Núñez, Manuel. *Computadoras y sentimientos: discordia entre investigadores de inteligencia artificial.* **Tecnología en marcha** Vol. 12 no. 1 p3 - 7.

COMPUTADORAS Y SENTIMIENTOS: DISCORDIA ENTRE INVESTIGADORES DE INTELIGENCIA ARTIFICIAL

Manuel Núñez*

En este artículo se discuten argumentos generales relacionados con la posibilidad de tener sistemas capaces de entender sentimientos humanos, siguiendo el punto de vista de dos autores: J. Haugelan y T. Edelson. Después de explicar la teoría, se procede a estudiar un desacuerdo entre ambos autores con respecto a si tal sistema debería tener sentimientos por sí mismo.

INTRODUCCION

Muchos autores de Ciencia Ficción han usado los temas de los sentimientos y los robots como un medio para desarrollar nuevas perspectivas en el futuro incierto de las relaciones hombre–máquina. Por ejemplo, Isaac Asimov, en su serie de novelas de robots, muestra su personaje R. (Robot) Daneel Olivaw como un robot detective, capaz de tener motivaciones y entender sentimientos humanos, aunque él carece de ellos.

El tema es analizado por John Haugeland[2] en su libro **Artificial Intelligence: The Very Idea** (Inteligencia Artificial: La Mera Idea) y por Thomas Edelson en su artículo[1] **Can a system be intelligent if it never gives a damn?** (¿Puede ser un sistema inteligente si nunca maldice?). Ambos autores están de acuerdo en que existen algunas clases de pensamiento que no pueden ser separados de afecciones y sentimientos, sin embargo, Haugeland considera que un sistema capaz de entender sentimientos humanos debe también poseer sentimientos por sí mismo, mientras que Edelson sostiene la idea contraria, es decir, que pueden existir tales sistemas sin poseer necesariamente sentimientos.

En este artículo se introduce la noción de **sistemas comprometidos** (*engaged systems*) con el fin de tener un marco de referencia que le permita al lector formar su propio criterio dentro de este fascinante campo de la Inteligencia Artificial. Luego, estudiamos la disputa entre los investigadores antes citados y, al mismo tiempo, presentamos una interesante analogía de estos sistemas con el concepto de **máquinas virtuales** del área de Sistemas Operativos.

SISTEMAS COMPROMETIDOS

En general a aquellos sistemas con facultades afectivas se les llama **sistemas comprometidos**. Por lo tanto, un sistema comprometido es uno capaz de poseer motivos, de tal manera, que puede explicar las acciones que realiza. Además, vamos a permitir que esta clase de sistemas tenga múltiples motivos

* Máster en Computación. Profesor de la Maestría en Computación. Instituto Tecnológico de Costa Rica.

independientes que pueden variar con el paso del tiempo y que pueden incorporar o perder motivos.

Los temas de los sentimientos y los robots han sido usados como un medio para desarrollar nuevas perspectivas en el futuro incierto de las relaciones hombre-máquina . Por ejemplo, los seres humanos son sistemas comprometidos. Sin embargo, como Edelson[1] observa, hasta el momento ningún sistema basado en computadoras es un sistema comprometido. En otras palabras, cuando decimos que un sistema de preguntas–respuestas **desea** encontrar la respuesta correcta, o que un programa que juega ajedrez **quiere ganar**, lo que obtenemos son metáforas y no frases literales. De hecho, estos sistemas no poseen ningún motivo.

En Why **robots will have emotions?** (¿Por qué los robots tendrán emociones?) por Aaron Sloman y Monica Croucher, se argumenta que cualquier sistema comprometido debe tener emociones y, por otro lado, un sistema con emociones debe ser un sistema comprometido. El argumento que se sigue consiste en que el éxito o fracaso en la obtención de una meta en un motivo produciría emociones. Por ejemplo, si X es un sistema comprometido y Y se sospecha que es responsable por violar uno de los motivos de X, entonces X estaría enojado con Y y X podría desear dañar o herir a Y (un nuevo motivo).

De hecho, vamos a argumentar que un modelo cognoscitivo humano debería incluir un modelo de facultades afectivas humanas. Consideremos la siguiente anécdota que le ocurrió a una de las amigas de Haugeland:

Cuando se hallaba en la universidad, la amiga de Haugeland solía tener una rata blanca como mascota. La mascota era lo suficientemente dócil como para seguir a la muchacha calladamente detrás de sus talones. Pero un día, asustada por un perro, la rata tomó refugio rápidamente muy arriba dentro de los pantalones de ella. Desafortunadamente, la rata se acomodó de tal manera que quedó atrapada sin poder volverse atrás y, mientras tanto, la muchacha no se atrevía a moverse por temor de maltratar a su mascota. Así que, después de un momento de duda, ella resignadamente se bajó sus pantalones, liberó a su tembloroso roedor y se ganó una ronda de aplausos proveniente de los deleitados transeúntes.

La mayoría de las personas encuentra esta anécdota divertida. Cuando se nos pregunta, ¿cómo se sintió la amiga de Haugeland?, o cuando se consideran divertidas tales historias, confrontamos gente hipotética en situaciones hipotéticas. Estamos envueltos en un proceso cognoscitivo, donde usamos nuestras facultades afectivas. Imaginando cómo nos veríamos en una situación como esa, podemos sentir sus reacciones como si nos sucediera a nosotros mismos.

Este proceso de imaginación puede ser pensado como un mecanismo que nos permite producir mensajes neurales falsos y hacer que nuestro cerebro responda exactamente como lo haría en la situación real. Esta clase de reconocimiento no puede ser adquirido por razonamiento y es diferente de la mayoría de los conocimientos de sentido común. Cada persona se usa a sí misma como un modelo de la persona en que está pensando.

MAQUINAS VIRTUALES

Edelson explica una interesante analogía entre el concepto de sistemas comprometidos con el concepto de 'maquina virtual' encontrado en la literatura de sistemas operativos:

Usamos nuestras creencias, esperanzas y temores para entender creencias, esperanzas y temores de otros sistemas comprometidos.

Cada vez que una máquina virtual tiene que ejecutar un programa debe simular una máquina compatible con la máquina para la cual el programa fue diseñado. Sin embargo, la ejecución es llevada a cabo en el *hardware* de la máquina real. Solo las entradas y salidas son simuladas o redireccionadas[6]. De hecho, la máquina simulada puede tener su *hardware* completamente distinto al de la máquina simuladora. (Un ejemplo típico de este tipo de máquinas[4] lo podemos encontrar en el sistema VM/370 de IBM).

De manera similar, un ser humano puede generar entradas (representando la situación imaginada), usar su propio *"hardware"* (las facultades afectivas) y producir una "salida", la cual es redireccionada al cerebro para entender lo que otra persona siente.

Las máquinas virtuales se utilizan para experimentar con nuevas arquitecturas de computadoras o para hacer compatibles (facilitar la comunicación) entre distintas computadoras. Análogamente, los seres humanos se comunican entre sí y se comprenden unos a otros por medio de su capacidad de interpretar los sentimientos de los demás.

De todos estos argumentos inferimos, que si se quieren modelar afecciones y sentimientos, es necesario construir un sistema comprometido, el cual debería tener múltiples motivos complejos y facultades afectivas, tal que realice el truco de la 'persona virtual' para entender gente a través de analogías consigo mismo.

Este enfoque es básicamente sugerido por ambos autores. Haugeland se refiere a él como *Metacognición*, en el sentido que usamos nuestras creencias, esperanzas y temores para entender creencias, esperanzas y temores de otros sistemas comprometidos.

Edelson lo sugiere como un enfoque llamado **de la caja negra**, que consiste en pensar de los seres humanos como funciones enviando preguntas hacia respuestas y predicciones. De esta manera, se crea una teoría de observaciones que, a su vez, crea un sistema sin las facultades afectivas, pero razonando sobre la teoría de la naturaleza humana.

La diferencia de los dos enfoques radica en que la persona virtual estudia la naturaleza humana desde adentro, mientras que el enfoque de la caja negra es un estudio exterior. En todo caso, ambos enfoques parecen ser la misma cosa.

PEDRO VRS JOSE

Hasta el momento, ambos autores coinciden en la idea de un sistema que posee un sistema comprometido (persona virtual) que le permite entender la naturaleza humana, es decir, afecciones, emociones, motivos, etc. La idea no es nueva, de hecho Marvin Minsky[3] en **Why people think computers can't?** (¿Por qué la gente piensa que las computadoras no pueden -pensar-?) nos dice que lo que nosotros llamamos "yo" es en realidad un modelo de nuestra mente útil para entender sentimientos de otras personas o los nuestros mismos.

Pero el punto de fisura entre los argumentos de ambos autores se halla precisamente en si tales sistemas, capaces de entender la naturaleza humana, deben o no ser sistemas comprometidos por sí mismos.

De acuerdo con Haugeland, tales sistemas deben ser también comprometidos, y para justificar su afirmación utiliza el siguiente argumento: Supongamos que tenemos un sistema llamado PEDRO capaz de entender a la persona (o sistema) llamada JUAN.

Entonces PEDRO debe poseer un sistema comprometido que llamaremos JOSE, el cual es un modelo de JUAN. Cada vez que PEDRO quiere entender algún estado de JUAN, PEDRO le da a JOSE las entradas apropiadas para que JOSE simule a JUAN. El resultado de la simulación le es dado a PEDRO y, como consecuencia, PEDRO entiende a JUAN.

Haugeland señala el hecho de que en realidad todo el trabajo lo realiza JOSE y no PEDRO. Más aún, JOSE tiene ahora un "yo" propio como consecuencia del axioma de independencia del medio de Inteligencia Artificial: "simular uno es ser uno". Por lo tanto, ya que JOSE realiza todo el trabajo, PEDRO por sí mismo, conteniendo a JOSE, debe ser un sistema comprometido.

Por el otro lado, Edelson señala que Haugeland se halla en un error y argumenta que el hecho de que PEDRO contiene a JOSE y de que JOSE sea un sistema comprometido no necesariamente implica que PEDRO es también un sistema comprometido. Por ejemplo, supongamos que en la simulación de PEDRO a través de JOSE, JOSE descubre que JUAN está enojado, entonces JOSE le comunica a PEDRO que JUAN está enojado y PEDRO toma alguna acción, no necesariamente ponerse también enojado. Inclusive, ni siquiera es necesario que PEDRO manifieste alguna emoción o afección.

Aún más, PEDRO puede tener metas programadas de tal manera que simplemente podría desestimar las salidas de JOSE que no estén relacionadas con esas metas. En conclusión, PEDRO no es un sistema comprometido, aunque contiene un modelo de uno.

CONCLUSION

La decisión sobre quién tiene la razón acerca de este pequeño desacuerdo queda a criterio del lector. En todo caso, la diferencia pareciera más semántica que técnica y, por ende, sin mucha importancia.

Sin embargo, la verdadera importancia de molestarse en pensar argumento es el más razonable radica plantearse la posibilidad de creación tales sistemas comprometidos, creac que ambos autores han dado como u hecho programable.

Tal vez, el mejor acercamiento cómo programar uno de tales sistema halla en el enfoque de la caja negra Edelson, que es desarrollado con ma detalle en el artículo de Sloman y Croucher mencionado anteriormente ese artículo se indica que lo que se necesita para crear uno de estos siste comprometidos es definir una serie de restricciones y especificaciones nece sarias para modelar afecciones huma Así que la verdadera dificultad desde punto de vista práctico se encuentra diseñar tales restricciones, aunque la manera de hallarlas sería muy seme a como estudiamos fenómenos físicos como funciones que producen ciertas salidas con base en ciertas entradas. Entonces el problema de creación de sistema comprometido se convierte problema de aproximación de funcion para el cual existen gran cantidad de métodos clásicos de solución. Tambi podrían utilizarse métodos de redes neurales para encontrar tal función.

Las dificultades que surgen se en la escogencia de una representac de los estados (sentimientos, motivos etc.) de forma que se facilite el anál distintas situaciones y la amplia gam reacciones que pueden darse bajo variaciones de una misma situación. pesar de todo, se recomienda el artíc de Sloman y Croucher, donde se introducen (tentativamente) estructur

datos capaces de llevar a cabo esta tarea.

Indudablemente, como sucede con la mayoría de los estudios propios de Inteligencia Artificial, nos hallamos en pañales cuando llega el momento de especificar sistemas de este tipo en concreto, e incluso, se corre el riego de caer en problemas usuales de simulaciones demasiado realistas, tales como crear un sistema tan complejo que es imposible utilizarlo como herramienta para entender el fenómeno original. Este problema, así como las implicaciones de la existencia de uno de tales sistemas, cae dentro de los planteamientos filosóficos y morales de la Inteligencia Artificial y no nos corresponde analizarlo en este artículo.

REFERENCIAS BIBLIOGRAFICAS

1) Edelson, Thomas (1986). *Can a System be Intelligent if it Never gives a Damn?* **Proceedings of the American Association for Artificial Intelligence**. p. 298-302.

2) Haugeland, John (1985). **Artificial Intelligence: The Very Idea.** Massachussetts: MIT Press.

3) Minsky, Marvin (1982). *Why People think computers Can't?* **AI Magazine**. p. 3-15.

4) Seawright, L. H. y MacKinnon R. A. (1979). *VM/370 – A Study of Multiplicity and usefulness* **IBM System Journal**, Vol. 18, no. 1. p. 4-17.

5) Sloman, Aaron y Croucher, Monica (1981). *Why Robots will have emotions?* **Proceedings IJCAI**. p. 197-202.

6) Goldberg, Robert (1979). *Survey of Virtual Machine Research.* **Computer Magazine**. p. 34-45.