Hacia las Máquinas Conscientes

Ricardo Sanz

Autonomous Systems Laboratory
Universidad Politécnica de Madrid

Introducción

Este artículo propone una interpretación –quizá el bosquejo de una teoría– sobre los hechos que rodean el fenómeno de la consciencia con el propósito de aplicarla en la construcción de máquinas conscientes. Esta interpretación está necesariamente basada en la perspectiva disciplinaria del autor, los sistemas de control automático, dado que este es el objetivo final de esta línea de investigación. Aunque puede sonar –y suena– prepotente, la teoría de sistemas dinámicos, que subyace a la ingeniería de control, proporciona un marco conceptual muy adecuado para la descripción del fenómeno consciente biológico y la postulación de teorías que permiten interpretar muchos aspectos puntuales y globales de dicho fenómeno.

Esta teoría surge como efecto secundario del análisis realizado en nuestro grupo de investigación sobre el fenómeno de la consciencia y sobre algunas de las teorías existentes con el propósito de identificar mecanismos que pudieran servir para dotar de consciencia a las máquinas. La búsqueda de tecnología de consciencia artificial puede racionalizarse de muchas formas, pero baste decir que, dado que todos preferimos los taxistas conscientes a los inconscientes, debería ser igual para elegir entre robots conscientes o inconscientes.

En el estado actual del conocimiento sobre el fenómeno consciente, no parece posible decir si las máquinas lo son en la actualidad (de hecho parece imposible decidir sobre este mismo aspecto en muchos de los animales, tanto más próximos a nosotros). Pero quizá debido a ello, los ingenieros desean disponer de una teoría científica que permita decidir e implementar el tipo de consciencia que cada máquina concreta precise. Es por ello que estudiamos la consciencia humana y creemos posible realizar alguna contribución a la comprensión del fenómeno.

Algunas ideas básicas sobre control automático

El objetivo único del control automático es el dotar a las máquinas de los mecanismos necesarios para que realicen su función incluso cuando sufren perturbaciones no esperadas. Nuestra tecnificada vida está soportada por un gran número de sistemas de control automático que, aunque no observamos al quedar ocultos en el interior de las máquinas, constituyen, en la mayor parte de los casos, un componente integral de ellas.

Los ejemplos mas inmediatos son los controles de temperatura que se emplean en la calefacción de las casas —los famosos termostatos— o en el control de los frigoríficos. A veces un frigorífico nos sorprende haciendo un ruido gutural repentino. Esto es debido a que el sistema de control ha conectado el compresor. Esto lo hace porque,

habiendo detectado que la temperatura interior del frigorífico es mayor de lo deseado —por ejemplo si abrimos la puerta repetidamente— necesita que se eliminen algunas calorías del interior, para lo que emplea el circuito de refrigeración del frigorífico.

Este ejemplo es muy interesante porque nos permite describir algunos de los elementos fundamentales de todo sistema de control:

- El *sensor*: en nuestro caso un termopar o un bimetal, que mide la temperatura interior y permite decidir cuando hay que actuar.
- El *actuador*: en nuestro caso el compresor, que permite cambiar la situación que se quiere controlar.
- El *controlador* propiamente dicho: que establece la estrategia a seguir en la operación del actuador en función de los valores del sensor.

En el caso del frigorífico la estrategia de control es muy simple; se denomina estrategia *todo-o-nada*: si el frigorífico está por debajo de la temperatura objetivo establecida el usuario no se hace nada; si está por encima de la temperatura objetivo, se conecta el compresor. Este se desconectará cuando, a base de eliminar calor con el sistema de refrigeración, se vuelva a alcanzar la temperatura objetivo.

En los frigoríficos con congelador independiente, este sistema está duplicado: dos sensores, dos compresores, dos controladores y dos ruedas de ajuste para establecer la temperatura deseada en cada parte.

En el ámbito de los sistemas de control a veces se emplea el término controlador para referirse al conjunto sensor-controlador-actuador. Otras veces se considera que el sensor y el actuador, dada su característica física, forman parte del sistema controlado, al que se le suele denominar *planta* (ver Figura 1).

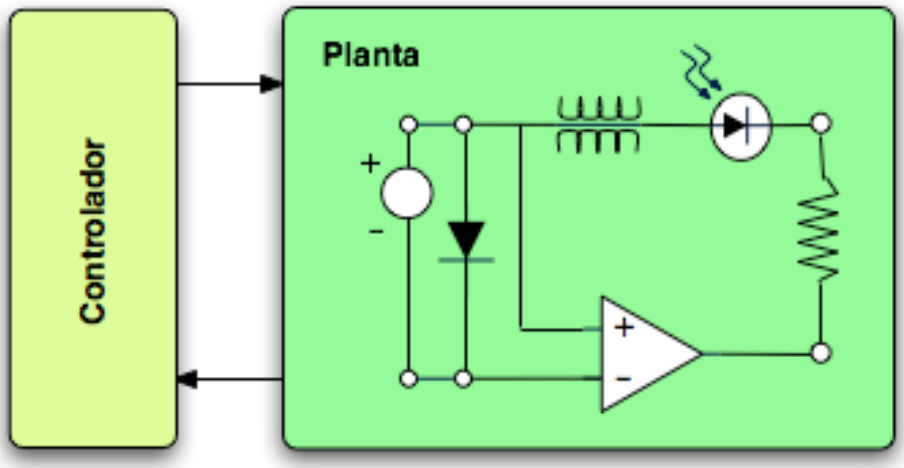

Figura 1: *En un sistema de control distinguimos dos partes: el subsistema que se controla, al que se denomina "planta" y el subsistema que controla, al que denominamos "controlador". El sensor y el actuador se encuentran entre ambos: el sensor proporciona información al controlador y el actuador proporciona acción —típicamente mediante aportación de masa o energía— a la planta.*

Esta forma de control constituye la estructura de control fundamental conocida como *retroalimenación*[1]. Sin embargo existen muchas otras estrategias de control mas complejas que permiten realizar un mejor control (*e.g.* optimizando el consumo de energía, realizando un control extremadamente preciso o llevando a la planta rápidamente al punto de funcionamiento) o controlar plantas mas complejas (*e.g.* cuando se tienen varias variables a controlar o cuando se quiere superponer un controlador a otro controlador mas elemental).

Hoy en día, casi todos los sistemas de control se implementan por medio de computadores, en los que el controlador es un programa informático que analiza los datos procedentes de los sensores y calcula los valores que deben aplicar los actuadores a la planta. La programación de este tipo de sistemas es una tarea de expertos, ya que requiere no sólo que el cálculo de la acción de control se realice de forma correcta, sino que, además, se calcule en el instante adecuado. Un cálculo correcto pero retrasado puede inutilizar un sistema de control e incluso producir daños catastróficos a la planta controlada. Por esta necesidad de comportamiento temporal correcto, a estos sistemas informáticos se los denomina sistemas de *tiempo-real*. El cómputo debe realizarse no en el tiempo virtual del computador sino en el tiempo de la realidad.

En el mundo de los sistemas industriales se emplean sistemas de control en todo tipo de instalaciones; de hecho muchas de ellas no pueden funcionar sin los sistemas de control: generadores eléctricos, refinerías, aviones, reactores nucleares, *etc*. En la

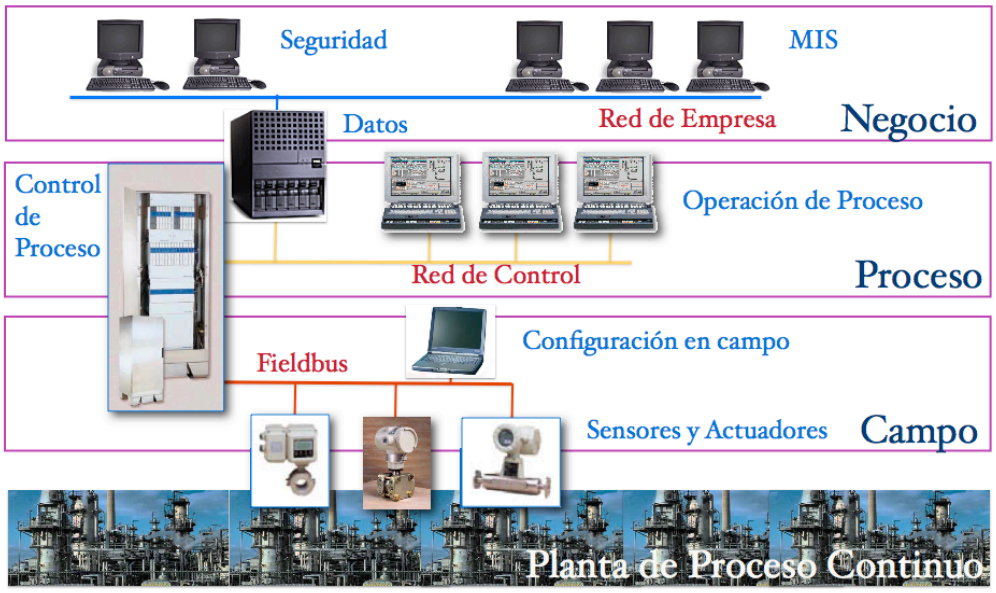

Figura 2: *Un ejemplo muy simplificado de un sistema de control de procesos continuos. El sistema se compone de muchos elementos interrelacionados, organizados en una estructura jerárquica. Se implementan estrategias de control en todos los niveles con diferentes características de abstracción, alcance y temporalidad. Los bucles de control del nivel inferior —nivel de campo— podrían denominarse* homeostáticos *y los del nivel mas alto —nivel de negocio—* cognitivos.

mayor parte de los casos, los sistemas de control industrial son mucho mas complejos que el sistema realimentado elemental de la Figura 1. Por ejemplo, la Figura 2 representa un sistema de control industrial en una planta química. En estas plantas se manipulan muchas sustancias para obtener algunos productos que en general se emplean como materias primas en otras fábricas: reactivos, disolventes, polímeros, etc.

[1] También se la conoce como *realimentación* o mediante el término inglés *feedback*.

Cada una de estas plantas está compuesta de gran cantidad de elementos y procesos que requieren de mecanismos de control para poder realizar la producción requerida: control de la temperatura en los reactores, control de la presión en los tanques de almacenamiento, control de la composición química de los productos, *etc*. Todo ello obliga a implantar múltiples bucles de control (cientos, miles en algunos casos) para controlar todas las magnitudes de la planta. Estos sistemas de control se organizan de forma jerárquica; desde los bucles de control elementales que controlan una presión o una temperatura en el nivel mas bajo, hasta los sistemas de control de la factoría que deciden que es lo que se debe producir y cuándo, dependiendo de las condiciones del mercado. Los sistemas de control intermedio se encargar de realzar el control de unidades —los diferentes órganos de la planta— que realizan tareas concretas —reactores, filtros, separadores, columnas de destilación, *etc*.— y que involucran a múltiples magnitudes del primer nivel.

Es interesante observar el estrecho paralelismo entre este tipo de sistema de control industrial y los mecanismos de control biológico del cuerpo humano. En el nivel mas bajo tenemos los sistemas de control homeostático que regulan magnitudes fisiológicas individuales (temperatura, presión sanguínea, azúcar en sangre, etc.). Sobre ellos se apoyan los sistemas que controlan los diversos órganos y sistemas y en el nivel mas alto se sitúan los sistemas de control cognitivo que regulan la actividad global del cuerpo humano. Este paralelismo es tanto mas interesante cuanto que los sistemas de control industrial complejo no se han inspirado en los sistemas de control biológicos para su estructuración, sino que esta ha surgido como una evolución técnica autónoma, desde los sistemas de control mas elementales.

Este mismo tipo de complejidad del sistema de control aparece también en los sistemas de control avanzado empleados en robótica. En este caso el paralelismo es menos sorprendente —dada la similitud corporal— máxime cuando muchos de los sistemas de control de robots son sistemas bioinspirados.

La búsqueda de la consciencia artificial

Pero, por muy complejos que sean los sistemas de control o los sistemas de procesamiento de información en general, existen capacidades que tienen los sistemas biológicos y de las que carecen los sistemas técnicos. Hay muchas de ellas que han sido objeto de debate —sobre todo en la literatura de inteligencia artificial— como es el caso de la capacidad de aprender o la capacidad de innovar. Sin embargo hay una característica que destaca, sobre todo desde el punto de vista de la capacidad del sistema para cumplir su misión y es la capacidad de *comprender* lo que sucede a su alrededor.

Obviamente, incluso el caso de humanos, no tenemos muy claro que quiere decir comprender lo que sucede, pero sería deseable que las máquinas fueran capaces de ello. Esta capacidad de comprensión de la realidad se ha venido asociando al concepto de *ser consciente* de lo que sucede en torno a uno, y así, la investigación sobre la mejora de las capacidades de las máquinas ha derivado en la búsqueda de la *consciencia artificial*. Deseamos que las máquinas se puedan dar cuenta de lo que está sucediendo globalmente, mas allá de la evolución de una serie de variables concretas.

En cierta medida, podríamos decir, que los sistemas de control actuales —de robots o de cualquier otra máquina— parecen operar en un nivel de automatismo alejado de lo ideal en condiciones de alta incertidumbre. En situaciones cambiantes, mal definidas, los sistemas de control se comportan habitualmente de forma demasiado rígida. Es por esto que en los sistemas industriales complejos siempre hay un ser humano en lo mas

alto de la jerarquía[2]. El humano bien preparado puede ser consciente de las posibles consecuencias de una determinada situación y obrar para minimizar los problemas o maximizar el éxito. Investigaciones clásicas de la inteligencia artificial en torno al *problema del marco* o el *sentido común* están en estrecha relación con esta capacidad.

La apariencia —por analogía con los seres humanos— es que los sistemas de control automáticos operan en modo *automático* (en el peor sentido del término), como diríamos de un ser humano que realiza su tarea *sin pararse a pensar*, *sin prestar la debida atención*, *sin ser consciente de lo que está haciendo*. Nadie quiere se operado por un cirujano que no presta la debida atención; o ser llevado al aeropuerto por un taxista que no se para a pensar; o que un maestro eduque a sus hijos sin ser consciente de lo que está haciendo.

Quizá sea un error o quizá solo una interpretación errónea de los hechos, pero un pequeño grupo de investigadores en distintos lugares del mundo, está persiguiendo el objetivo ambicioso de crear máquinas conscientes. De los múltiples estados de consciencia posibles en los seres humanos, quisiéramos que nuestras máquinas estuvieran, permanentemente, en un óptimo estado de vigilia[3].

En gran medida, la ambición de crear máquinas conscientes aparece con la inteligencia artificial. Gran parte de los investigadores del área han sido motivados por películas como *Planeta Prohibido*, *2001: Una Odisea espacial*, *La Guerra de las Galaxias* o el mismísimo *Terminator*. En todas ellas las máquinas son protagonistas no por su influencia en la trama sino porque constituyen los verdaderos *personajes* de la película. Personajes como personas, *i.e.* con todas esas características que nos hacen tan humanos (quizá no en *Terminator* pero obviamente sí en *Terminator II*).

Ahora hemos vuelto a aquella ambición de los 60, pero por necesidad. La máquina consciente, otrora un sueño de la inteligencia artificial, hoy aparece como un objetivo convergente de investigación en múltiples disciplinas que van desde la pura y clásica inteligencia artificial robótica, hasta los sistemas de control industrial mas avanzados o incluso la investigación biológica sobre la mente humana o la lucha contra el ciberterrorismo.

En este artículo, estudiaremos las motivaciones, comentaremos algunas de estas tendencias y formularemos algunas hipótesis sobre la investigación futura en el área de las máquinas conscientes, sus posibles resultados y el impacto potencial de esta tecnología emergente.

Sobre la necesidad de consciencia artificial

Si analizamos el espectro de actividades investigadoras en el ámbito de la consciencia artificial, observamos tres motivaciones principales que podríamos describir con los siguientes slogans:

a. *Artefactos como nosotros*: es la motivación básica que inspira a muchos de los constructores de robots. La frontera inalcanzada de la robótica es crear personas

[2] Aunque esto no es garantía de que se haga todo correctamente como lo demuestran las catástrofes —producidas por humanos— de Chernobyl o Bhopal. Cuanto más arriba en la jerarquía, mayor es la potencialidad para el daño.
[3] Vigilia frente a alternativas como el *sueño*, el *estupor*, el *coma* o la amplia variedad de *estados alterados* producidos por las drogas, el fútbol o la religión (aunque este caso es, cuando menos, discutible).

artificiales, que manifiesten consciencia, emoción y afecto, experiencia, estados fenoménicos, imaginación, *etc*. El paradigma disciplinario es, obviamente, la *robótica*.

b. *Modelos de sistemas naturales*: es lo que motiva a los investigadores en ciencias biológicas y humanas. Quieren crear implementaciones técnicas de laboratorio que plasmen en realidades sus modelos de la cognición humana. El paradigma disciplinario es la *ciencia cognitiva*.

c. *Máquinas más eficaces*: en este caso se buscan mecanismos de control que permitan obtener mejoras en el comportamiento de los sistemas técnicos en condiciones de alta incertidumbre. El paradigma disciplinario es en este caso el *control inteligente*.

Este último caso es el que mueve la investigación en nuestro Laboratorio de Sistemas Autónomos, aunque es obvio que los resultados de la investigación, de haberlos, serán de inmediata utilidad a los investigadores en robótica y posiblemente proporcionen ideas y modelos a los científicos cognitivos.

Existen múltiples sistemas técnicos que precisan que sus sistemas de control mejoren en esa capacidad de comprensión que mencionábamos antes. Por poner un ejemplo que está resultando tristemente célebre estos últimos años, podemos mencionar el caso de los sistemas de producción y distribución de energía eléctrica.

Este tipo de sistemas se organizan de forma distribuida y jerárquica, con sistemas de

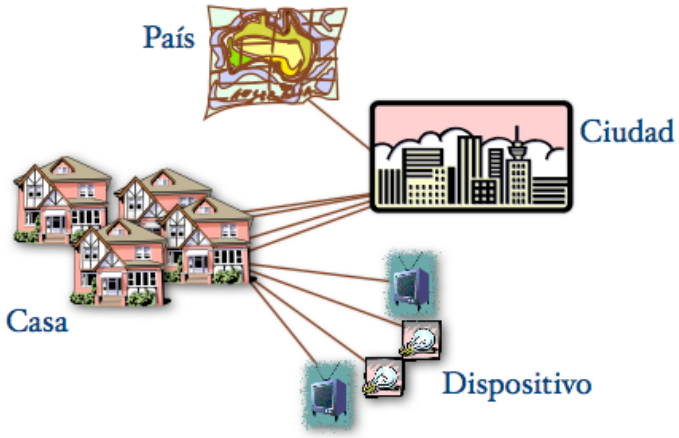

Figura 3: *En las redes de distribucion y consumo de energia eléctrica están apareciendo mecanismos de control inteligente en cada nivel de la jerarquía para optimizar el uso de las redes. Sin embargo es precisa la emergencia de una consciencia global que permita hacer frente a fenómenos sistémicos de gran escala.*

control en cada uno de los nodos que tienen objetivos locales y complementarios. El funcionamiento global del sistema y, por tanto, la prestación del servicio surge, en condiciones normales, de la conjugación de los comportamientos locales (ver Figura 3).

En condiciones anormales, por ejemplo cuando se produce el fallo de un generador o de una línea de transporte de energía, el sistema de control distribuido debe reaccionar para compensar la situación y que los usuarios de la red eléctrica no se vean afectados (se pondrán en marcha otros generadores, se redirigirá la energía eléctrica por otras líneas de alta tensión, *etc*). Esto casi siempre funciona bien pero hay situaciones que se pueden producir y que no se pueden valorar adecuadamente desde una perspectiva estrictamente local ni estrictamente general. Es preciso disponer de una visión global e integrada para poder darse cuenta de lo que esta sucediendo y obrar en consecuencia.

Ejemplos de este tipo de fenómenos se dan, por ejemplo, cuando no se dispone de mucho margen de maniobra (pocos generadores disponibles para emergencias) o cuando se produce una cascada de disparos de las protecciones de las líneas de distribución por la sobrecarga de éstas.

Es imposible manejar adecuadamente estas situaciones a menos que se disponga de la capacidad de *ser consciente de lo que esta sucediendo a nivel global*. Los humanos bien preparados son capaces de darse cuenta de la situación pero, lamentablemente, no son lo suficientemente rápidos como para poder manejar estas situaciones por ellos mismos. El resultado final es un apagón a gran escala que deja sin energía a grandes regiones durante períodos prolongados (dado que volver a recuperar el estado de funcionamiento es un proceso complicado de generación progresiva del difícil equilibrio que es un sistema eléctrico regional).

Desde un punto de vista estrictamente técnico, no se necesita implantar consciencia artificial al estilo humano salvo en aquellos casos en los que el sistema técnico deba ser considerado en cierta medida humano. Este será el caso de asistentes cognitivos que deben compartir la forma de pensar humana o en sistemas robóticos humanoides. Desde esta perspectiva, no es necesario analizar las posibilidades de emular sistemas con consciencia humana sino que lo que necesitamos es una implemenación limpia, técnicamente justificada de los mencanismos fundamentales de la autocosnciencia. En palabras de John McCarty "*the useful forms of computer agent self-awareness will not be identical with the human forms. Indeed many aspects of human self-awareness are bugs and will not be wanted in computer systems*".

Aparte de la aparente soberbia de crear mentes humanas, se pueden enumerar cuatro razones fundamentales para tratar de implementar *consciencia de acceso* —usando la terminología de Block— en sistemas artificiales no humanoides que les permitan ser auto-conscientes:

Prestaciones: Para maximizar las prestaciones es preciso que el sistema disponga de bucles de control interno que puedan ajustar su funcionamiento para utilizar óptimamente los recursos (que en una máquina están muy lejos de las 10^{10} neuronas o las 10^{15} sinapsis). Esta necesidad es manifiesta en los sistemas con aprendizaje declarativo o en controladores con optimización. Pero en realidad, mecanismos —elementales— de autoconsciencia ya aparecen en los sistemas computerizados mas simples: desde la computación consciente de la energía hasta los sistemas de middleware adaptativo-reflexivo o los planificadores con realimentación de los sistemas operativos de altas prestaciones.

Confianza: En los sistemas complejos de software se requiere en muchos casos —por parte de los usuarios humanos— que implementen mecanismos de justificación, *i.e.* de mecanismos que permitan explicarle al humano lo que el sistema está haciendo o como ha llegado a tomar determinadas decisiones. Para poder justificarse, el sistema necesita disponer de mecanismos de representación y de introspección que le permita representar y monitorizar sus procesos internos de razonamiento. Dos ejemplos paradigmáticos de esta situación son los sistemas con autodiagnosis y los sistemas basados en conocimiento; en particular los sistemas expertos.

Robustez: En muchos casos la adaptación automática no se realiza para mejorar las prestaciones sino para hacer frente a los cambios no deseados en el sistema informático o en el contexto de ejecución. Los sistemas informáticos robustos observan su propio comportamiento en su contexto de uso y son capaces de identificar anomalías y comprender qué es lo que está sucediendo en ellos mismos o en el entorno que les rodea en términos del servicio que deben proporcionar. Los sistemas deben disponer de introspección —autoconsciencia—

para poder identificar patrones de comportamiento anómalos y ser capaces de autodiagnosticarse e incluso autorepararse.

Coste: Los sistemas que se autogestionan son —a priori— mas económicos en términos de operación, al requerir de menos personal humano para su explotación y mantenimiento. Por ejemplo, IBM establece claramente el objetivo estratégico de su computación autonómica: *"Quite simply, it [autonomic computing] is about freeing IT professionals to focus on higher- value tasks by making technology work smarter, with business rules guiding systems to be self-configuring, self-healing, self-optimising, and self-protecting"*. La reducción de coste por la eliminación del personal humano en sistemas que deben tener un alto nivel de disponibilidad solo es posible si los sistemas son auto-conscientes.

Todos estos enfoques se basan en la implementación de mecanismos de introspección y de bucles de control interno —*homeostáticos* que diría Cannon— dentro del sistema hardware-software para mantenerlo en funcionamiento óptimo. Es interesante observar las similitudes entre las arquitecturas de sistema que estas necesidades plantean y los sistemas de control inteligente avanzado industrial.

Una visión sistémica de la consciencia

Desde una perspectiva de ingeniería de sistemas de control, la consciencia nos proporciona —o quizá es— una capacidad de prestar la atención debida a las cosas que nos rodean. Para Taylor, sin el mecanismo de atención nuestro cerebro quedaría colapsado por el flujo informacional procedente de nuestro sentidos. Mediante la consciencia somos capaces de serializar nuestro pensamiento en torno a determinados hechos de relevancia para nosotros. Las diferencias que hacen una diferencia, que dijo Bateson. La consciencia es el último árbitro y canal de nuestra necesidad de saber.

Desde el punto de vista de los sistemas autónomos hay tres aspectos básicos en esta necesidad de saber:

- Saber del **mundo**, o lo que es lo mismo, *percibir el estado del mundo* para poder actuar adecuadamente en él. Tuviera Dawkins razón o no en que tan sólo somos vehículos de nuestro genes, podemos decir, sin género de duda, que estos vehículos necesitan ver el camino para poder llevar los genes a destino.

- Saber del **yo**, o lo que es lo mismo conocer *el propio estado* y *la propia capacidad* para obrar en el mundo en persecución de nuestros fines.

- Saber de **otros**, o lo que es lo mismo *conocer el estado físico y sobre todo mental de otros* con los que interaccionamos porque, como agentes activos, pueden determinar o constreñir enormemente nuestro deambular por el mundo.

Desde este análisis somero de las necesidades del sistema de percepción en tiempo real de un sistema autónomo, podemos decir que la consciencia presenta tres aspectos básicos o tres tipos de consciencia:

- **Perceptual**: la capacidad de recibir información del mundo, tanto exterior (percepción) como interior (propiocepción) constituye el mecanismo fundamental por el que un sistema autónomo se encuentra conectado a la realidad.

- **Autoconsciencia**: la capacidad de introspección y observación de los propios mecanismos de pensamiento. El desarrollo del concepto de "yo" y su uso en los

procesos mentales es básico para la realización robusta de tareas y para la interacción con otros agentes cognitivos.

- **Qualia**: para muchos el aspecto crucial del fenómeno consciente es la capacidad de tener experiencias fenoménicas, incluso la de experimentar emociones.

El fenómeno de la consciencia nos muestra que hay distancia entre *fotografiar* el mundo y *ver* el mundo. Pudiérase decir que hay tres variantes de interacción de entrada con la realidad para un agente cognitivo: a) *medir* del mundo por medio de nuestros sentidos, b) *percibir* el mundo por medio de nuestros sistemas de interpretación y generación de sentido y c) *experimentar* el del mundo por medio de nuestros mecanismos neurales fenomenológicos —cualquiera que estos sean— que tanto se buscan en el mundillo de los correlatos neuronales de la consciencia.

Nos podemos replantear la pregunta de ¿por qué construir máquinas conscientes?, que tratamos de justificar someramente con anterioridad, desde esta triple perspectiva de la consciencia —perceptual, autoconsciencia y qualia.

La experiencia del mundo proporciona funcionalidad básica a un sistema como demuestra —matemáticamente— la teoría de control. La retroalimentación sensorial es la única forma práctica de hacer frente a la incertidumbre del mundo y para poder tenerla se precisan sensores y mecanismos perceptuales que los utilicen.

La observación y experiencia del yo —y de los otros, que para el caso cabe agruparlas bajo el mismo paraguas— proporciona los fundamentos para el comportamiento autónomo robusto. Del mismo modo, la experiencia de los otros proporciona mecanismos de planificación multiagente profunda que ayudará a la realización de tareas tanto en cooperación como en competición.

Respecto a la fenomenología y los qualia, la vieja discusión sobre si se trata de epifenómenos o no, se traslada al campo de la consciencia artificial en la forma de una falta de justificación funcional clara. Nadie puede decir para qué serviría que una máquina tuviera qualia —excepto, posiblemente, para poder comunicar la propia experiencia a otros (con lo que habríamos de asociar fenomenología y socialización).

También podríamos responder a la pregunta ¿por qué máquinas conscientes? al estilo gallego: ¿y por qué no ?. Esto nos lleva a considerar los posibles problemas éticos derivados de esta actividad. No vamos a entrar en ello —la ética no entra dentro de las competencias estándar del ingeniero— pero el filósofo alemán Thomas Metzinger afirma que llegado cierto punto en la implementación de mecanismos de autoconsciencia, las máquinas pueden llegar a sufrir: *"suffering starts at the level of phenomenal self model (PSM)"*. Es por ello que el filósofo recomienda ciertas cotas éticas al desarrollo de tecnología de consciencia artificial, limitándonos a la implementación de máquinas que no tengan este problema: *"we should ban all attempts to create artificial and postbiotic PSMs"*.

Antes dimos algunas de las razones para buscar consciencia artificial (artefactos como nosotros, modelado de sistemas naturales o construcción de máquinas más eficaces). Pero, aparte de los posibles problemas éticos mencionados por Metzinger, otros sostienen la imposibilidad real de construir tales máquinas. Los argumentos son variados y heterogéneos; algunos con las mismas raíces de la crítica a la inteligencia artificial al estilo de Dreyfus o Searle. Estos opositores de base podríamos clasificarlos en: a) misterianos y demás familia, que creen profundamente en la naturaleza especial del ser humano, b) los que defienden la emergencia de la consciencia en mecanismos autopoiéticos de auto-organización en sistemas biológicos cognitivos complejos y c) técnicos del experimento mental y la retórica que presentan argumentos como los de

de *qualia ausente*, diversos tipos de zombies, neurocientíficos disminuidos sensoriales o naciones chinas.

El problema duro —por usar la terminología de Chalmers— de la producción de fenomenología sintética todavía permanece vivo, candente; en parte debido a la carencia de motivación técnica pero principalmente debido a la carencia de teorías plenamente aceptadas sobre los correlatos neuronales de la consciencia biológica. Los programas de investigación al estilo del de Koch prometen proporcionar esta explicación, pero, hasta ahora la impresión más generalizada es que se puede realizar una implementación (o dar una explicación) de la funcionalidad sensorial-perceptual o incluso del "yo", pero no se puede —con el conocimiento actual— realizar una implementación de la capacidad experiencial.

Aunque el *lugar de la consciencia* está siendo sistemáticamente buscado con técnicas experimentales mejoradas —sobre todo de neuroimagen—aún se está lejos de alcanzar algún resultado objetivo globalmente aceptado. La prueba tangible de esta carencia de explicación *oficial* la tenemos en la aparición de teorías *anormales*, que reclaman un fundamento nuevo para explicar la consciencia, siendo los fenómenos de coherencia cuántica a nivel cerebral el caso mas flagrante. La racionalización irónica que hace Steven Pinker de este enfoque cuántico no deja de ser brillante: *Si la consciencia es misteriosa y la mecánica cuántica es misteriosa entonces tienen algo que ver*. En los últimos tiempos, lo que en cierta medida se está aceptando como evidente es que la búsqueda del lugar de la consciencia —al estilo del área de Broca o el córtex visual— está destinada al fracaso y que se precisan modelos sistémicos globales.

En resumidas cuentas, la construcción de máquinas conscientes se enfrenta al problema de que todavía *no se tiene claro qué es la consciencia biológica*. Sin embargo, la investigación en consciencia artificial puede contribuir a mejorar este conocimiento mediante aportaciones en esta línea de modelos sistémicos globales (al fin y al cabo esta es la especialidad de la ingeniería de sistemas) y la clarificación de conceptos teóricos fundamentales. En este contexto, analizaremos algunas de las estructuras técnicas comunes en sistemas de control y veremos como la evolución de los diseños en busca de la pura mejora técnica —en los ya mencionados aspectos de prestaciones, confianza, robustez y coste— nos lleva a estructuras que son singularmente similares a las postuladas en el caso de la mente natural.

Evolución de los sistemas de control

Los sistemas de control tienen como objetivo el lograr que un sistema físico tenga un comportamiento determinado, prefijado de antemano. Habitualmente este comportamiento deseado es diferente del que tendría el sistema abandonado a su suerte. Por ejemplo, en el caso de un frigorífico sin sistema de control (esto es, sin algo que conecte/desconecte el compresor) acabará o bien con el interior a temperatura ambiente o bien totalmente congelado a la mínima temperatura que se pueda alcanzar (un comportamiento totalmente análogo al de una calefacción sin termostato: al final se está a la temperatura de la calle o con un calor insoportable). Podríamos decir que el sistema de control es lo que mantiene al sistema físico en una situación útil pero

Figura 4: *El sistema de control retroalimentado mas elemental está compuesto por un sensor y un actuador gobernado por los valores del sensor. Es el caso del frigorífico, en el que el sensor de temperatura — usualmente un bimetal— conecta y desconecta el compresor.*

físicamente inestable (por ejemplo un frigorífico manteniendo un desequilibrio térmico en un punto concreto de temperatura). La forma mas elemental de control se establece por medio de un sensor y un actuador que actúa en base al valor medido por dicho sensor (Figura 4).

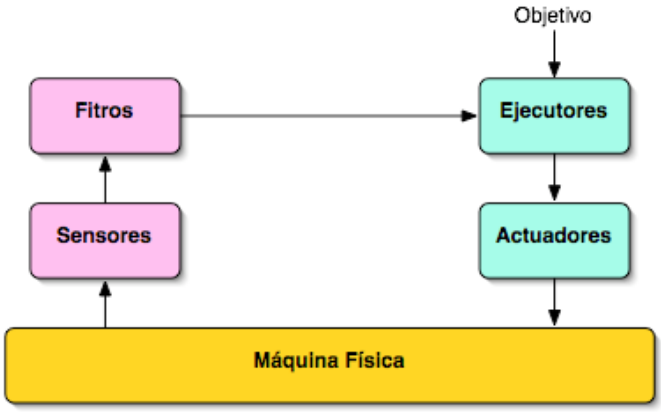

Figura 5: *Un control mejorado con preprocesamiento de datos sensoriales, toma de decisiones de control y establecimiento de objetivos de control a perseguir.*

Un control ligeramente mas avanzado permite por una parte el establecimiento de consignas de operación —objetivos de control— y por otra la eliminación de información innecesaria en el flujo de datos de los sensores por medio de filtros y la mejora en la toma de decisiones de control a ejecutar por medio de los actuadores (Figura 5).
En algunos casos el sistema de control necesita tener algo de memoria para poder operar correctamente. En estos casos se dice que el control debe tener un cierto *estado* dinámico que le permita saber qué debe hacer después.

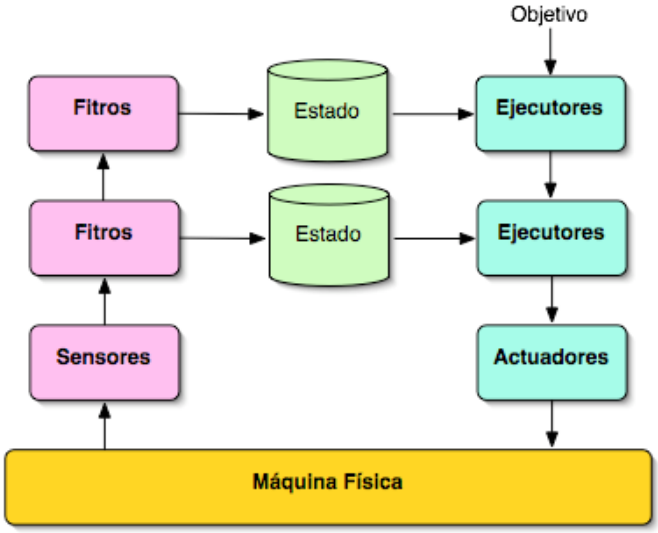

Figura 6: *Un control jerárquico en el que dos controladores se anidan para lograr un objetivo complejo de control.*

Otra variación estructural importante se produce cuando se emplea un sistema de control anidado. En este caso un sistema de control externo controla un sistema que dispone ya de un sistema de control. Esto se produce por ejemplo, cuando se emplea un control de presión sobre un sistema que ya dispone de un control de temperatura. Ambos controles podrían ser independientes, pero también puede darse la

circunstancia de que exista una dependencia jerárquica ente ellos. Este es el caso del complejo sistema de control de la Figura 2 o mas sencillamente, el representado en la Figura 6.

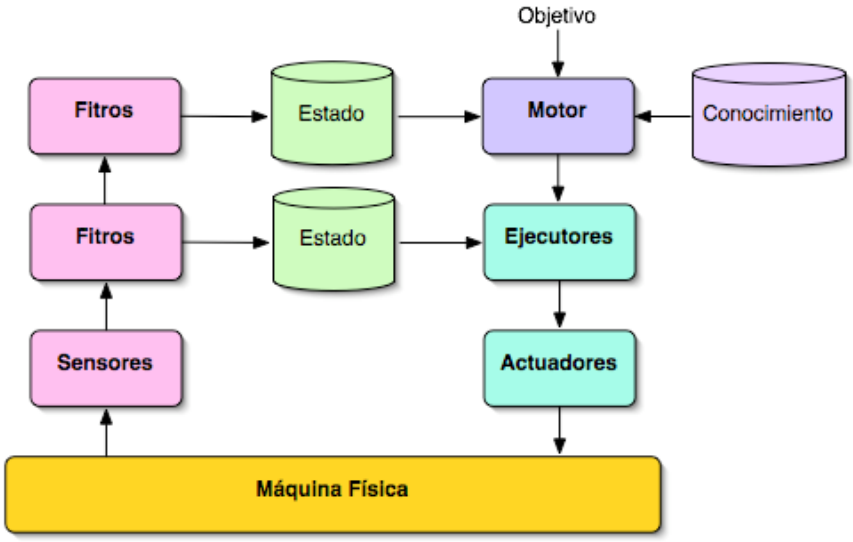

Figura 7: *Un sistema de control jerárquico en el que el controlador de mas alto nivel es un controlador basado en conocimiento (se separa el conocimiento concreto de la tarea del mecanismo de aplicarlo que puede ser genérico).*

Una variante interesante se produce cuando la estrategia específica de control a seguir para un proceso concreto se describe de forma declarativa y es aplicada por un sistema genérico. En la Figura 7 vemos un sistema de este tipo, en el que el ejecutor de un bucle de control se divide en dos entes diferentes: el *conocimiento* de control y el *motor* de aplicación de dicho conocimiento a la tarea de control.

Esta separación es similar a la empleada en los sistemas expertos, una clase de sistemas basados en el conocimiento en los que un motor de inferencia genérico aplica conocimiento declarativo específico del problema hasta alcanzar unas conclusiones. En

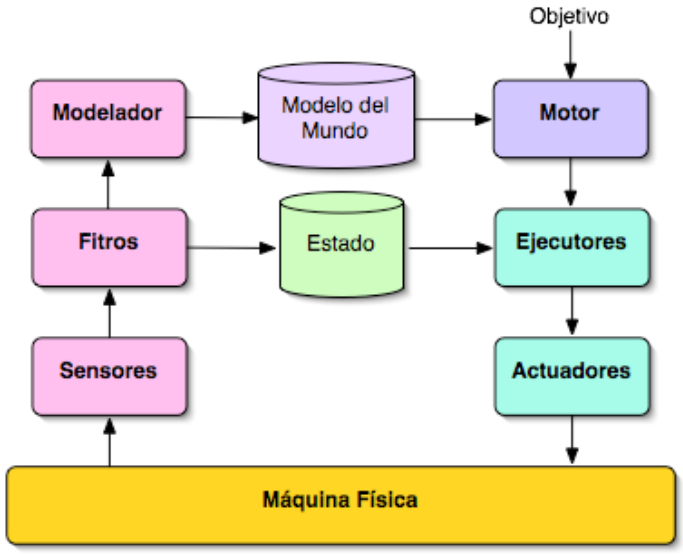

Figura 8: *En un control con modelo se integra la información perceptual con la memoria para mantener una representación actualizada de una parte de la realidad que resulta de interés para el controlador —un modelo del mundo.*

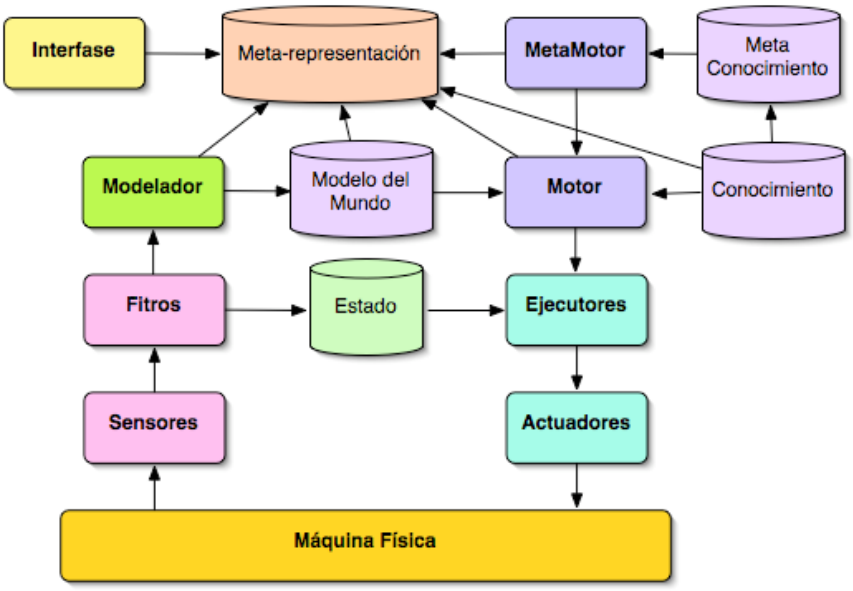

Figura 9: *La representación del mundo puede incluir al propio sistema de control. Se dice en este caso que se tiene una metarepresentación y que se emplea metaconocimiento para realizar actuaciones sobre el propio sistema de control.*

nuestro casos esas conclusiones son acciones a realizar sobre el sistema controlado y el proceso de inferencia debe realizarse en tiempo-real. En esta situación podemos decir que el sistema de control pasa de ser puramente reactivo a ser deliberativo[4].

Como se aprende rápidamente, es conveniente almacenar toda la información que se obtiene sobre y a través del cuerpo físico de una forma integrada constituyendo un *modelo del mundo* (Figura 8). Así en un sistema con múltiples sensores se puede fusionar la información de varios para obtener un dato de mejor calidad o un dato imposible de obtener con un solo sensor. Los ejemplos abundan: la percepción estereoscópica, la percepción de la velocidad de cambio, *etc*.

El uso de los modelos puede extenderse hasta la propia representación del sistema de control dentro del modelo del mundo —o en un modelo separado de más alto nivel (Figura 9). En este tipo de sistemas se emplean los automodelos junto con *metaconocimiento* del propio sistema para implementar mecanismos de introspección y reflexión que permitan al sistema de control realizar algunas tareas de adaptación.

Como vemos, y en cierta medida, se observa que la evolución de los controladores complejos replica la evolución de la mente biológica; desde el sistema autonómico y el cerebro reptiliano hasta los sistemas cognitivos y conscientes de más alto nivel.

La complejidad creciente de los sistemas de control, que se implementan por medio de agentes semiautónomos en todos los roles de las jerarquías de control —sobre todo en el casos de sistemas distribuidos de control—nos ha conducido a una situación en la que tenemos un gran problema: no hay teorías ni métodos formales usables para sistemas de tiempo-real con agentes cognitivos que me permitan tener una seguridad en que su funcionamiento va a ser correcto. Es por ello que los sistemas de control complejos con agentes cognitivos son poco robustos.

La teoría y la práctica del control han desarrollado diversas estrategias para incrementar la robustez de los sistemas de control:

[4] Aunque el sistema de control reactivo que tiene un estado dinámico es también, aunque implícitamente, un sistema que almacena y usa conocimiento.

- **Control adaptativo**: Hace frente a cambios en la planta debidos a derivas en la misma —alteraciones físicas que hacen que la planta vaya cambiando con el paso del tiempo.

- **Control robusto**: El sistema de control se calcula para tolerar pequeños desplazamientos de las condiciones de diseño.

- **Control tolerante a fallos**: Puede incrementar la robustez hasta un cierto límite, haciendo frente a cambios en la planta debidos a fallos —habitualmente por medio de redundancias.

Frente a estas estrategias mas o menos convencionales, se plantea una nueva posibilidad al emplear mecanismos de introspección para hacer **control reflexivo**.

Añadir mecanismos de reflexión supone incorporar representaciones de (algunos aspectos de) el propio sistema de control así como metaconocimiento. Este modelo y conocimiento se emplea en implementar un control autoreflexivo basado en modelos en el que la auto-representación está causalmente conectada con la operación del propio sistema de control. El sistema mantiene una representación precisa de sí mismo que emplea para controlarse. El paradigma de investigación es el control *reflexivo predictivo multiresolucional basado en modelos*.

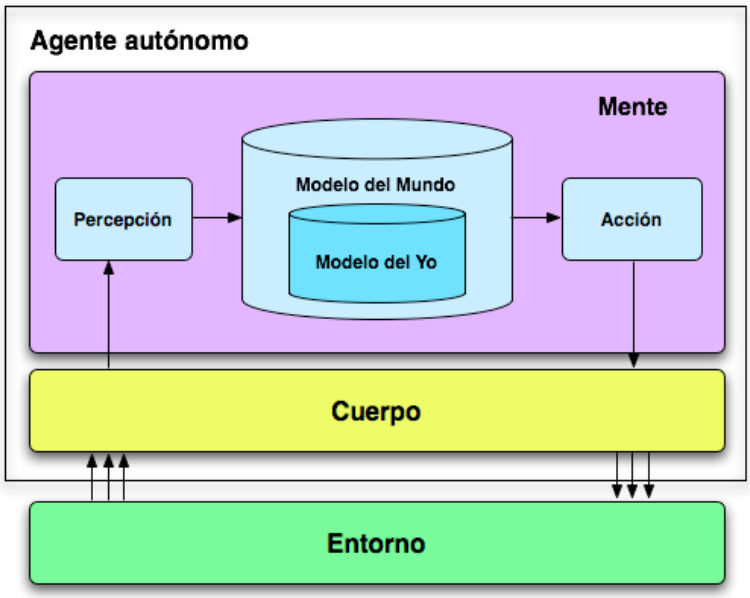

Figura 10: *Los agentes autónomos incorporan modelos de sí mismos en sus modelos del mundo para poder autogestionarse.*

En resumen, podríamos decir que se está en el camino de crear controladores (auto)conscientes, aunque la mayoría de los ingenieros de control se negaría a reconocerlo en público (o incluso en privado).

Máquinas conscientes

La investigación en máquinas conscientes es algo relativamente reciente, en parte por la relegación de la consciencia natural en sus propios ámbitos, como por la falta de ideas claras sobre el fenómeno natural.

Sin embargo, desde hace algunos años, se observa un interés creciente en este tema debido en parte a la reconsideración que la investigación en consciencia ha sufrido y en parte por una sorprendente confluencia de intereses y tópicos en diversas áreas: redes neuronales globales, robótica cognitiva, modelado de emociones, modelos cerebrales, sistemas anti-intrusión, sistemas autonómicos, protección de infraestructuras, control robusto, etc.

La mayor parte de la investigación en consciencia mecánica se da en los ámbitos de la robótica y la ciencia cognitiva. Pero surge una duda, ¿qué es, exactamente, consciencia mecánica?. La respuesta estándar es bastante elusiva: *conciencia mecánica es lo que quiera que una persona piense que es consciencia, pero sucediendo en una máquina* y no en un animal. Una respuesta similar a la resultante del viejo debate sobre la inteligencia artificial dura.

La literatura clasificatoria de los tipos de consciencia está en permanente producción y las clasificaciones abundan; por ejemplo Armstrong los clasifica en "*minimal*, *perceptual* and *introspective* consciousness", mientras que Anthony distingue entre "*phenomenal* consciousness, *access* consciousness, *state* consciousness, *creature* consciousness, *introspective* consciousness and *self*-consciousness".

Desde nuestra perspectiva de lo artificial, y en base a nuestro intereses objetivos, la clasificación que hacemos divide los mecanismos de consciencia en tres grandes tipos: percibir el mundo, percibirse a sí mismo y experimentar qualia. En el mundo de la consciencia artificial son interesantes los cinco axiomas —o síntomas— que propone Aleksander para determinar si una máquina es consciente: percepción, imaginación, atención, planificación y emoción.

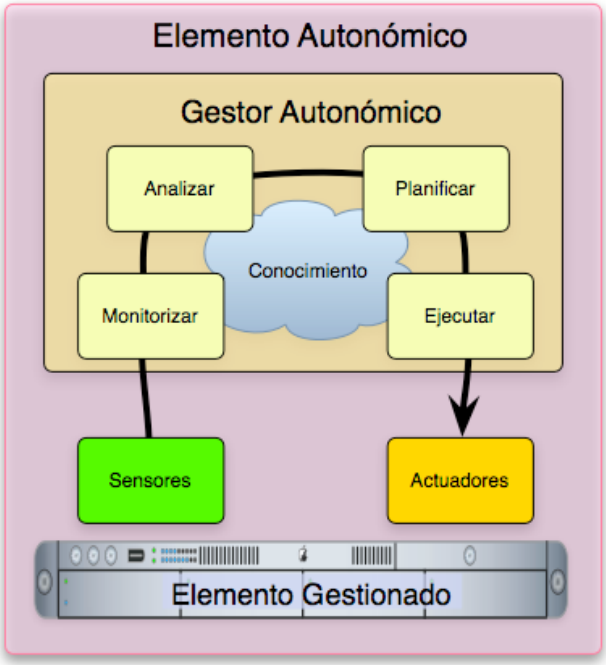

Figura 11: *La visión del cómputo autónomico de IBM se basa en la implantación de bucles de control homeostático en la infraestructura de los sistemas de información.*

Hoy en día hay varios intentos de avanzar en la implementación de consciencia artificial. Franklin en la Universidad de Memphis construye un agente "consciente" para manejo de asignaciones en la marina norteamericana siguiendo la *Global Workspace Theory* de Baars. Aleksander en el Imperial College de Londres trata de

escalar arquitecturas neuronales hasta el nivel de consciencia. Taylor en el King's College focaliza su trabajo en los modelos de atención consciente mediante redes neuronales. Manzotti en la Universidad de Milán trata de evolucionar una arquitectura cognitiva para robots basado en una nueva ontología perceptual que incluye la consciencia. Holland en la Universidad de Essex construye robots con control basado en modelos autoreferentes.

Aunque no específicamente centrados en la construcción de consciencia artificial, los postulados que guían la tecnología de cómputo autonómico de IBM, son cercanos a la motivaciones de la consciencia artificial: "*to design and build computing systems capable of running themselves, adjusting to varying circumstances, and preparing their resources to handle most efficiently the workloads we put upon them. These autonomic systems must anticipate needs and allow users to concentrate on what they want to accomplish rather than figuring how to rig the computing systems to get them there*".

La visión de la computación autonómica se basa en el comportamiento del sistema nervioso autonómico. Se busca la creación de sistemas que se manejan a sí mismos de acuerdo con objetivos prefijados por el administrador. Los nuevos componentes se integran sin esfuerzo en un sistema autonómico de la misma forma que las células nuevas se integran en un cuerpo vivo. Los objetivos del cómputo autonómico se centran en dotar a los sistemas informáticos cuatro capacidades básicas: autoconfiguración, autooptimización, autocuración y autoprotección.

Siguiendo a Holland, podemos decir que en la ingeniería de la máquina consciente se pueden seguir múltiples estrategias, pero destacan dos: de inspiración biológica y de inspiración técnica.

En la estrategia de inspiración técnica, trataremos de diseñar y construir el robot más inteligente que podamos, con la misión más precisa en el entorno más preciso y —con suerte— será consciente. En general esta estrategia está condenada al fracaso porque habitualmente desconocemos la complejidad de la misión y del entorno —esa es una de las motivaciones para buscar consciencia— y además no sabremos a priori si el robot es suficientemente inteligente. Un problema adicional de esta estrategia es que, incluso en el hipotético caso de que el robot fuera consciente y pudiéramos demostrarlo, no sabríamos por qué (dada la estrategia de construcción de caja negra).

En la estrategia de inspiración biológica, trataríamos de construir un robot tonto, con una misión sencilla, en un entorno sencillo. Si esta estrategia tiene éxito, el siguiente paso es hacer la misión y el objetivo mas difíciles hasta que el robot deje de tener éxito. Después incrementaremos la inteligencia del robot hasta que vuelva a tener éxito y por iteración de este proceso hasta alcanzar la consciencia. En cierta medida esta estrategia es mejor que la anterior porque posiblemente refleja lo que la evolución ha hecho con nuestra propia consciencia y porque podemos empezar ya mismo —porque lo que de verdad sabemos construir son robots demasiado tontos para ser conscientes. Una ventaja adicional es que si conseguimos que aparezca la consciencia tendremos una oportunidad de saber como y por qué.

En la literatura se describen muchas otras estrategias: construir una réplica del cerebro humano neurona a neurona, implementar facultades atómicas usando redes neuronales, implementar teorías psicológicas sobre la consciencia, *etc.* Nuestra estrategia se basa en desarrollar mecanismos de control basado en modelos que incorporen el propio control del controlador. Una estrategia a mitad de camino, podría decirse. Controles tontos pero no tanto.

Nos proponemos desarrollar y usar teoría de control autoreflexivo basado en modelos (también conocido como enfoque de imaginación o simulación en otros ámbitos más biológicos, por ejemplo Hesslow o Ziemke). Frente a aprender los modelos de la

interacción con la realidad (incluyendo la realidad interior) buscamos el uso de un proceso basado en diseño profundo que permita dotar al sistema técnico de un nivel de autoconsciencia muy superior —en calidad— al humano.

Conclusiones

Nos encontramos en el camino *hacia la consciencia artificial movidos por múltiples razones tanto científicas* (empezar a vislumbrar una posible explicación a la consciencia) *como puramente técnicas* (se precisan arquitecturas mentales mas eficaces para entornos de alta incertidumbre). Se está retomando el viejo sueño de la inteligencia artificial de los años 60: fabricar personas.

En este contexto intelectual existen muchas palabras calientes y es perfectamente posible que, el intento de implementación de teorías que supone la investigación en máquinas conscientes contribuya a una clarificación de los conceptos que hay tras estas palabras: percepción, significado, valor, consciencia, autoconsciencia, yo, emoción, imaginación, qualia, sabiduría, *etc.* caerán de su pedestal de confusión.

Este es un campo muy nuevo enfrentado a un problema de miles de años de antigüedad, y por ello existen muchas cuestiones abiertas:

- ¿Es posible la replicación de la consciencia humana en una máquina?
- ¿Qué tipos de consciencia sirven para algo?
- ¿Es ético construir máquinas conscientes?
- ¿Cual es el mejor camino? ¿La ontogénesis o el diseño?
- ¿Son posibles otras consciencias ?
- ¿Cómo será una mente alienígena?

En los últimos años se han venido celebrando diversas reuniones internacionales[5] sobre la consciencia artificial pero todavía está por consolidar una comunidad y establecer unos medios de diseminación de los avances.

Se puede obtener más información en el website de Holland sobre máquinas conscientes http://www.machineconsciousness.org, o en los websites de algunos de los eventos que hemos organizado (como http://www.aslab.org/public/events/moc, o http://www.aslab.org/public/events/mcc).

Queda mucho por hacer y mucha incertidumbre por disipar; y, aunque podamos tener ciertas ideas sobre como hacer máquinas conscientes, es posible que nos quede siempre una duda. Parafraseando a Nagel, podríamos plantearnos la pregunta clave: *What is it like to be a model-based reflective predictive controller?*

Referencias

Aleksander, I. y Barry Dunmall, B. (2003) Axioms and Tests for the Presence of Minimal Consciousness in Agents. *Journal of Consciousness Studies*, Vol 10, No. 4-5.

Baars, B. (1997) *In the Theatre of Consciousness*. Oxford University Press.

[5] Cold Spring Harbor, USA (2001), Skövde, Suecia (2001), Memphis, USA (2002), Birmingham, UK (2003), Turin, Italia (2003), Antwerp, Bélgica (2004), Hertfordshire, UK (2005) y Agrigento, Italia (2005).

Block, N.(1995) On a confusion about a function of consciousness. *Behavioral and Brain Sciences*, 18:227-287.

Brachman, R.J. (2002). Systems that know what they're doing. *IEEE Intelligent Systems*, 17(6):67–71.

Canon, W. (1939) *The Wisdom of the Body*. W.W. Norton.

Chalmers, D. (1997*) The Conscious Mind: In Search of a Fundamental Theory*. Oxford University Press.

Doyle, J. (1980). *A model for deliberation, action, and introspection*. Technical Report AI TR 581, Massachussets Institute of Technology Artificial Intelligence Laboratory.

Forrest, S., Hofmeyr, S.A., Somayaji, A. y Longstaff, T.A. (1996). A sense of self for a Unix processes. In Proceedings of 1996 *IEEE Symposium on Computer Security and Privacy*.

Franklin, S., y A. Graesser. A. (1999) A Software Agent Model of Consciousness. *Consciousness and Cognition* 8:285-305.

Hesslow, G. (2002) Conscious thought as simulation of behaviour and perception. *Trends in Cognitive Science* 6, 242–247.

Holland, O. (2004). *Machine Consciousness*. Imprint Academic.
Horn, P (2001). *Autonomic computing: IBM'sperspective on the state of information technology*. IBM.

Koch, C. (2004) *The Quest for Consciousness: a Neurobiological Approach*. Roberts and Co.

Metzinger, T.(2003). *Being No One*. MIT Press.

Nagel, E. (1974). What is it like to be a bat?, *The Philosophical Review*.

Sanz, R. (2004). Envisioning conscious controllers. In Proceedings of the *International Workshop on Software Systems*, IWSS'2004, Istambul, Turkey.

Sanz, R. y A. Meystel (2002). Modeling, self and consciousness: Further perspectives of ai research. In Proceedings of PerMIS '02, *Performance Metrics for Intelligent Systems Workshop*, Gaithersburg (MD), USA.

Sanz, R., I. López, J. Bermejo, R. Chinchilla y R.P. Conde (2005). Self-X: The control within. In Proceedings of the 16th *IFAC World Congress*, Prague, Czech Republic.

Taylor, J. (2003). Paying Attention to consciousness. *Progress in Neurobiology* 71(4):305-35.

Ricardo Sanz es el coordinador del Grupo de Investigación *Autonomous Systems Laboratory* de la Universidad Politécnica de Madrid.

Se puede contactar con él en:

Ricardo.Sanz@upm.es
Ricardo.Sanz@aslab.org
Ricardo.Sanz@ieee.org

http://www.aslab.org/~sanz

Este artículo se ha elaborado a partir del material empleado en las *III Jornadas de Teoría y Psicología. Aspectos de la consciencia*. Menorca, 9 – 11 de junio de 2005.

El autor desea agradecer a Toni Gomila su invitación a participar en estas jornadas que resultaron ser, aparte de agradables, muy provocadoras.