D. 2006, "A unified cognitive architecture for physical agents", *Proceedings of the Twenty-First National Conference on Artificial Intelligence*AAAI Press, Boston, pp. 1469-1474.
Laureys, S., Owen, A.M. & Schiff, N.D. 2004, "Brain function in coma, vegetative state, and related disorders", *The Lancet Neurology*, vol. 3, no. 9, pp. 537-546.
Lehky, S.R. & Sejnowski, T.J. 1999, "Seeing White: Qualia in the Context of Decoding Population Codes", *Neural computation*, vol. 11, no. 6, pp. 1261-1280.
Lennie, P. 2003, "The Physiology of Color Vision" in *The Science of Color (Second Edition)*, ed. Steven K. Shevell, Elsevier Science Ltd, Amsterdam, pp. 217-246.
Levine, J. 1983, "Materialism and Qualia: The Explanatory Gap.", *Pacific Philosophical Quarterly*, no. 64.
Lewis, L. 2007, "The robots are running riot! Quick, bring out the red tape", *The Times*, [Online], pp. March 223 2010. Available from: http://www.timesonline.co.uk/tol/news/world/asia/article1620558.ece.
Lewis, M. 2003, "The Emergence of Consciousness and Its Role in Human Development", *Annals of the New York Academy of Sciences*, vol. 1001, no. 1, pp. 104-133.
Linaker, F. 2000, "Time Series Segmentation Using an Adaptive Resource Allocating Vector Quantization Network Based on Change Detection", *IJCNN '00: Proceedings of the IEEE-INNS-ENNS International Joint Conference on Neural Networks (IJCNN'00)-Volume 6*IEEE Computer Society, Washington, DC, USA, pp. 6323.
Livingstone, D. 2006, "Turing's test and believable AI in games", *Comput.Entertain.*, vol. 4, no. 1, pp. 6.
Llinás, R., Ribary, U., Contreras, D. & Pedroarena, C. 1998, "The neuronal basis for consciousness", *Philosophical Transactions of the Royal Society - Biological Sciences*, vol. 353, no. 1377, pp. 1851-1859.

Lyshevsky, L. 2002, "Neuroscience – Neuroarchitectronics – Nanocomputers – and – Nanotechnology", *NANO 2002*, IEEE, Washington D.C.
Manson, N. 2000, "State consciousness and creature consciousness: a real distinction", *Philosophical Psychology*, vol. 13, pp. 405-410.
Manzotti, R., Metta, G. & Sandini, G. 1998, "Emotion and learning in a developing robot", *Proc. of Emotion, Consciousness and Qualia*Naples and Ischia (Italy), pp. 19-24.
Marina, J.A. 2002, *El laberinto sentimental*, Anagrama, Barcelona,.
Markram, H. 2006, "The Blue Brain Project", *Nature reviews. Neuroscience*, vol. 7, no. 2, pp. 153-160.
Marques, H.G. 2009, *Architectures for Embodied Imagination*, University of Essex.
Marques, H.G. & Holland, O. 2009, "Architectures for functional imagination", *Neurocomputing*, vol. 72, no. 4-6, pp. 743-759.
McFarland, D. & Bösser, T. 1993, *Intelligent behavior in animals and robots*, MIT Press, Boston, MA.
Merker, B. 2005, "The liabilities of mobility: A selection pressure for the transition to consciousness in animal evolution", *Consciousness and cognition*, vol. 14, no. 1, pp. 89-114.
Metzinger, T. 2003, *Being No One: The Self-Model Theory of Subjectivity*, The MIT Press, Cambridge, MA.
Microsoft Corp. 2006, *Microsoft Robotics Studio.* Available: *http://msdn.microsoft.com/robotics/* [2006. Dec, 18].
Minsky, M. 1988, *Society of Mind*, Simon & Schuster, New York.
Monti, M.M., Vanhaudenhuyse, A., Coleman, M.R., Boly, M., Pickard, J.D., Tshibanda, L., Owen, A.M. & Laureys, S. 2010, "Willful Modulation of Brain Activity in Disorders of Consciousness", *The New England journal of medicine*.
Montz Andrée, R., Jiménez Vicioso, A., Coullaut Jáuregui, J., López-Ibor Aliño, J.J. & Carreras Delgado, J.L. 2002, "PET en neurología y psiquiatría I. PET con FDG en el estudio del SNC", *Revista Española de Medicina Nuclear*, vol. 21, no. 5, pp. 370-386.
Moor, J.H. 1988, "Testing Robots for Qualia" in *Perspectives on Mind*, eds. Herbert R. Otto & James A. Tuedio, Reidel Publishing Company, Dordrecht/Boston/Lancaster/Tokyo.
Moravec, H. 1990, *Mind Children: The Future of Robot and Human Intelligence*, Harvard University Press, Cambridge, Massachusetts.
Morin, A. 2002, "Self-awareness review Part 1: Do you “self-reflect” or “self-ruminate”?", *Science & Consciousness Review*, no. 1.
Moura, I. & Bonzon, P. 2004, "A Computational Framework for Implementing Baars Global Workspace Theory of Consciousness", *Brain Inspired Cognitive Systems 2004*. Scotland.
Muckli, L., Kriegeskorte, N., Lanfermann, H., Zanella, F.E., Singer, W. & Goebel, R. 2002, "Apparent Motion: Event-Related Functional Magnetic Resonance Imaging of Perceptual Switches and States", *Journal of Neuroscience*, vol. 22, no. 9, pp. 219RC.
Nagel, T. 1974, "What Is It Like To Be a Bat?", *The Philosophical Review*, vol. 83, no. 4, pp. 435-450.
Nielsen, H.F. & Chrysanthakopoulos, G. 2006, *Decentralized Software Services Protocol - DSSP*, Microsoft Corporation.
Nii, H.P. 1986, "Blackboard Application Systems, Blackboard Systems and a Knowledge Engineering Perspective", *AI Magazine*, vol. 7, no. 3, pp. 82-107.
Noë, A. 2002, "Is the Visual World a Grand Illusion?", *Journal of Consciousness Studies*, vol. 9, no. 5-6, pp. 1-12.
O’Regan, J. 2007, "How to Build Consciousness into a Robot: The Sensorimotor Approach" in *50 Years of Artificial Intelligence*, pp. 332-346.
O'Regan, J.K. 2010, "Explaining what people say about sensory qualia" in *Perception, Action, and Consciousness: Sensorimotor Dynamics and Two Visual Systems*, eds. N. Gangopadhyay, M. Madary & F. Spicer, Oxford University Press.
O'Regan, J.K. & Noë, A. 2001, "A sensorimotor account of vision and visual consciousness", *Behavioral and Brain Sciences*, vol. 24, no. 5, pp. 939-1031.
Parker, S.T., Mitchell, R.W. & Boccia, M.L. 2006, *Self-Awareness in Animals and Humans: Developmental Perspectives*, Cambridge University Press.
Penrose, R. 1994, *Shodows of the Mind*, Oxford Press, London.
Perner, J. & Lang, B. 1999, "Development of theory of mind and executive control", *Trends in Cognitive Sciences*, vol. 3, no. 9, pp. 337-344.
Pietarinen, A. 2002, "Awareness in logic and cognitive neuroscience", *First IEEE International Conference on Cognitive Informatics*IEEE, pp. 155-162.
Prem, E. 1997, "Epistemological Aspects of Embodied Artificial Intelligence", *Cybernetics and Systems*, vol. 28, no. 6, pp. 3.

Prinz, J. 2003, "Level-Headed Mysterianism and Artificial Experience", *Journal of Consciousness Studies*, vol. 10, pp. 111-132.
Raichle, M.E. 2006, "NEUROSCIENCE: The Brain's Dark Energy", *Science*, vol. 314, no. 5803, pp. 1249-1250.
Rakovic, D. 1997, "Prospects for Conscious Brain-like Computers: Biophysical Arguments", *Informatica (Slovenia)*, vol. 21, no. 3.
Ramachandran, V.S. & Hubbard, E.M. 2001, "Synaesthesia -- A window into perception, thought and language", *Journal of Consciousness Studies*, vol. 8, pp. 3-34.
Ramamurthy, U. 2008, "Might a LIDA Controlled Robot be Phenomenally Conscious?", *Proceedings of the Nokia Workshop on Machine Consciousness 2008*, pp. 32-33.
Ramamurthy, U., Baars, B., D'Mello, S.K. & Franklin, S. 2006, "LIDA: A Working Model of Cognition", *Cognitive Modeling*, eds. F.D.M. Danilo Fum & Andrea Stocco, Edizioni Goliardiche, , pp. 244.
Rao, A.S. & Georgeff, M.P. 1995, "BDI-agents: from theory to practice", *Proceedings of the First Intl. Conference on Multiagent Systems*San Francisco, pp. 312-319.
Rao, A.S. & Georgeff, M.P. 1991, "Modeling Rational Agents within a BDI Architecture", *Proceedings of the 2nd International Conference on Principles of Knowledge Representation and Reasoning*, eds. A. James, R. Fikes & E. Sandewall, Morgan Kaufmann publishers Inc.: San Mateo, CA, USA, pp. 473-484.
Rao, A.S., Georgeff, M.P. & Sonenberg, E.A. 1992, "Social Plans: A Preliminary Report", *Proceedings of the Third European Workshop on Modelling Autonomous Agents in a Multi-Agent World*. Elsevier Science B.V., pp. 57-76.
Raymond, J.E., Shapiro, K.L. & Arnell, K.M. 1992, "Temporary suppression of visual processing in an RSVP task: an attentional blink?", *Journal of Experimental Psichology. Human Perception and Performance.*, vol. 18, no. 3, pp. 849-860.
Rees, G., Kreiman, G. & Koch, C. 2002, "Neural correlates of consciousness in humans", *Nature Reviews Neuroscience*, vol. 3, no. 4, pp. 261-270.
Revonsuo, A. 2005, *Inner Presence: Consciousness as a Biological Phenomenon*, MIT Press, Cambridge, MA.
Riba, R. 1998, "How small is the circle? The question of animal consciousness.", *Mind Matters*, vol. 1, no. 1.
Richter, J. 2006, "Concurrent Affairs: Concurrency and Coordination Runtime", *MSDN Magazine*, , no. 10.
Rosenthal, D.M. 2005, "Thinking that One Thinks." in *Consciousness and Mind* Oxford University Press.
Rosenthal, D.M. 2000a, "Consciousness, Content, and Metacognitive Judgements", *Consciousness and cognition*, vol. 9, pp. 203-214.
Rosenthal, D.M. 2000b, "Metacognition and Higher-Order Thoughts", *Consciousness and Cognition*, vol. 9, no. 2, pp. 231-242.
Rosow, & Manberg, 2001, "Bispectral index monitoring.", *Anesthesiology Clinics of North America*, vol. 19, no. 4, pp. 947.
Rubia, F.J. (ed) 2000, *El cerebro nos engaña*, Ediciones Temas de Hoy, Madrid.
Samsonovich, A.V. & De Jong, K.A. 2005, "Designing a self-aware neuromorphic hybrid", *AAAI Technical Report WS-05-08*, eds. K.R. Thorisson, H. Vilhjalmsson & S. Marsela, AAAI Press, Menlo Park, CA, pp. 71-78.
Samsonovich, A.V. & Jong, K.A.D. 2004, "A General-Purpose Computational Model of the Conscious Mind", *Proceedings of the International Conference on Cognitive Modelling, ICCM 2004*, pp. 382.
Sanz, R., López, I. & Hernández, C. 2007, "Self-awareness in Real-time Cognitive Control Architectures", *AI and Consciousness: Theoretical Foundations and Current Approaches. AAAI Fall Symposium 2007*, pp. 136-141.
Sanz, R., López, I., Rodríguez, M. & Hernández, C. 2007, "Principles for consciousness in integrated cognitive control", *Neural Networks*, vol. 20, no. 9, pp. 938-946.
Sasaki, S., Hongo, T., Naitoh, K. & Hirai, N. 2008, "The process of learning a tool-use movement in monkeys, with special reference to vision", *Neuroscience Research*, vol. 60, no. 4, pp. 452-456.
Savage-Rumbaugh, S., Mintz Fields, W. & Taglialatela, J. 2000, "Ape Consciousness-Human Consciousness: A Perspective Informed by Language and Culture", *Integrative and Comparative Biology*, vol. 40, no. 6, pp. 910-921.
Schacter, D.L. 1989, "On the relation between memory and consciousness" in *Varieties of memory and consciousness: Essays in honor of Endel Tulving* Erlbaum Associates, Hillsdale, NJ, pp. 355-389.
Schacter, D.L., Reiman, E., Uecker, A., Roister, M.R., Yun, L.S. & Cooper, L.A. 1995, "Brain regions associated with retrieval of structurally coherent visual information", *Nature*, vol. 376, pp. 587-590.

Schnakers, C., Vanhaudenhuyse, A., Giacino, J., Ventura, M., Boly, M., Majerus, S., Moonen, G. & Laureys, S. 2009, "Diagnostic accuracy of the vegetative and minimally conscious state: Clinical consensus versus standardized neurobehavioral assessment", *BMC Neurology*, vol. 9, no. 1, pp. 35.
Searle, J.R. 1992, *The Rediscovery of the Mind*, MIT Press, Cambridge, Massachusetts.
Searle, J.R. 1980, "Minds, brains, and programs", *Behavioral and Brain Sciences*, vol. 3, no. 03, pp. 417-424.
Senkowski, D., Talsma, D., Grigutsch, M., Herrmann, C.S. & Woldorff, M.G. 2007, "Good times for multisensory integration: Effects of the precision of temporal synchrony as revealed by gamma-band oscillations", *Neuropsychologia*, vol. 45, no. 3, pp. 561-571.
Seth, A.K. 2009, "The strength of weak artificial consciousness", *International Journal of Machine Consciousness*, vol. 1, no. 1, pp. 71-82.
Seth, A.K. 2007, "Models of Consciousness", *Scholarpedia*, vol. 2, no. 1, pp. 1328.
Seth, A. 2009, "Explanatory Correlates of Consciousness: Theoretical and Computational Challenges", *Cognitive Computation*, vol. 1, no. 1, pp. 50-63.
Seth, A.K., Dienes, Z., Cleeremans, A., Overgaard, M. & Pessoa, L. 2008, "Measuring consciousness: relating behavioural and neurophysiological approaches", *Trends in Cognitive Sciences*, vol. 12, no. 8, pp. 314-321.
Seth, A.K., Izhikevich, E., Reeke, G.N. & Edelman, G.M. 2006, "Theories and measures of consciousness: An extended framework", *Proceedings of the National Academy of Sciences*, vol. 103, no. 28, pp. 10799-10804.
Seth, A., Baars, B. & Edelman, D. 2005, "Criteria for consciousness in humans and other mammals", *Consciousness and Cognition*, vol. 14, no. 1, pp. 119-139.
Shanahan, M. 2006, "A cognitive architecture that combines internal simulation with a global workspace", *Consciousness and Cognition*, vol. 15, no. 2, pp. 433-449.
Shanon, B. 2008, "A Psychological Theory of Consciousness", *Journal of Consciousness Studies*, vol. 15, pp. 5-47.
Shoemaker, S. 1982, "The Inverted Spectrum", *The Journal of Philosophy*, vol. 79, no. 7, pp. 357-381.
Siegwart, R. & Nourbakhsh, I.R. 2004, *Introduction to Autonomous Mobile Robots*, MIT Press.
Singer, W. & Gray, C.M. 1995, "Visual Feature Integration and the Temporal Correlation Hypothesis", *Annual Review of Neuroscience*, vol. 18, no. 1, pp. 555-586.
Singh, M.P. & Huhns, M.N. 2005, *Service-Oriented Computing. Semantics, Processes, Agents.* John Wiley and Sons.
Sloman, A. 2007, "Why some machines may need qualia and how they can have them: Including a demanding new Turing test for robot philosophers.", *AAAI Fall Symposium on AI and Consciousness: Theoretical Foundations and Current Approaches.* AAAI Press, pp. 9-16.
Sloman, A. & Chrisley, R. 2003, "Virtual Machines and Consciousness", *Journal of Consciousness Studies*, vol. 10, pp. 133-172.
Sloman, A. & Scheutz, M. 2002, "A Framework for Comparing Agent Architectures", *In UKCI'02: Proceedings of the UK Workshop on Computational Intelligence*, pp. 169-176.
Sloman, A. 2001, "Varieties of Affect and the CogAff Architecture Schema", *Proceedings Symposium on Emotion, Cognition, and Affective Computing AISB'01 Convention*, pp. 39.
Spence, C. & Squire, S. 2003, "Multisensory Integration: Maintaining the Perception of Synchrony", *Current Biology*, vol. 13, no. 13, pp. R519-R521.
Sperling, G. 1960, "The information available in brief visual presentations", *Psychological Monographs,* , no. 498, pp. 1-29.
Stanley, R.P. 2000, *Enumerative Combinatorics*, Cambridge University Press, Cambridge, UK.
Stening, J., Jacobsson, H. & Ziemke, T. 2005, "magination and abstraction of sensorimotor flow: Towards a robot model", *Proceedings of the AISB05 symposium on next generation approaches to machine consciousness*, eds. R. Chrisley, R. Clowes & S. Torrance.
Sugiyama, S. 2000, "Reflected method for having consciousness", *International Conference on Systems, Man, and Cybernetics.* IEEE, pp. 3141-3146.
Sun, R. 2006, "CLARION cognitive architecture: extending cognitive modeling to social simulation" in *Cognition and Multi-Agent Interaction. From Cognitive Modeling to Social Simulation.* Rensselaer Polytechnic Institute, New York.
Sun, R. 1997, "Learning, Action and Consciousness: A Hybrid Approach Toward Modelling Consciousness", *Neural Networks*, vol. 10, no. 7, pp. 1317-1331.
Takeno, J., Inaba, K. & Suzuki, T. 2005, "Experiments and examination of mirror image cognition using a small robot", *Proceedings of the 2005 IEEE International Symposium on Computational Intelligence in Robotics and Automation*, pp. 493-498.

Taylor, J.G. 1999, "Towards the networks of the brain: from brain imaging to consciousness", *Neural Networks*, vol. 12, no. 7-8, pp. 943-959.
Taylor, J.G. 2003, "The CODAM model of attention and consciousness", *Proceedings of the International Joint Conference on Neural Networks.* IEEE, pp. 292.
Tegmark, M. 2000, "Why the brain is probably not a quantum computer", *Information Sciences*, vol. 128, no. 3-4, pp. 155-179.
Thrun, S. 2000, "Probabilistic Algorithms in Robotics", *AI Magazine*, vol. 21, no. 4, pp. 93-109.
Tononi, G. 2008, "Consciousness as Integrated Information: a Provisional Manifesto", *The Biological bulletin*, vol. 215, no. 3, pp. 216-242.
Tononi, G. 2004, "An information integration theory of consciousness", *BMC Neuroscience*, vol. 5, no. 1, pp. 42.
Tononi, G. & Edelman, G.M. 1998, "Consciousness and Complexity", *Science*, vol. 282, no. 5395, pp. 1846-1851.
Towaga, T. & Otsuka, K. 1998, "A model of cortical neural network structure", *Engineering in Medicine and Biology Society. Proceedings of the 20th Annual International Conference of the IEEE*, pp. 2066-2069.
Tozour, P. 2002, "First-Person Shooter AI Architecture" in *AI Game Programming Wisdom*, ed. S. Rabin, Charles River Media, pp. 387-396.
Tran Duc Thao 1971, *Phenomenologie et materialisme dialectique*, Publications gramma, London, UK.
Trehub, A. 2007, "Space, self, and the theater of consciousness", *Consciousness and cognition*, vol. 16, no. 2, pp. 310-330.
Tsagarakis, N.G., Metta, G., Sandini, G., Vernon, D., Beira, R., Becchi, F., Righetti, L., Santos-Victor, J., Ijspeert, A.J., Carrozza, M.C. & Caldwell, D.G. 2007, "iCub: the design and realization of an open humanoid platform for cognitive and neuroscience research", *Advanced Robotics*, vol. 21, pp. 1151-1175.
Turing, A. 1950, "Computing Machinery and Intelligence", *Mind*, , no. 49, pp. 433-460.
Vanderwolf, C.H. 2000, "Are neocortical gamma waves related to consciousness?", *Brain research*, vol. 855, no. 2, pp. 217-224.
Ventura, D. 2001, "On the utility of entanglement in quantum neural computing", *International Joint Conference on Neural Networks.* IEEE, pp. 1565-1570.
Vinge, V., Brin, D. & Goertzel, B. 2009, "POLL: Is A Terminator Scenario Possible?", *h+ Magazine*, [Online], pp. 23 March 2010. Available from: http://www.hplusmagazine.com/articles/ai/poll-terminator-scenario-possible.
Vygotsky, L.S. 1980, *Mind in Society: The Development of Higher Psychological Processes*, Harvard University Press.
Wandell, B.A. 1995, *Foundations of Vision*, Sinauer Associates.
Wang, P., Goertzel, B. & Franklin, S. (eds) 2008, *Artificial General Intelligence 2008, Proceedings of the First AGI Conference, AGI 2008, March 1-3, 2008, University of Memphis, Memphis, TN, USA*, IOS Press.
Wei, Z., Wu, C. & Chen, L. 2000, "Symbolism and connectionism of artificial intelligence", *The 2000 IEEE Asia-Pacific Conference on Circuits and Systems, 2000. IEEE APCCAS 2000*IEEE, pp. 364-366.
Weiß, G. 1994a, "Neural networks and evolutionary computation. Part I: Hybrid approaches in artificial intelligence", *Proceedings of the IEEE International Conference on Evolutionary Computation*IEEE Press, pp. 268-272.
Weiß, G. 1994b, "Neural networks and evolutionary computation. Part II: Hybrid approaches in the neurosciences", *Proceedings of the IEEE International Conference on Evolutionary Computation*IEEE Press, pp. 273-277.
Weizenbaum, J. 1966, "ELIZA a computer program for the study of natural language communication between man and machine", *Commun.ACM*, vol. 9, no. 1, pp. 36-45.
Wierzbicka, A. 1986, "Human Emotions: Universal or Culture-Specific?", *American Anthropologist*, vol. 88, no. 3, pp. 584-594.
Wooldridge, M. 1999, "Intelligent Agents" in *Multiagent Systems: A Modern Approach to Distributed Artificial Intelligence*, ed. G. Weiss, The MIT Press, pp. 27-78.
Zeki, S., Watson, J., Lueck, C., Friston, K., Kennard, C. & Frackowiak, R. 1991, "A direct demonstration of functional specialization in human visual cortex", *Journal of Neuroscience*, vol. 11, no. 3, pp. 641-649.
Zhou, J. 2003, "Automatic detection of premature ventricular contraction using quantum neural networks", *Third IEEE Symposium on Bioinformatics and Bioengineering*IEEE, pp. 169-173.

Ziemke, T., Jirenhed, D. & Hesslow, G. 2005, "Internal simulation of perception: a minimal neuro-robotic model", *Neurocomputing*, vol. 68, pp. 85-104.
Zlatev, J. 2000, "The Mimetic Origins of Self-Consciousness in Phylo-, Onto- and Robotogenesis", *26th Annual Confjerence of the IEEE Industrial Electronics Society, 2000. IECON 2000.*