ARQUITECTURAS EMOCIONALES EN INTELIGENCIA ARTIFICIAL: UNA PROPUESTA UNIFICADORA

Las emociones pueden ser entendidas de diferentes modos dependiendo de las disciplinas desde las que nos planteemos estudiarlas (neurofisiología, filosofía, etc.). En Inteligencia Artificial el aspecto que interesa del estudio de las emociones es el de buscar qué tipos de requisitos estructurales son satisfechos por los estados emocionales y qué mecanismos funcionales subyacen a los procesos emotivos con el fin de reproducirlos en arquitecturas artificiales. En este artículo se hace una revisión de las arquitecturas emocionales más importantes a lo largo de la historia de la Inteligencia Artificial lo que permite, por otro lado, poner en evidencia el problema de las estrategias tradicionales en la implementación de modelos emocionales: En todas subyace el carácter dualista emoción-razonamiento que se ha asignado tradicionalmente a la cognición humana. Los trabajos más modernos en neurofisiología de las emociones han llevado a los investigadores en Inteligencia Artificial a mostrar un aumento de interés por el diseño de sistemas emocionales y a plantearse nuevas arquitecturas, sin proponerse si los sistemas artificiales pueden tener realmente emociones como las de los humanos, y concentrándose en cómo pueden ser reproducidas las emociones exclusivamente desde una perspectiva funcional. Finalizamos el artículo sentando las bases de una arquitectura integrada dónde las emociones serían interpretadas cómo procesos de ayuda a la decisión y no exclusivamente como mecanismos de alarma.

Palabras clave: Emociones, arquitecturas inteligentes y emocionales, funcionalismo neurofisiológico.

UNIFYING EMOTIONS AND REASONS:A NEUROSCIENTIFIC-INSPIRED ARCHITECTURE

Emotions can be explained in different ways depending on the issues we consider (psychology, neuroscience, philosophy, etc). In particular, from an artificial intelligence viewpoint, studying emotions is searching which structural conditions corresponds to emotional states and which functional processes underlie emotional behavior in living beings, especially in humans, with the aim of reproducing them in artificial architectures. This paper sets out a review of the main architectures in Artificial Intelligence and explains where are the errors and constraints in classical models: basically, a dualistic approach to understand cognition that has considered emotions and reasons as antagonists. In the second part of the paper, is introduced the recent interest that researchers show on works about emotions that neurophysiology are involved in. In the next section it is proposed a complex architecture that allows implementing some neurophysiological features with results that generate important expectations. Finally, the paper concludes with a discussion about emotions and Artificial Intelligence, beyond issues of engineering, trying to understand what it means for humans to have emotional behaviours from an adaptive approach.

Keywords: Emotions, intelligence and emotional architectures, neurophysiology.

ARCHITECTURES ÉMOTIONNELLES EN INTELLIGENCE ARTIFICIELLE: UNE PROPOSITION UNIFICATRICE

Les émotions peuvent s´entendre de différentes façons selon les disciplines d´étude (neurophysiologie, philosophie, etc.). En intelligence Artificielle, l´angle qui intéresse l´étude des émotions est la recherche des différents types de requis structurels satisfaits par les états émotionnels et les mécanismes fonctionnels sous-jacents aux processus émotifs dans l´objectif de les reproduire en architectures artificielles. Cet article présente les plus importantes architectures émotionnelles définies tout au long de l´histoire de l´Intelligence Artificielle, ce qui permet de mettre en évidence le problème des stratégies traditionnelles dans la mise en application des modèles émotionnels. Le caractère dualiste émotion-raisonnement, assigné traditionnellement à la connaissance humaine, est sous-jacent dans toutes les stratégies. Les travaux les plus récents en neurophysiologie des émotions ont conduit les chercheurs en Intelligence Artificielle à démontrer un intérêt croissant pour la définition des systèmes émotionnels et à se questionner sur la définition de nouvelles architectures, en écartant le postulat que les systèmes artificiels peuvent ressentir réellement des émotions exclusivement depuis une perspective fonctionnelle. Enfin, nous établirons les bases d´une architecture intégrée où les émotions seraient interprétées comme des processus d´aide à la décision et non exclusivement comme mécanisme d´alerte.

Mots clés: Emotions, architectures intelligentes et émotionnelles, fonctionnalisme neurophysiologique.

ARQUITECTURAS EMOCIONALES EN INTELIGENCIA ARTIFICIAL: UNA PROPUESTA UNIFICADORA

Manuel González Bedia[1] & Joaquín García Carrasco[2]
[1]mgbedia@inf.uc3m.es
[2]carrasco@usal.es
[1]Departamento de Informática
Universidad Carlos III de Madrid

[2]Departamento de Teoría e Historia de la Educación
Universidad de Salamanca

1.- INTRODUCCIÓN

¿Por qué nos interesan en la Inteligencia Artificial los sistemas emocionales? Fundamentalmente por tres razones:

1. Existen problemas en Inteligencia Artificial para los que la aplicación de métodos clásicos no funcionan: Necesitamos herramientas de diseño más efectivas y existen teorías emocionales que pueden inspirar su desarrollo. [Damasio, 1996.]

2. Existen algunas teorías sobre las emociones en Neurología que necesitan aún soporte empírico. La implementación en modelos artificiales junto a experimentos de simulación pueden ser útiles para contrastar este tipo de teorías sobre comportamiento emocional en animales y humanos.

3. El objetivo fundamental de la Inteligencia Artificial es el desarrollo de sistemas artificiales que actúen de manera similar a los humanos. Si las emociones tienen funciones biológicas importantes, como así parecen asegurarlo numerosos estudios [LeDoux, 2000.], el diseño de sistemas emocionales artificiales podrían incorporar nuevas funcionalidades que permitirían mejores competencias y mayor autonomía.

2.- UN REPASO DE LAS EMOCIONES DESDE EL ENFOQUE DE LA INTELIGENCIA ARTIFICIAL

En este artículo nos proponemos presentar las distintas concepciones que se han ido planteando sobre las emociones a lo largo de la historia, y también las estrategias emocionales más representativas que, en el contexto de la Inteligencia Artificial se han desarrollado desde la década de los setenta hasta la actualidad. A continuación, relacionaremos las hipótesis implícitas en el diseño de tales sistemas con los principios de diferentes teorías emocionales en áreas como la Psicología o Neurociencia. Por último, presentaremos las razones por las que tales estrategias no han obtenido el éxito esperado y propondremos una posible vía de solución.

2.1. Teorías emocionales

Existen fundamentalmente tres clases de teorías para explicar el comportamiento de las emociones:

Teorías No cognitivas (o anti-cognitivas)

Tradicionalmente se han considerado las emociones como estados mentales, conscientes y por tanto identificables mediante el lenguaje. La mente desde esta perspectiva sería la combinación, por un lado, de razones y, por otro, de emociones, ámbitos separados y con estatus diferentes. Para este tipo de explicaciones, las emociones serían esencialmente un tipo de respuestas de tipo reactivo, codificadas en nuestros genes, que funcionarían como alarmas ante la percepción de situaciones que peligrosas para nuestra supervivencia. Ejemplos de este enfoque son los trabajos de [Izard, 1977; Zajonc, 1980]. Este enfoque no es satisfactorio por diversas razones y no resuelve algunas preguntas sobre las emociones y la inteligencia.

Teorías Cognitivas

El primero en considerar que las emociones no constituyen un módulo al margen de la cognición fue William James [James, 1884]. Por primera vez se plantea una teoría que interpreta las emociones como la toma de conciencia de las reacciones viscerales (y no las reacciones mismas). Este carácter cognitivo de las emociones permite entender emociones complejas que no podían explicarse con el modelo anti-cognitivo --como emociones nostálgicas acerca de algo que ya ocurrió-- [Lazarus, 1984] pero no resuelve algunos aspectos que sí explicaban las teorías anteriores.

Teorías interactivas

Con este tipo de teorías se pretende encontrar un modelo que integre los resultados de ambas teorías previas. En este caso, las emociones son una entidad producto de dos aspectos hasta el momento separados: activación visceral (anticognitivista) y valoración

cognitiva de la alteración (cognitivista) [Schachhter & Singer, 1962, Arnold, 1960]. El modelo interactivo más completo (y satisfactorio) es el modelo de Frijda [Frijda, 1986]. Permite integrar y explicar emociones primarias (reactivas), secundarias (evaluadas cognitivamente) y meta-emociones de carácter contrafactual (no existen ni han existido los motivos que las provocan).

2.2. Arquitecturas emocionales

Vamos a mostrar algunas arquitecturas emocionales que podemos situar como ejemplos represntativos de las teorías antes presentadas.

Arquitectura No cognitiva [Tomkins, 1984]

El comportamiento emocional en este tipo de sistemas no está relacionado con ningún tipo de sistema cognitivo-deliberativo. La arquitectura establece relaciones entre estímulos y respuestas pre-establecidas mediante un principio de excitación basado en la "densidad de disparos" recibidos (intensidad del estímulo por tiempo). En esta arquitectura, las emociones que pueden generarse cumplen las siguientes características:

1) Las emociones como un tipo de reacciones automáticas
2) La emoción que se manifiesta es la que responde a mayor estimulación de la siguiente manera: (i) si la estimulación aumenta se representan estados como miedo o interés, (ii) si la estimulación decrece se tienen estados como alegría, y (iii) si la estimulación se nivela tenemos estados como la angustia o el enojo.

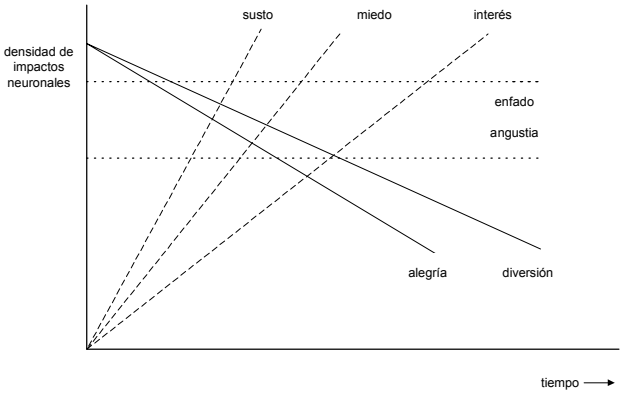

Fig. 1. Modelo de Tomkins

Arquitectura cognitiva [Simon, 1982]

En este modelo, las emociones no se entienden como reacciones automáticas sino más bien como elementos evaluadores del entorno con la capacidad para interrumpir el proceso deliberativo en curso y forzar al sistema para que centre su atención y sus recursos en una nueva situación. Las únicas emociones que pueden representarse son solamente

aquellas que se pueden entender como mecanismos de atención urgente. El sistema básicamente está formado por:

1. Un módulo central de gestión de objetivos de tipo serial: los nuevos estímulos se ponen a la cola y esperan turno para ser evaluados organizándose en forma de jerarquía de objetivos.

2. Un módulo emocional representado por un sistema de vigilancia con capacidad para interrumpir el procesamiento del módulo central si observa contingencias que requieren atención urgente.

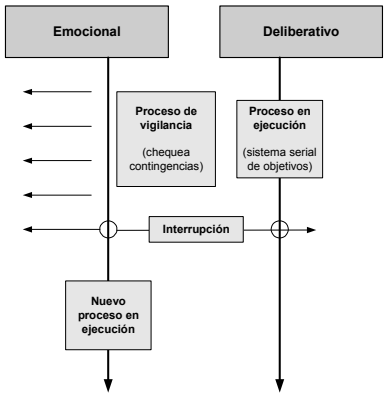

Fig.2. **Modelo de Simon**

Arquitectura interactiva [Sloman, 1987]

Este modelo es más complejo aunque se deriva del anterior. Presenta emociones que aparecen en el modelo de Simon pero incorpora:
1. Un mecanismo de filtro: las actividades en curso usan recursos, tanto cognitivos como físicos, que son limitados. Por tanto, es necesario un filtro para que estén protegidos y abiertos a la interrupción.

2. De umbral variable: La variabilidad del umbral permite al nivel de protección ser dependiente del contexto.

El modelo de Sloman, por tanto, además de tener la capacidad de interrumpir los objetivos actuales del sistema puede representar la disposición a la interrupción sin llegar a hacerlo.

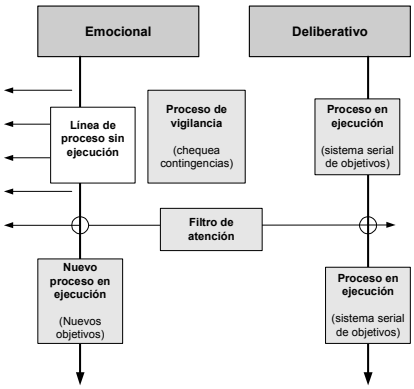

Fig.3. **Modelo de Sloman**

En este modelo, mucho más rico, podemos reproducir comportamiento emocional con las características de las emociones de Simon, es decir, *emociones como mecanismos de atención urgente que interrumpen el procesamiento actual,* pero además con las siguientes propiedades:

1. Emociones como mecanismos disposicionales
2. Emociones como mecanismo graduales
3. Emociones y disposiciones que pueden coexistir

Arquitectura basada en el modelo de Frijda [Frijda, 1986]

Siguiendo el modelo de Frijda se desarrolla la propuesta más completa hasta el momento. Se estructura a partir del modelo de Sloman añadiendo un canal para el flujo de información continuo y bi-direccional, entre proceso de atención, vigilancia y filtro.

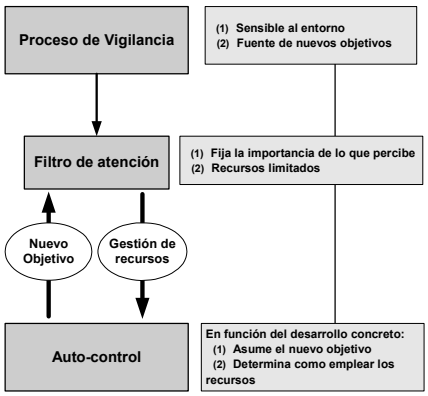

Fig.4. Modelo de Frijda

Bajo estas condiciones podemos tener flujos de información, a los que podemos asociar significado emocional, dadas las propiedades que reflejan:

- Pueden interrumpir el proceso de atención si detectan situaciones de emergencia (característica del modelo de Simon)
- Pueden representarse como disposiciones a la acción, coexistir varias y presentan un carácter gradual (característica del modelo Sloman)
- Su manifestación depende de una articulación entre percepción, contexto y experiencia (característica del modelo de Frijda)

Una de las arquitecturas emocionales de mayor éxito, TABASCO (Tractable Appraisal-Based Architecture for Situated Cognizers) [Petta, 2002], sigue en su diseño un "modelo emocional interactivo" basado en la teoría de Frijda.

3.- PROBLEMAS Y LIMITACIONES

En el conjunto de todas las arquitecturas presentadas encontramos una serie de elementos que se repiten: las emociones son estados, es decir, son emociones prefijadas y formalizadas como conceptos, y las emociones actúan esencialmente como perturbaciones, esto es, interrumpen el proceso de deliberación en curso. Todos los ejemplos son casos de arquitecturas emocionales basadas en conceptos. Esta condición define un comportamiento prefijado para las distintas emociones y una serie de características típicas:

- *Contenido representacional:* Emociones formalizadas como conceptos
- *Contenido no-representacional*: Emociones modeladas como señales de alarma
- *Capacidad reactiva de tiempo real*: Mecanismo de vigilancia e interrupción
- *Disposiciones sin interrupción*: Emociones como motivaciones a partir de un mecanismo de filtro variable

Sin embargo, nos gustaría que presentasen otras características que también son rasgos esenciales de las emociones, como que:

- Generen estructuras del *sentimiento* ("mood") de las emociones, es decir, un mecanismo de valoración emocional.
- Generen procesos de adaptación emocional (*aprendizaje emocional*).
- Favorezcan procesos de decisión. Según teorías neurofisiológicas actuales la falta de emociones en sujetos inteligentes puede constituir una fuente importante de conducta irracional. Se ha comprobado [Damasio, 1996] que pacientes con daños en el lóbulo central, volviéndose emocionalmente planos, pierden su capacidad para tomar decisiones racionales.

4.- POSIBLES SOLUCIONES

Para reproducir artificialmente comportamiento emocional más complejo y más ajustado al que observamos en la realidad, sugerimos adentrarnos en *las teorías Neurofisiológicas* acerca de las emociones. En estas teorías se defiende un enfoque completamente diferente al de los modelos de emociones basadas en conceptos. Veamos sus postulados más importantes:

4.1. Dualismo vs. Funcionalismo

Mientras que los modelos emocionales basados en conceptos presentan habitualmente una visión dualista entre la cognición y las emociones, la Neurofisiología no establece diferencias entre ambos: los dos son casos de procesamiento informacional. Las emociones son simplemente procesos informacionales que representan funciones mentales. Esta posición de cara al estudio en la Inteligencia Artificial implica que:

- No nos interesarán explicaciones causales sobre los mecanismos que las generan (¿*Cómo ocurren*?)
- Tampoco nos preocuparemos por teorías evolutivas que expliquen su relación con las funciones mentales (¿*Por qué ocurren así*?)
- Ni cuestiones acerca del carácter innato, aprendido, o las modificaciones a lo largo del desarrollo del individuo (¿*Cuándo*?).

Nuestro interés se centra en entender *Para qué sirven*, es decir, cómo las emociones hacen a los organismos sobrevivir o adaptarse. De otra forma, nos interesará responder a la cuestión siguiente: *Cuál es su función.*

4.2. Perturbaciones vs. Mecanismos de apoyo a la razón

Los trabajos de [Damasio, 1996; De Sousa, 1987], a partir de sus investigaciones, defienden que las emociones *nos ayudan a tomar decisiones,* pues, por un lado, deshacen el empate en los casos de *indiferencia, y por otro lado,* Constituyen un criterio de selección en casos de *indeterminación*, al hacer posible que nos centremos en los rasgos más destacados de la situación presente. Por tanto, pasamos de entender las emociones como interrupciones a emociones como sistemas de toma de decisiones que, en ciertas ocasiones, mejoran nuestra conducta.

5.- MODELOS INTEGRADOS DE "RAZONES-EMOCIONES"

La propuesta que hacemos, tras el estudio realizado, es buscar arquitecturas emocionales con las siguientes características:

- Neutras (sin emociones pre-definidas)
- Emociones como procesos funcionales
- Emociones como mecanismos de supervivencia (mejora del comportamiento del sistema)
- Emociones como mecanismos en procesos de toma de decisión (planificación) complementarios a la deliberación

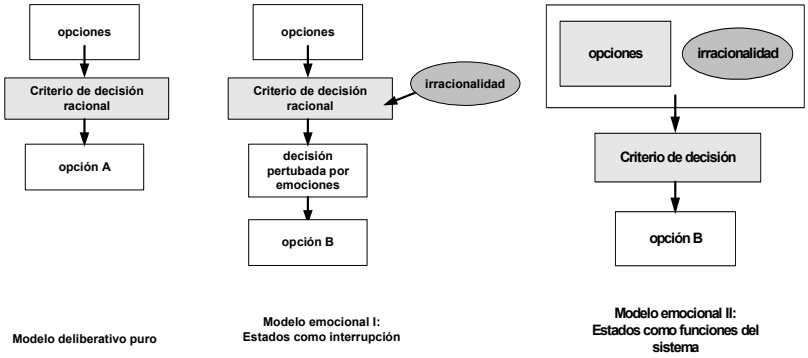

Fig.5. Modelo unificado emociones-razones

En la figura 5 mostramos un modelo de "alto nivel" donde se refleja el cambio necesario en la mentalidad ingenieril para abordar el diseño de artefactos inteligentes en sintonía con la neurofisiología actual. La irracionalidad, considerada como la etiqueta de los procesos regulados por instintos y corazonadas en lugar de basados en razones, debe ser un mecanismo interno que participe en los procesos de toma de decisión.

6.- CONCLUSIONES

Un problema que ha hecho del análisis de las emociones algo complicado ha sido la iterpretación naïve que los investigadores han asumido a partir de la vision popular que se tiene de ellas. Como consecuencia, ha existido mucha vaguedad en torno a su análisis. En este artículo el propósito ha sido mostrar las emociones de manera distinta a cómo hasta ahora se han venido considerando para dar nuevas directrices en el desarrollo de arquitecturas de inteligencia artificial. La principal característica en este nuevo enfoque es que consideramos la relevancia de las emociones en términos del rol de la autonomía que añaden a los sistemas inteligentes: los sistemas pueden ser más adaptativos si utilizan estrategias de carácter emocional en sus mecanismos de toma de decisiones.

El objetivo de incluir emociones en sistemas artificiales no es, por tanto, simular comportamiento emotivo ni emular sentimientos humanos en los mecanismos de procesamiento de información. Si eliminamos los prejuicios populares sobre un posible carácter antagónico entre emociones y razones, tendremos entonces la posibilidad de entender la contribución de las emociones en los artefactos inteligentes como un mecanismo que controla la prioridad entre el conjunto elevado de aspectos del entorno sobre los que concentramos nuestra atención para asegurarnos, en ciertas situaciones, una respuesta inteligente en un tiempo razonablemente breve.

7.- BIBLIOGRAFÍA

ARNOLD, M. (1960). *Emotions and Personality.* Nueva York. Columbia University Press.

DAMASIO, A. (1996). *Descartes' error*. Drakontos, Crítica.

DE SOUSA, R. (1987). *The Rationality of Emotion*. Cambridge, Mass., MIT Press.

FRIJDA, N. (1986). *The emotions. Sudies in Emotion and Social Interactions*. Cambridge University Press, Cambridge, UK.

GAZZANIGA, M.S. & LEDOUX, J. (1978). *The Integrated Mind*, Nueva York, Plenum.

IZARD, C. E. (1977). *Human Emotions*. Nueva York. Plenum Press.

JAMES, W. (1884). *What is emotion?,* Mind, 1884, pp. 59

LAZARUS, R.S. (1991). Progress on a Cognitive-Motivational-Relational theory of emotion. *American Psychologist*, **46**, pp. 819-834.

LEDOUX, J. (2000). *The Emotional Brain*, Nueva York, 2000

PETTA, P. (2002). The role of emotions in tractable architectures for situazed cognizers. In TRAPPL, R., PETTA, P. & PAYR, S. (Eds.). *Emotions in Humans and Artifacts*. Cambridge, M.A., MIT Press.

SCHACHHTER, S. & SINGER, J. (1962). Cognitive, social and physiological determinants of emotional state. *Psychological Review*, **59**, pp. 379-399.

SIMON, H.A. (1982). A← ect and Cognition: Comments. In CLARK, M.S. & FISKE, S.T. (Eds.). *The Seventeenth Annual Carnegie Symposium on Cognition: Affect and Cognition*. London: Lawrence Erlbaum Associates, pp. 333-342.

SLOMAN, A. (1987). Motives mechanisms and emotions. *Cognition and Emotion*, **1** (3), pp. 217-234. Reprinted in Boden, M.A. (Ed.). The philosophy of Artificial Intelligence, OUP.

TOMKINS, S. (1984). Affect Theory. In SCHERER, K. & EKMAN, P. (Eds.). Approaches to Emotion. Hillsdale, NJ: Lawrence Erlbaum Associates.

ZAJONC, R. B. (1980). Feeling and thinking: preferences need no inferences. *American Psychologist*, **35**, pp. 151-175.

Para citar esta artículo puede utilizar la siguiente referencia:

BEDIA, Manuel (2006): Arquitecturas emocionales en inteligencia artificial: una propuesta unificadora. En GARCÍA CARRASCO, Joaquín (Coord.) Estudio de los comportamientos emocionales en la red [monográfico en línea]. *Revista Electrónica Teoría de la Educación: Educación y Cultura en la Sociedad de la Información.* Vol. 7, nº 2. Universidad de Salamanca. [Fecha de consulta: dd/mm/aaaa].
<http://www.usal.es/~teoriaeducacion/rev_numero_07_02/n7_02_manuel_bedia.pdf>
ISSN 1138-9737