Jornadas de Hum.H.A.
Bahía Blanca - República Argentina
11 al 13 de agosto de 2005
Dpto. de Humanidades
Área Historia del Arte

Representación del conocimiento en inteligencia artificial

Erica Frontini [1]
Gustavo Bodanza[2]
(Dpto. Humanidades – UNS)

Introducción

La Inteligencia Artificial (AI) intenta crear sistemas de computación que desarrollen actividades que puede hacer la mente humana. Es, por consiguiente, una disciplina que involucra tanto a las ciencias de la computación como a la psicología, la neurobiología, la lingüística y la filosofía en el estudio del comportamiento inteligente. En sus desarrollos interactúan dos objetivos que se retroalimentan: comprender el funcionamiento de la mente humana (aunque también puede extenderse el objetivo al comportamiento animal) y producir máquinas útiles para las diversas tareas humanas.

Los orígenes de esta disciplina se remontan a un artículo del año 1943 de McCulloch y Pitts titulado "A logical calculus of the ideas immanent in nervous activity".[3] En dicho artículo se integran tres poderosos desarrollos de comienzos del siglo XX: la lógica proposicional, la teoría neuronal de Charles Sherrington y las máquinas de Turing. Este *paper* hizo posible el surgimiento de la inteligencia artificial de tres maneras: por un lado, influenció a Von Neumann en la utilización de las lógicas y aritméticas binarias para el diseño de la computadora digital. Por otro, les dio a los psicólogos y técnicos la confianza para modelar razonamiento proposicional en computadoras basadas en la lógica. Además inspiró el comienzo del estudio de las propiedades computacionales de diversos tipos de redes neuronales.

No obstante ello, la expresión "inteligencia artificial" fue utilizada por primera vez por John McCarthy, en 1956 en una propuesta escrita para un encuentro en Darmouth College, New Hampshire. Es este mismo autor quien considera que la Inteligencia Artificial comenzó con el artículo de Turing *Computing Machinery and Intelligence* de

[1] ericafrontini@yahoo.com
[2] ccbonz@criba.edu.ar
[3] Aparece compilado en Boden, M.A. (ed.), *The Philosophy of Artificial Intelligence*, Oxford, Oxford University Press, 1990.

1950 y con la discusión de Shannon acerca de cómo hay que programar una máquina para jugar ajedrez.[4]

Dentro de esta disciplina dominan dos enfoques: el Tradicional (que engloba la Inteligencia Artificial Clásica y el Conexionismo) y la Nouvelle AI, también conocida como Robótica Situacional o Antirrepresentacionalismo.[5] Dicha división se plantea en base al rol que la representación juega en cada uno de los dos enfoques: mientras que la Inteligencia Artificial Tradicional en sus dos versiones (Conexionismo y Clásica) considera la representación interna como esencial a la inteligencia, postulando estructuras identificables de datos en la "mente" que son distinguibles de los procesos del sistema y que están en lugar de las cosas que hay en el mundo, la Robótica Situacional pretende ser biológicamente más realista enfatizando la noción de agentes autónomos específicamente adaptados a su entorno, en lugar de la de computadoras de propósito general controlables por diferentes programas, al mismo tiempo que rechaza la utilización de representaciones internas del mundo exterior (objetivo), aunque algunos sistemas usan representaciones temporarias de su propio entorno. A causa de que el entorno es asumido como pleno de interferencias, dinámico y problemático la detallada modelación del mundo propuesta por la Inteligencia Artificial Clásica es abandonada. La Robótica Situacional evita la descomposición funcional abstracta de las tareas empleada por la Inteligencia Artificial Tradicional analizando inteligencia en términos de comportamiento.

De todos modos, si bien esta propuesta supera muchas de las dificultades de los enfoques representacionistas (como el conocido problema del marco o *frame problem*, que desarrollaremos más adelante) no resulta completamente satisfactoria ya que es muy difícil sostener que la inteligencia humana (y animal) pueda ser modelada sin representaciones internas.

En lo que resta del trabajo nos centraremos, por lo tanto, en algunos de los desarrollos de la Inteligencia Artificial Tradicional cuyo eje es la representación del conocimiento y, en particular, el razonamiento de sentido común.

Inteligencia Artificial Tradicional

Como mencionamos anteriormente este enfoque tiene dos vertientes: la Inteligencia Artificial Clásica y el Conexionismo. Los investigadores de ambas trabajan bajo dos supuestos fundamentales: el primero es que un proceso inteligente puede ser descripto por algoritmos, que son secuencias de operaciones realizables en un tiempo finito en las

[4] Cfr. McCarthy, J. y Hayes, P., "Some philosophical problems from de standpoint of AI", en *Machine Intelligence* **4**, 1969, pp. 463-502.

que cada paso (denominado también primitiva) es tan claro y simple que puede ser hecho automáticamente, sin inteligencia. El segundo es que todos los algoritmos pueden ser implementados en alguna computadora de propósito general.

La Inteligencia Artificial Clásica comenzó asumiendo que la lógica simbólica es un modelo normativo del razonamiento de los humanos y los autómatas. Este supuesto cuaja en algunas formas de razonamiento, como la demostración de teoremas, pero los humanos razonamos mayormente de modo aproximado y cualitativo y podemos utilizar el conocimiento acerca del mundo para resolver problemas lógicos y decidir si el razonamiento estrictamente lógico, opuesto a la heurística falible basada en la experiencia, es apropiado en un determinado caso dado. Es por eso que la investigación en Inteligencia Artificial se ha centrado cada vez más en el razonamiento de sentido común, superando los problemas que tenían los primeros enfoques basados estrictamente en la lógica clásica, aunque sin eliminarlos completamente.

Paralelamente (aunque de modo tímido hasta mediados de los ochenta) se desarrolló la otra vertiente, denominada Conexionismo. Los modelos conexionistas son sistemas de procesamiento paralelo que involucran cómputos mutuamente basados en interacciones locales entre unidades conectadas. Cada cómputo individual es mucho más simple que una instrucción típica en un programa de Inteligencia Artificial Clásica. Aun así las unidades conexionistas y los cómputos varían significativamente.

La distinción entre ambas vertientes se puede definir en términos de la presencia o ausencia de constituyentes lógicos causalmente eficaces y semánticamente evaluables. Tales constituyentes simbólicos tipifican la Inteligencia Artificial Clásica. Las unidades "subsimbólicas" (conexionistas) no poseen un interpretación fija, sino que la misma es dependiente del contexto puesto que su efecto en el procesamiento varía de acuerdo a la activación simultánea de otras unidades. Así, en la forma más exitosa de Conexionismo, el procesamiento de distribución paralela (PDP) un concepto no es guardado por una sola unidad, sino que es representado por un patrón global de activación dispersado a través de la red entera, las diferentes unidades hacen alguna contribución. Es imposible asignar una interpretación específica, independiente del contexto a una unidad dada. Más aun, ninguna unidad individual es necesaria o suficiente para la red entera para representar algún concepto particular. Estos sistemas tienen el problema de que no pueden modelar estructuras jerárquicas o procesamientos secuenciales, habilidades que sí poseen los programas de la Inteligencia Artificial Clásica. Como algunos tipos de razonamiento

[5] Garnham, A., *Artificial Intelligence. An Introduction,* London, Routledge and Keagan Paul, 1988.

humano (por ejemplo ciertos aspectos del lenguaje y la resolución de problemas) requieren dichas habilidades, existe un interés creciente en el desarrollo de modelos híbridos, que tratan de tomar lo mejor de cada vertiente.

Dado que la vertiente clásica es la que ha dado lugar a la mayor cantidad de desarrollos en lo referente a representación de conocimiento, nos centraremos en esta de ahora en más para efectuar nuestro análisis.

La noción de representación en la Inteligencia Artificial Clásica

La Inteligencia Artificial Clásica se desarrolla en base a ciertos presupuestos filosóficos y científicos, entre los que se destaca el de la existencia de un mundo objetivo. Así McCarthy[6] nos dice que los hechos de la matemática y la física son independientes de que exista gente para conocerlos y los robots deben conocer los mismos hechos al mismo tiempo que necesitan creer que el mundo existe independientemente de ellos. Y un robot representa lo que cree acerca del mundo a través de expresiones lógicas. Algunas de esas creencias las construimos, otras provienen de sus observaciones y otras arriban por inducción a través de la experiencia. En cada caso se trata de diseñarlo de modo tal que lo que crea acerca del mundo sea tan acertado como sea posible, aunque no todo lo detallado que sea posible. Existen tres formas de adecuación para una representación.[7]

1) Adecuación metafísica: una representación es metafísicamente adecuada si se da el caso en que el mundo podría tener esa forma sin contradecir los hechos que pertenecen al aspecto de la realidad que nos interesa. Las representaciones metafísicas adecuadas son útiles para construir teorías generales en ciencia.

2) Adecuación epistemológica: una representación es epistemológicamente adecuada si puede ser usada por una persona o máquina para expresar los hechos que pertenecen a determinado aspecto del universo. El lenguaje ordinario es un ejemplo paradigmático de un instrumento adecuado para expresar los hechos que la gente comunica entre sí. Sin embargo, no lo es para expresar lo que la gente sabe acerca de cómo reconocer una cara.

3) Adecuación heurística: una representación es heurísticamente adecuada si puedo plantear nuevos problemas en términos de su lenguaje.

Una buena parte de los trabajos en Inteligencia Artificial tratan de generar modelos formalizables del razonamiento de sentido común que cumplan con estos requisitos de adecuación. Puesto que un agente debe razonar acerca de algo, cualquier consideración

[6] McCarthy, J., "Philosophical and scientific presupositions of AI", en Levesque H.J y F. Pirri (eds.), *Logical Foundations for Cognitive Agents: Contributions in Honor of Ray Reiter*, Springer-Verlag, 1999.

sobre la naturaleza del razonamiento requiere una preocupación concomitante con el modo en el que el agente representa sus conocimientos o creencias. En este sentido podemos afirmar que el trabajo de muchos investigadores de AI es la representación de acuerdo a estos tres cánones.

El razonamiento de sentido común y sus aproximaciones formales

El modelado del sentido común de un agente abarca básicamente la representación de:

1) El contenido de su conocimiento

1.1) Sus creencias acerca de los hechos

1.2) Sus creencias acerca de cómo inferir unas creencias factuales de otras

2) La representación del mecanismo que determina las inferencias justificables en el sistema.[8]

Dicha representación debe correr a cargo de estructuras de datos interpretables como fórmulas lógicas de algún tipo, como expresiones en un lenguaje con una teoría verdadera. Debe ser posible que tomemos una de esas expresiones o fórmulas y digamos cómo debería ser el mundo para que sea verdadera.

Entre los diferentes patrones de razonamiento humano que incluyen el razonamiento probabilístico, el difuso (*fuzzy*), el inductivo y el deductivo, ha cobrado cada vez más fuerza, a la hora de modelar el razonamiento de sentido común, el razonamiento no–monótono. La no–monotonía es una propiedad que se cumple cuando ciertas consecuencias de un conjunto de creencias dejan de derivarse cuando se agregan nuevas creencias. El ejemplo (recalcitrantemente) canónico de la literatura especializada es el siguiente:

Normalmente, las aves vuelan.

Tweety es un ave.

Luego, Tweety vuela.

Puesto que no partimos de la premisa “todas las aves vuelan” se puede decir que hemos “saltado” a la conclusión o que hemos supuesto “por defecto” (*default*) que las aves vuelan. Cuando los humanos razonamos hacemos todo el tiempo este tipo de “saltos” que en un sistema deductivo fuerte como el de la lógica clásica no sería admitido. Por otro lado, y aquí viene específicamente la propiedad de no monotonía, si agregásemos a las

---

[7] McCarthy, J. y Hayes, P., *Op. Cit.,* 1969.

[8] Cfr. Bodanza, G.,*Un Sistema de Argumentación Rebatible Suposicional*. Tesis de doctorado. Bahía Blanca, UNS, 1999.

anteriores una nueva premisa que nos dice que Tweety es un pingüino, nosotros revisaríamos la conclusión y terminaríamos afirmando que Tweety no vuela. Hecho que tampoco es reflejado en la lógica clásica que se caracteriza por ser monótona, ya que al agregar una premisa más a las que ya poseía antes mi conclusión, si fue válidamente inferida, no debería cambiar.

El interés por desarrollar sistemas no–monótonos fue alimentado por la necesidad de superar el denominado "problema del marco" o "*frame problem*" que concierne a la representación de esos aspectos de la realidad cambiante que permanecen sin variación ante cambios de estado. Por ejemplo, el hecho de que yo tipee en la computadora no altera el color de las paredes de la habitación en la que me encuentro. En una representación de primer orden de tales mundos es necesario representar explícitamente todas esas cosas que no varían bajo todos los cambios de estado a través de los denominados 'axiomas de marco'. Obviamente esto es dificultoso si no imposible.

Para captar esta propiedad de no–monotonía se han elaborado desarrollos con dos enfoques diferentes: el primero de ellos se podría denominar "logicista" y abarca, entre otras, la lógica *default* de Reiter,[9] la lógica modal no–monótona de McDermmot[10] y la lógica autoepistémica de Moore.[11] El restante se podría denominar enfoque "argumentativo" y es deudor de las obras de Toulmin, Lorenzen, Rescher y Pollock y abarca desarrollos como los de Poole,[12] Simari y Loui[13] y Dung.[14]

Resulta interesante aclarar que muchos de los desarrollos de la vertiente logicista operan con nociones que no son computacionalmente implementables, lo cual ha dado más impulso al enfoque argumentativo. Por limitaciones de espacio y alcance del presente trabajo, desarrollaremos esquemáticamente sólo un representante de cada vertiente para ver de qué modo juega allí la noción de representación.

4.1 Lógica default

En la versión de Reiter de 1980 la lógica *default* se presenta como una extensión de la lógica de primer orden para dar cuenta de ese "salto" dado en la conclusión. Esto se

---

[9] Reiter, R., "A logic for default reasoning", en *Artificial Intelligence* **13**, 1980, pp. 81-132.
[10] McDermott, D. & J. Doyle; "Non-Monotonic Logic I", en *Artificial Intelligence* **13**, 1980, pp. 41-72.
[11] Moore, R., "Semantical considerations on nonmonotonic logic", en *Artificial Intelligence* **25**, 1985, pp. 75-94.
[12] Poole, D., "On the comparison of theories: preferring the most specific explanation", *Proc. of the Ninth IJCAI*, Los Altos, 1985, pp.144-147.
[13] Simari G. y Loui R., "A mathematical treatment of defeasible reasoning", en *Artificial Intelligence* **53**, 1992, pp. 125-157.
[14] Dung, P. M., "On the acceptability of arguments and its fundamental role in nonmonotonic reasoning, logic programming and n-person games", en *Artificial Intelligence* **77** (2), 1995, 321-357.

logra mediante el añadido de reglas específicas de inferencia o *defaults* que poseen la siguiente forma:

$$\frac{A(x) : B(x)}{C(x)}$$

De modo más informal: si para cierto caso x, A(x) es inferible a partir de lo que se conoce y B(x) se puede asumir consistentemente, entonces se concluye por defecto que C(x). En el ejemplo de Tweety, podríamos analizar la regla que dice que las aves vuelan como un *default* que expresa que si es posible inferir que Tweety es un ave, y se puede asumir consistentemente que Tweety vuela, entonces se infiere que Tweety vuela. El *default* que representa esta inferencia sería:

$$\frac{ave(x): vuela(x)}{vuela(x)}$$

Una teoría *default* es un par <W, D> en el que W es un conjunto de fórmulas de lógica de primer orden que representan el conocimiento estricto (no modificable, por ejemplo, el conocimiento de que Tweety es un ave, expresado como 'ave(Tweety)') y D es un conjunto de *defaults*, que representan las inferencias tentativas. Un conjunto de *defaults* induce una o más *extensiones* de las fórmulas en W. Una extensión puede ser vista como un conjunto aceptable de creencias que uno puede sostener acerca del mundo W adecuadamente argumentadas por los *defaults* D. Aquí la propiedad de no–monotonía es reflejada en que añadiendo *defaults* a una teoría las extensiones pueden cambiar.

4.2. Argumentación abstracta

Desde la otra vertiente, el sistema de Phan. M. Dung[15] propone definir extensiones a partir de la competencia de argumentos restringidos a un marco que se atacan o derrotan sin importar cuál sea su estructura interna. La noción central es la de *marco argumentativo* (*argumentation framework*) que se define como *AF* = <*AR*, *attacks*>, donde *AR* es un conjunto de argumentos y *attacks* es una relación binaria en *AR*, i.e. *attacks* $\subseteq$ *AR* $\times$ *AR*. Para representar los diferentes conjuntos de argumentos que un agente consideraría justificados y por lo tanto parte de su conjunto de creencias, Dung va a presentar distintas nociones de extensión, cada una de las cuales reflejará diversas *actitudes epistémicas*, esto es, posiciones de los agentes ante el conocimiento que involucran la credulidad y el escepticismo.

[15] Dung, M., *Op. Cit.*

Por ejemplo, consideremos un marco donde *AR* = {*A*, *B*} y *attacks*={(*A*,*B*), (*B*,*A*)}. Este ejemplo, conocido como "Diamante de Nixon",[16] se puede entender con la siguiente interpretación: *A*: "Nixon es pacifista puesto que es cuákero"; *B*: "Nixon no es pacifista puesto que es republicano". Si queremos representar una actitud epistémica crédula, podríamos definir una noción de extensión que nos entregue dos extensiones en este caso: una que contiene el argumento *A* y la otra que contiene el argumento *B*. Esto representará el hecho de que la creencia acerca del pacifismo de Nixon es indiferente dados esos argumentos, es decir, las dos extensiones pueden ser razonablemente admisibles desde un punto de vista crédulo. Por otra parte, si queremos representar una actitud epistémica escéptica, podríamos definir una noción de extensión que en este ejemplo no contenga a ninguno de los argumentos. Esto representará el rechazo de la creencia tanto en el pacifismo como en el no pacifismo de Nixon.[17]

En los sistemas argumentativos la no–monotonía surge en que las distintas extensiones pueden cambiar en virtud de una ampliación del marco, es decir, al incorporar nuevos argumentos.

Lo interesante de este sistema es que rescata los logros de otros sistemas argumentativos y logicistas sin algunos de sus defectos y es computacionalmente implementable.

Conclusiones

Para cerrar nuestra exposición apuntaremos primero a lo que consideramos la diferencia fundamental desde el punto de vista de lo representativo entre el enfoque de las lógicas no–monótonas y los sistemas argumentativos.

Por un lado, la ontología supuesta es distinta. Los elementos básicos del razonamiento de sentido común en el enfoque logicista son las reglas metalingüísticas llamadas *default*, que conectan unas creencias con otras tentativamente. Las extensiones son conjuntos infinitos de creencias que se obtienen aplicando tales reglas más las inferencias deductivas. La racionalidad mínima está dada por la consistencia interna. Nótese que éste no deja de ser un criterio lógico. Desde el punto de vista argumentativo, en cambio, los argumentos, que son los elementos básicos, pueden justificarse conjuntamente requiriendo como criterio mínimo de racionalidad que no se ataquen entre sí. El ataque no depende sólo de cuestiones lógicas, ya que la relación de ataque podría comprender cualquier tipo de incompatibilidad. Es decir, el ataque entre argumentos

[16] El término 'diamante' refiere a la forma de cierta representación gráfica de ese marco argumentativo.

abarca tanto inconsistencias sintáctica o semánticamente determinadas, como incompatibilidades pragmáticamente determinadas. Las extensiones aquí no son ya conjuntos de fórmulas tentativamente inferidas, sino conjuntos de argumentos que pueden, en algún sentido, defenderse entre sí.

Finalmente, las lógicas *default* no dicen nada acerca de cuál de las extensiones contiene las creencias mejor justificadas. Los sistemas argumentativos en cambio, permiten definir distintos criterios según los cuales o se obtiene una única extensión de argumentos justificados, o las múltiples extensiones ofrecen alternativas igualmente justificadas. Así, estos sistemas son más ricos en tanto permiten distintas caracterizaciones de actitudes epistémicas, en un rango que va de la credulidad al escepticismo.

Deberíamos agregar, para concluir, que los avances logrados en representación de conocimiento, aún cuando podamos determinar que algunos enfoques se muestran más prometedores que otros, no son más que los primeros pasos en un intento por comprender apenas una cara de la multifacética inteligencia. Desde un punto de vista crítico, aún parece una exageración hablar de "representar el conocimiento" cuando nadie sería capaz de decir, con suficiente rigor, de qué se trata el conocimiento. El hecho de que podamos simular en una computadora un juego de *flipper* de un modo muy realista, se debe a que las leyes que gobiernan los rebotes de una bola en una superficie elástica son fácilmente representables por un conjunto de ecuaciones bien probadas, y tales ecuaciones son perfectamente computables. Si programamos la computadora apropiadamente haciendo que resuelva tales ecuaciones para ciertos *inputs* apropiados, tendremos una representación de una bola real, simplemente porque tanto la bola real como la simulada siguen las mismas ecuaciones. Del mismo modo, si queremos simular el sentido común humano, primero necesitamos una lógica que sea capaz de explicar porqué razonamos como razonamos. Pero mientras para simular los rebotes de una bola real, usamos las mismas ecuaciones por las que se rigen los rebotes reales, para simular el razonamiento humano no disponemos aún de la lógica que lo gobierna.

Aún si algún día conociéramos la lógica del sentido común, eso no nos garantizaría que podamos representarlo en una computadora, pues bien podría ocurrir que esa lógica no fuera computable.[18] Por eso es que los desarrollos mencionados en representación del conocimiento se encuentran en un terreno intermedio entre la lógica filosófica y la

[17] Dung define *preferred extensions* y *grounded extensions* para representar las actitudes crédula y escéptica, respectivamente.
[18] Esto significa que exista un algoritmo que pueda finalizar en un tiempo finito.

computación. Para buscar el sistema artificial debemos a la vez buscar la teoría que le da sustento. Y es por eso que en este terreno pueden verse, trayendo y llevando ideas, a la psicología cognitiva, las ciencias de la computación, la lógica, la filosofía o la lingüística.

BIBLIOGRAFÍA

BODEN, Margaret A. (ed.), *The Philosophy of Artificial Intelligence*, Oxford, Oxford University Press, 1990.

BODANZA, Gustavo, *Un Sistema de Argumentación Rebatible Suposicional*. Tesis de doctorado. Bahía Blanca, UNS, 1999.

DELGRANDE, James y W. Ken JACKSON, "Default logic revisited", en *Proceedings of the Second International Conference on the Principles of Knowledge Representation and Reasoning,* 1991, pp. 118-127.

DUNG, Phan Minh, "On the acceptability of arguments and its fundamental role in nonmonotonic reasoning, logic programming and n-person games", en *Artificial Intelligence* **77** (2), 1995, 321-357.

Garnham, Alan, *Artificial Intelligence. An Introduction,* London, Routledge and Keagan Paul, 1988.

MCCARTHY, John y Patrick HAYES, "Some philosophical problems from de standpoint of AI", en *Machine Intelligence* **4**, 1969, pp. 463-502.

MCCARTHY, John, "Philosophical and scientific presupositions of AI", en LEVESQUE, Héctor y Fiora PIRRI (eds.), *Logical Foundations for Cognitive Agents: Contributions in Honor of Ray Reiter*, Springer-Verlag, 1999.

MCDERMOTT, Drew y Jon DOYLE; "Non-Monotonic Logic I", en *Artificial Intelligence* **13**, 1980, pp. 41-72.

MOORE, Robert, "Semantical considerations on nonmonotonic logic", *Artificial Intelligence* **25**, 1985, pp. 75-94.

POOLE, David, "On the comparison of theories: preferring the most specific explanation", *Proc. of the Ninth IJCAI*, Los Altos, 1985, pp.144-147.

REITER, Raymond, "A logic for default reasoning", en *Artificial Intelligence* **13**, 1980, pp. 81-132.

……..., “Nonmonotonic reasoning”, en *Ann. Rev. Comput. Sci.* `87, **2**, 1987, pp. 147-86.

SIMARI, Guillermo y Ronald LOUI, “A mathematical treatment of defeasible reasoning”, en *Artificial Intelligence* **53**, 1992, pp. 125-157.