Conciencia, lógica e inteligencia artificial

ALFREDO BURRIEZA MUÑIZ
Universidad de Málaga

En este artículo haremos un breve exposición del tratamiento lógico de la conciencia («awareness») en el campo de la Inteligencia Artificial (IA) [NOTA 1]. Este tema ha sido tratado en el contexto de las lógicas del conocimiento y de la creencia iniciadas por von Wright (1951) y Hintikka (1962). Para adentrarnos en el tema de la conciencia en sus aspectos lógicos arrancaremos de una distinción entre «creencia racional» y la creencia corriente o «psicológica» (ver, por ejemplo, Rescher (1968)). Para esta última, Rescher apela a la noción intuitiva de lo que es comúnmente creer; en cambio, caracteriza la creencia racional como un tipo de creencia según la cual el sujeto («el agente racional») no tiene necesidad de «ser consciente» de algo para creerlo, basta con que sea una *consecuencia lógica* de su base de creencias. De acuerdo con esta visión, resulta que el agente creería en todas las consecuencias lógicas de aquello en lo que cree, lo cual es bastante ideal y dista mucho de modelar la creencia humana. A nadie se le escapa que aunque ψ se siga lógicamente de φ y el agente crea que φ, no tiene por qué tener la habilidad deductiva para creer que ψ, especialmente si la cadena deductiva es suficientemente larga. No obstante, Rescher arguye que, aunque el sujeto no crea expresamente algo, si se deduce de lo que cree realmente, está comprometido racionalmente a creerlo. Es decir, el agente se halla en el compromiso racional de creer en cosas de las que no tiene por qué reconocer conscientemente que cree en ellas; bastaría con que se dedujeran a partir de lo que cree confesadamente. En esto consiste lo que suele denominarse «creencia racional».

Como puede apreciarse, en este planteamiento, las consecuencias lógicas de lo que el agente racional cree se convierten en creencias de

las que se pueden deducir nuevas creencias. Tenemos así una base de creencias *cerrada* bajo la operación de consecuencia lógica. Debido a esta operación de cierre se amalgaman creencias de las que el sujeto es consciente (en el sentido de que confesaría sostener tales creencias si se le preguntara) y otras de las que no lo es o bien incluso no admitiría (si igualmente se le preguntara). Como vemos, desde este punto de vista se aúnan cosas en las que se cree de modo explícito, por así decirlo, y otras que pueden ser implícitamente creídas en el sentido expuesto.

La forma en que hemos tratado la creencia (vale igualmente para la noción de conocer o saber) revela una cuestión ampliamente tratada en la literatura de la IA. Que un agente sepa (crea en) todas las consecuencias lógicas de lo que sabe (cree) y, en particular, que conozca (crea en) todas las fórmulas válidas se conoce con el nombre de «omnisciencia lógica» (OL en lo sucesivo). También se llama a tal tipo de agentes «razonadores perfectos o ideales». El término «omnisciencia lógica», introducido por Hintikka (1975), se refiere, en realidad, y dicho de una manera general, a una propiedad de cierre de la base de conocimiento (o creencias) de un agente y comprende una familia de tales condiciones de cierre. Si X es un operador de conocimiento o creencia, podemos hacer una lista de ciertas formas de OL para X como las siguientes [NOTA 2]:

Si $\models \varphi \rightarrow \psi$, entonces $\models X\varphi \rightarrow X\psi$	(cierre bajo implicación válida)
Si $\models \varphi$, entonces $\models X\varphi$	(conocimiento o creencia de fórmulas válidas)
Si $\models \varphi \leftrightarrow \psi$, entonces $\models X\varphi \leftrightarrow X\psi$	(cierre bajo equivalencia válida)
$(X\varphi \ \& \ X(\varphi \rightarrow \psi)) \rightarrow X\psi$	(cierre bajo implicación material)
$(X\varphi \ \& \ X\psi) \rightarrow X(\varphi \ \& \ \psi)$	(cierre bajo conjunción)

De un modo más preciso, podemos expresar las propiedades anteriores como auténticas propiedades de cierre, es decir, supongamos que Γ es un conjunto de creencias (conocimientos) de un agente, entonces, por ejemplo, Γ contendrá cualquier fórmula válida (conocimiento o creencia de fórmulas válidas); si $X\varphi$ y $X(\varphi \rightarrow \psi)$ pertenecen a Γ, entonces también $X\psi$ pertenecerá a Γ (cierre bajo implicación material), etc.

La OL es inofensiva para tratar el conocimiento (creencia) en ciertas áreas de la IA, como el análisis de *sistemas distribuidos* Halpern (1987), donde los agentes (procesadores) no computan su propio conocimiento ni tienen que responder a cuestiones acerca de éste. El conocimiento se atribuye «externamente» al sistema. Pero hay aplicaciones donde no es conveniente sostener la OL (e.g. *razonamiento local* o *agentes con recursos limitados*). La aceptación de las diversas formas de OL

depende, en realidad, de la clase de marcos (modelos) aceptada en la semántica. La OL surge de modo natural cuando se considera la semántica kripkeana [NOTA 3] como base para modelar conocimiento (o creencia). Todas las formas de OL expuestas anteriormente son válidas en cualquier marco (o modelo) de una semántica de Kripke.[NOTA 4]

Hay diversas formas de superar o al menos suavizar los efectos de la OL, pero no es mi intención explorar este campo, así que remito al lector a consultar Fagin et al. (1995) para una exposición sobre el tema. Sin embargo, una de dichas formas de tratar la OL es precisamente el motivo de la redacción de este artículo.

I. Creencias explicitas e implicitas

Hay un punto de vista que representa las creencias (el conocimiento) mediante conjuntos de oraciones. En este tipo de planteamientos se diferencia claramente entre creencias explícitas e implícitas, las creencias explícitas constituyen el elemento primitivo y las implícitas provienen como consecuencia lógica de las primeras a partir de un conjunto de reglas de inferencia. La simple distinción entre creencias explícitas e implícitas es una primera aproximación al tema de la conciencia desde la perspectiva lógica dentro del campo de la IA. Si empleamos la palabra «consciente» en el sentido intuitivo o vulgar del término, esto es, como sinónimo de «tener noticia de» o bien «tener en cuenta», podemos expresar esta situación diciendo que aquello de lo que el agente es consciente (lo explícito) se convierte en un generador de lo que puede no serlo (lo implícito). El punto de partida son las creencias explícitas (las que un agente cree o profesa realmente), siendo las creencias implícitas consecuencias lógicas de las primeras. La relación entre ambos tipos de creencia es de inclusión: creer explícitamente algo implica creerlo también implícitamente, aun cuando puede haber creencias «estrictamente implícitas» (de las que no se es consciente).

Un ejemplo típico de este planteamiento de carácter puramente sintáctico es el *modelo de deducción* de Konolige (1984 y 1985). Konolige presenta un sistema donde el agente razona con un conjunto, usualmente incompleto, de reglas de inferencia a partir de una base de creencias (explícitas). La pretensión es modelar el comportamiento de agentes que presentan ciertas incapacidades o limitaciones, ya sean deductivas o computacionales. En este planteamiento, sin embargo, el agente es un razonador ideal respecto de los recursos que posee, es decir, es capaz de extraer todas las consecuencias que se pueden deducir con el conjunto de sus reglas, aunque no sea capaz de extraer todas las consecuencias de

lo que sabe o cree (este sería el caso si el conjunto de sus reglas fuera incompleto). Konolige pone el ejemplo de un estudiante que puede ser un perfecto razonador, pero no sabe cuál es el valor de x en la ecuación $x + a = b$ porque desconoce la regla de «sustracción de cantidades iguales» a ambos lados del signo =.

Un planteamiento semántico para tratar creencias explícitas e implícitas se halla en Levesque (1984). Este parte también de las creencias explícitas como primarias. Las creencias explícitas son creencias de las que un agente es consciente, mientras que el agente puede no serlo de las implícitas. Levesque usa una sintaxis en la que diferencia entre operadores para la creencia explícita e implícita (para un agente) y la semántica propuesta es similar a la de los «mundos posibles» (Levesque usa lo que denomina *situaciones*). [NOTA 5]. Sin embargo, la solución propuesta por Levesque no es demasiado satisfactoria para tratar la OL (Vardi (1986)). Es cierto que, con este planteamiento, la creencia explícita no padece el problema de la OL (para la creencia implícita carece de importancia por tratarse de un tipo de creencia idealizada, no real). No obstante, el problema ataca por la espalda. El agente es un omnisciente lógico si usamos la nociones de validez y consecuencia lógica propia de la lógica de la relevancia. Tenemos, entonces, un agente que es un omnisciente lógico en otra lógica.

II. La conciencia como un operador sintáctico

El trabajo de Levesque dio pie a que Fagin y Halpern intentaran resolver la cuestión de la OL con un trabajo donde aparece expresada formalmente la noción de «conciencia». Para tratar esta cuestión, en Fagin y Halpern (1988) se presentan dos lógicas: «la lógica de la conciencia» (*awareness logic*), a la que denotaremos mediante LC, y la «lógica de la conciencia general» (*general awareness logic*), denotada LCG en lo sucesivo. De ambas merece especial atención la LCG.

La propuesta de la LCG pretende superar el problema de la OL modelando el razonamiento de agentes con recursos limitados. Queremos decir «agentes con recursos computacionales limitados» o agentes que – aunque tengan capacidad deductiva para poder obtener las consecuencias acerca de lo que creen (o saben)- se hallan limitados computacionalmente por cuestiones de tiempo, memoria, etc.

En la LCG se introduce expresamente un operador de conciencia A («awareness»), de modo que una expresión como $A_i\varphi$ se lee intuitivamente como «el agente i es consciente de que φ». Esta lectura no tiene por qué poseer el carácter cognitivo que pudiera pensarse. Aunque

podemos usar, si lo deseamos, la noción corriente de «consciente» como «tener noticia de algo» o «caer en la cuenta», lo cierto es que Fagin y Halpern pretenden que la noción de conciencia tenga gran flexibilidad y dan interpretaciones como «*i* está familiarizado con los enunciados que intervienen en φ», «*i* es capaz de comprender la verdad de φ» o -por darle una lectura más computacional- «*i* es capaz de computar la verdad de φ en un tiempo T». [NOTA 6]

El operador de conciencia se combina en el lenguaje formal con operadores de creencia explícita Be_i (el agente *i* cree explícitamente que ...) e implícita B_i (el agente *i* cree implícitamente que ...). También podemos usar operadores de conocimiento, siguiendo la exposición de Fagin et al. (1995). Denotaremos al operador de conocimiento explícito Ke_i y al implícito K_i.[NOTA 7]

Una cuestión fundamental es el *carácter sintáctico* del operador A_i, definido por Fagin y Halpern. Dicho informalmente, esto quiere decir que el significado de A_i viene dado simplemente por un listado de fórmulas. Concretamente, se define una función de conciencia A_i que recoge todas las fórmulas de las que el agente es consciente, no las que sabe efectivamente. Así, dado un estado (o mundo posible) *e*, $A_i(e)$ indica el conjunto de fórmulas de las que el agente *i* es consciente en *e*. Tenemos entonces:

$A_i\varphi$ es verdadera en *e* si y sólo si $\varphi \in A_i(e)$

El conjunto $A_i(e)$ está arbitrariamente elegido y el criterio para efectuar dicha elección es puramente sintáctico. Esta arbitrariedad permite, por ejemplo, que $A_i(e)$ pueda contener *p* y -*p* o bien sólo una de ellas o bien ninguna de las dos. También que *p* v *q* aparezca en $A_i(e)$ pero no *q* v *p*. Esto tiene sentido, por ejemplo, si identificamos $A_i\varphi$ con la lectura «*i* es capaz de computar la verdad de φ en un tiempo T» y pensamos en un computador que tratara ciertos problemas donde uno de los componentes de la disyunción tiene un elevado nivel de complejidad. En este caso el orden de la disyunción es muy importante. Pensemos en la diferencia entre computar primero «2 + 2 = 4» o bien «la suma de dos números primos mayores que dos es un número par» (conjetura de Goldbach).

Para expresar esto más formalmente, supongamos un lenguaje multimodal con *n* agentes, llamado L_n, con las conectivas booleanas habituales y los operadores Be_i, B_i, A_i (con $1 \leq i \leq n$). Para dotar de una semántica a dicho lenguaje definimos lo siguiente. Un *marco de conciencia general* es una tupla ordenada $(E, R_1, ..., R_n, A_1, ..., A_n)$, donde:

(a) E es un conjunto no vacío de «estados» (mundos posibles),
(b) cada R_i ($1 \leq i \leq n$) es un subconjunto de E x E, o relación de *accesibilidad* (la correspondiente al agente i) en E, que es serial, transitiva y euclídea. [NOTA 8] Los mundos a los que se accede desde uno dado se entiende que son los estados que el agente i considera como posibles desde dicho estado. Aunque la relación de accesibilidad de cada agente posee las propiedades mencionadas, podemos debilitar dichas restricciones dando mayor generalidad a los marcos de conciencia general (Thijsse (1992 y 1993) y Fagin et al. (1995)).
(c) cada A_i ($1 \leq i \leq n$) es una función de E en L_n. Intuitivamente, $A_i(e)$ es el conjunto de fórmulas de las que el agente i es consciente en el estado e.

Con estas ideas podemos definir un *modelo de conciencia general* como una secuencia de la forma $(E, R_1, ..., R_n, A_1, ..., A_n, V)$, donde:

(a) $(E, R_1, ..., R_n, A_1, ..., A_n)$ es un marco de conciencia general
(b) V es una función que asigna a cada átomo o conjunto de fórmulas atómicas primitivas de L_n (ATOM) un valor de verdad (1 ó 0) en cada estado de E, i.e. V es una función de ATOM x E en $\{0, 1\}$.

Extenderemos la función V –abusando de la notación- para evaluar toda fórmula del lenguaje de L_n en un estado cualquiera e del modelo. Para las conectivas clásicas procedemos como es habitual. Para las modales tenemos las cláusulas siguientes:

$V(B_i\varphi, e) = 1$ si y sólo si para todo e' tal que eR_ie', $V(\varphi, e') = 1$
$V(A_i\varphi, e) = 1$ si y sólo si $\varphi \in A_i(e)$
$V(Be_i\varphi, e) = 1$ si y sólo si $\varphi \in A_i(e)$ y para todo e' tal que eR_ie', $V(\varphi, e') = 1$
(i.e. $V(Be_i\varphi, e) = 1$ si y sólo si $V(A_i\varphi, e) = 1$ y $V(B_i\varphi, e) = 1$)

Las definiciones de validez son las usuales, es decir, una fórmula de L_n es *válida en un modelo de conciencia general* si y sólo si es verdadera en cada estado del modelo. Una fórmula es *válida* si y sólo si es válida en todos los modelos de conciencia general.

Nótese que en la semántica propuesta, el operador B_i se define como en los modelos de Kripke, el operador A_i se define sintácticamente y Be_i es un operador de carácter mixto. Si en lugar de los operadores de creencia usamos los de conocimiento, obtenemos cláusulas similares a las

anteriores sustituyendo simplemente B_i por K_i y B^e_i por K^e_i como presentan (Fagin et al., 1995). Así que podemos hablar indistintamente de creencia o conocimiento. No obstante, hay que tener en cuenta que, al tratar conocimiento, las condiciones de la relación R_i en el modelo pueden cambiar. Es usual que se trate en este caso de una relación de equivalencia [NOTA 9].

La idea básica de este planteamiento, expresado de manera intuitiva, es que el agente ha de ser consciente de algo para poderlo creer (o conocer) de modo explícito (no podemos esperar que alguien crea o no que Feynman fue premio Nobel de Física si carece por completo de noticia de la existencia de Feynman). Conviene notar que -contrariamente a los planteamientos de Levesque y Konolige- aquí se invierte el orden de prioridad entre lo explícito y lo implícito merced a la aparición autónoma de la conciencia. Se parte de las creencias implícitas y de lo consciente como elementos primitivos. Lo que se intenta caracterizar es el conjunto de creencias reales del agente. La intersección entre el conjunto de lo creído implícitamente y de aquello de lo que se tiene conciencia arroja como resultado la creencia explícita. La conciencia actúa, pues, como un filtro sintáctico que deja pasar ciertas creencias implícitas, que son precisamente las creencias explícitas. Podemos expresar esto en una máxima:

La creencia explícita es la creencia implícita de la que se es consciente.

Esto se refleja en la ley: $B^e_i\varphi <-> (A_i\varphi \ \& \ B_i\varphi)$ (alternativamente tenemos: $K^e_i\varphi <-> (A_i\varphi \ \& \ K_i\varphi)$. Esto es, el conocimiento explícito es el conocimiento implícito del que se es consciente).

La conciencia sirve para determinar o caracterizar la creencia real de acuerdo con los recursos de los que dispone el agente. Si pensáramos, por ejemplo, que los agentes poseen recursos ilimitados (son conscientes de todas las fórmulas), entonces coincidirían las nociones de creencia explícita e implícita: $B^e_i\varphi <-> B_i\varphi$.

La otra lógica desarrollada por Fagin y Halpern para tratar expresamente la conciencia, la LC, se caracteriza porque es una extensión de la lógica de Levesque, limita la conciencia a fórmulas atómicas primitivas y en ella no aparece definido el operador de conciencia, aunque sí la función de conciencia para modular el significado de la creencia explícita. No obstante, el operador de conciencia se puede introducir por definición. El agente es consciente de un hecho p cuando cree explícitamente que $p \vee -p$. Generalizando esto, si $p_1, ..., p_n$ son todos los átomos

que intervienen en una fórmula φ, entonces $A_i\varphi$ se define como $B^e_i(p_1 \vee -p_1) \& \ldots \& B^e_i(p_n \vee -p_n)$. Lo que se pretende recoger es que φ está en la conciencia del agente cuando éste tiene una creencia en la que φ se ve involucrada. Esto supone una noción recursiva de la conciencia, pues el agente posee suficiente capacidad para ascender a expresiones complejas a partir de sus átomos.

III. Propiedades de la conciencia

No parece muy razonable dejar que el conjunto $A_i(e)$ se defina arbitrariamente, por este motivo pueden imponerse diversas restricciones a dicho conjunto (Fagin y Halpern, 1988). Esto significa que tenemos que considerar diferentes formas de conciencia siguiendo el criterio sintáctico adoptado. Por ejemplo:

1- El agente *i* es consciente de todas las subfórmulas de las fórmulas con las que razona. Esto significa que el agente sería un magnífico analizador (completo) de aquello de lo que es consciente. El sentido que Fagin y Halpern dan a esta restricción es que podemos suponer que el agente es un programa que trabaja sobre una base de conocimiento y requiere computar todas las subfómulas de una fórmula para decidir acerca del valor de verdad de ésta. En este contexto, eliminar esta restricción nos conduce a lo contrario, es decir, a un programa que diera cuenta, por ejemplo, de que se da que $p \vee -p$ sin necesidad de computar la verdad de *p*.

La propiedad de que el conjunto $A_i(e)$ es cerrado bajo subfórmulas se corresponde con los siguientes axiomas:

$A_i -\varphi \rightarrow A_i\varphi$
$A_i(\varphi \& \psi) \rightarrow (A_i\varphi \& A_i\psi)$
$A_iB^e_i\varphi \rightarrow A_i\varphi$
$A_iB_i\varphi \rightarrow A_i\varphi$
$A_iA_j\varphi \rightarrow A_i\varphi$

Una consecuencia de esto es que el operador B^e_i es cerrado bajo la implicación material (una forma de la OL), lo que no resulta adecuado si se intenta representar una noción de conciencia propia de agentes con recursos limitados, donde tener conciencia de φ se entiende como ser capaz de computar la verdad de φ (tengamos en cuenta que el agente podría ser consciente sólo de ciertas subfórmulas de φ). Esto nos conduce a la restricción siguiente.

2- El agente *i* es consciente de algunas subfórmulas y no de otras. Queda por decidir acerca del carácter de dicha elección. Uno podría ser el siguiente: el agente *i* es consciente de todas las subfórmulas de una fórmula en forma conjuntiva. Esto es, si φ & ψ pertenece a $A_i(e)$, entonces φ y ψ también pertenecen a $A_i(e)$. Otra restricción de este estilo es que el agente podría ser sólo consciente de las subfórmulas generadas por un subconjunto previamente fijado (PRIM) del conjunto de fórmulas atómicas primitivas. Por tanto, en este caso, $A_i(e)$ contiene exactamente aquéllas fórmulas que se generan a partir de PRIM. Entonces se cumple:

$A_i\text{-}\varphi <\text{-}> A_i\varphi$
$A_i(\varphi \,\&\, \psi) <\text{-}> (A_i\varphi \,\&\, A_i\psi)$
$A_iBe_i\varphi <\text{-}> A_i\varphi$
$A_iB_i\varphi <\text{-}> A_i\varphi$
$A_iA_j\varphi <\text{-}> A_i\varphi$

Una consecuencia interesante de este supuesto es que se puede expresar el hecho de que la gente normalmente es consciente de sus creencias explícitas, en el sentido intuitivo de que se da cuenta de que cree algo cuando lo cree explícitamente. Es decir:

$Be_i\varphi \text{-}> A_iBe_i\varphi$

3- Propiedad de cierre de la función A_i: el conjunto $A_i(e)$ contendrá $A_i\varphi$ siempre que contenga φ. En este caso, los agentes son auto-reflexivos, en el sentido de que son conscientes de que son conscientes, i.e. valdría

$A_i\varphi \text{-}> A_iA_i\varphi$

Hay variados motivos para aceptar o rechazar el anterior axioma según interpretemos la noción de conciencia. Si apelamos a la noción psicológica de la conciencia como *percepción* (Huang y Kwast, 1991), según la cual ser consciente de algo es percibirlo, se puede rechazar el anterior axioma. En cambio, si acudimos a la noción psicológica corriente de conciencia para los agentes humanos como «darse cuenta» o a programas que respondan al punto 2 anterior, podemos aceptarlo.

No obstante, en este punto nos asalta una cuestión que planea sobre este planteamiento de la conciencia. Una expresión como $A_iA_i\varphi$ puede

interpretarse –en principio- de dos formas: como ser consciente de un acto mental, o bien como ser consciente de una oración como «el agente *i* es consciente de que φ». Esto nos retrotrae a la cuestión ya planteada desde hace largo tiempo de que los enunciados que expresan *actitudes proposicionales* (introducidas por verbos psicológicos como «saber», «creer», «pensar», «desear», etc.) pueden entenderse de diversas formas: como una relación entre un agente y una oración, o como una relación entre un agente y una representación mental, o bien como una relación entre un agente y una *proposición* (entendida como la *intensión* de una oración). En términos de una semántica de mundos posibles, la proposición sería un conjunto de mundos posibles (precisamente aquellos en los que la oración es verdadera). Este punto es muy controvertido, y con respecto a la conciencia nos hallamos ante diversas consideraciones. Sin embargo, autores como Konolige (1986) y Ule (2000) han señalado expresamente que el tratamiento sintáctico de la conciencia trabaja sobre oraciones, no sobre proposiciones.

4- Propiedades de *monotonía* de la función A_i: en este caso establecemos un puente entre la relación de accesibilidad y la función de conciencia. Podemos establecer que $A_i(e)$ está incluido (no estrictamente) en $A_i(e')$ cuando eR_ie' (*monotonía creciente*); o bien lo contrario. Es decir, que $A_i(e')$ está incluido (no estrictamente) en $A_i(e)$ cuando eR_ie' (*monotonía decreciente*).

De acuerdo con esto, el agente posee una creencia implícita de que es o no consciente de algo. Tendríamos entonces como leyes, según el tipo de monotonía admitida, las siguientes:

$A_i\varphi \rightarrow B_iA_i\varphi$ (monotonía creciente)
$-A_i\varphi \rightarrow B_i{-}A_i\varphi$ (monotonía decreciente)

Si aceptamos ambos axiomas, entonces imponemos a la función A_i la monotonía en ambas direcciones, es decir:

Si eR_ie', entonces $A_i(e) = A_i(e')$

Esta condición cobra sentido, por ejemplo, si consideramos que eR_ie' significa que el agente *i* no distingue cognitivamente el estado *e* del estado *e'*, entonces parece natural que el agente sea consciente en ambos estados de las mismas fórmulas. En sistemas multiagentes esto significa que el conjunto de lo que el agente es consciente se halla en función de su *estado local*. [NOTA 10] Asimismo, estos axiomas son también acep-

tables cuando el conjunto de lo consciente es generado por un subconjunto de fórmulas atómicas primitivas (punto 2 anterior).

5. El conjunto $A_i(e)$ puede componerse de aquellas fórmulas que un agente i puede determinar en alguna cantidad especificada de tiempo o espacio, tanto si aquéllas se siguen o no de la información que i posee. Esta limitación es discutida más ampliamente en Fagin et al. (1995) y dedicaremos a ella más atención al tratar el conocimiento algorítmico.

IV. Conciencia y OL

Una de las fuentes del problema de la OL -según Fagin y Halpern- es la carencia de conciencia. Con el uso de los modelos de conciencia, el problema de la OL desaparece para la creencia explícita (la OL queda recluida a la creencia implícita). Esto se debe a que la creencia implícita soporta las condiciones de la semántica kripkeana, mientras que la modulación sintáctica de la conciencia permite eliminar o rescatar diversas formas de la OL para la creencia explícita. Así, mientras que para el conjunto de creencias implícitas vale que es cerrado bajo la implicación válida y la regla de creencia de fórmulas válidas, en cambio, para el conjunto de creencias explícitas la cosa difiere. Por ejemplo, los agentes no creen explícitamente todas las fórmulas válidas, pues basta con que alguna fórmula válida no pertenezca a $A_i(e)$ por definición. Procediendo de forma similar, puede mostrarse que tampoco vale el cierre bajo implicación ni equivalencia válidas. Más aún, tampoco se sostienen otras formas de la OL. El conjunto de creencias explícitas no es cerrado bajo la implicación material ni bajo conjunción, es decir, no son leyes:

$(B^e_i\varphi \;\&\; B^e_i(\varphi \rightarrow \psi)) \rightarrow B^e_i\psi$
$(B^e_i\varphi \;\&\; B^e_i\psi) \rightarrow B^e_i(\varphi \;\&\; \psi)$

En general, manipulando simplemente el conjunto de lo que un agente es consciente conseguimos refutar todas estas características. En cambio, contamos con las condiciones siguientes:

Si $\models \varphi \rightarrow \psi$, entonces $\models A_i\psi \rightarrow (B^e_i\varphi \rightarrow B^e_i\psi)$
Si $\models \varphi$, entonces $\models A_i\varphi \rightarrow B^e_i\varphi$

Con esto se destaca claramente la condición de que el agente tenga que ser consciente previamente de las fórmulas válidas para tener una

creencia explícita de las mismas. Esto significa que el agente es un razonador ideal restringido al conjunto de enunciados de los que es consciente. Quizá esto no parezca muy adecuado para modelar agentes con recursos limitados. No obstante, existen planteamientos que atiende más satisfactoriamente a esta cuestión (el modelo computacional de la conciencia) y que abordaremos más adelante.

V. Axiomatizaciones relativas a la conciencia

Hay una serie de axiomas modales (hablando con rigor, son *esquemas de axiomas*) de conocimiento y creencia que conviene considerar. Son los siguientes:

K $X(\varphi \rightarrow \psi) \rightarrow (X\varphi \rightarrow X\psi)$
D $X\varphi \rightarrow \neg X\neg\varphi$
T $X\varphi \rightarrow \varphi$
4 $X\varphi \rightarrow XX\varphi$
5 $\neg X\varphi \rightarrow X\neg X\varphi$

donde X es un operador de conocimiento o de creencia. De estos axiomas, **K** indica la operación de cierre de la implicación material respecto del conocimiento o la creencia (una de las formas de OL). **D** indica que las creencias o conocimiento del agente no es inconsistente. **T** no es adecuado para tratar la creencia (todo lo que cree el agente es cierto) y no se utiliza en las axiomatizaciones al uso. En cambio, **T** sí es admisible para tratar conocimiento, ya sea implícito o explícito (según este esquema el agente no «sabe» cosas falsas). Por su parte, **4** y **5** se denominan respectivamente axiomas de *introspección positiva* e *introspección negativa* y pueden ser discutidos o aceptados (según las aplicaciones) tanto para conocimiento como para creencia. **4** informalmente dice que el agente sabe (cree) que sabe (cree) algo cuando lo sabe (cree). Por su parte, **5** intuitivamente dice que el agente sabe (cree) que desconoce (no cree) algo cuando lo desconoce (no lo cree).

Fagin y Halpern presentan un sistema axiomático correcto y completo para la LCG (o sea, respecto de la clase de modelos seriales, transitivos y euclídeos). Se trata del sistema modal $KD45_n$ (para el operador B_i) al que se añade el axioma $Be_i\varphi \leftrightarrow (A_i\varphi \,\&\, B_i\varphi)$. $KD45_n$ viene definido por lo siguiente:

- **Prop**. Todas las instancias de sustitución de las tautologías veritativo funcionales
- Los axiomas **K**, **D**, **4** y **5** (donde X es B_i)
- Las reglas:
  Si |- φ y |- φ -> ψ, entonces |- ψ (*modus ponens*)
  Si |- φ, entonces |- B_iφ [NOTA 11] (generalización doxástica)

En Fagin et al. (1995) se proponen sistemas axiomáticos correctos y completos respecto de clases de modelos de conciencia donde la relación de accesibilidad ya no posee las características de los modelos del planteamiento de Fagin y Halpern (1988). Estos sistemas usan operadores de conocimiento en lugar de operadores de creencia para el caso multiagente. Son los que resultan de añadir a los sistemas K_n o $S5_n$ (para el operador K_i) [NOTA 12] el axioma $K^e_i\varphi <-> (A_i\varphi \& K_i\varphi)$ y las equivalencias que expresan que la conciencia es generada por un subconjunto de fórmulas atómicas primitivas (son las equivalencias del apartado 2 de la sección III, pero sustituyendo B^e_i por K^e_i y B_i por K_i).

En Halpern (2000) se definen nuevos sistemas axiomáticos para el caso de un único agente (en este caso eliminamos el subíndice de los operadores modales). Tenemos la siguiente lista de axiomas:

A0. $K^e\varphi <-> (A\varphi \& K\varphi)$

A1.	$A\text{-}\varphi -> A\varphi$	A7.	$A\varphi -> A\text{-}\varphi$
A2.	$A(\varphi \& \psi) -> (A\varphi \& A\psi)$	A8.	$(A\varphi \& A\psi) -> A(\varphi \& \psi)$
A3.	$AK^e\varphi -> A\varphi$	A9.	$A\varphi -> AK^e\varphi$
A4.	$AK\varphi -> A\varphi$	A10.	$A\varphi -> AK\varphi$
A5.	$AA\varphi -> A\varphi$	A11.	$A\varphi -> AA\varphi$
A6.	$A\varphi -> KA\varphi$	A12.	$\text{-}A\varphi -> K\text{-}A\varphi$

Consideremos los conjuntos C_1 = {A1,.., A5}, C_2 = {A1, ..., A10}, C_3 = {A11}, C_4 = {A6, A12}. C_1 se compone de los axiomas que señalan que la conciencia es cerrada bajo subfórmulas, C_2 señala que la conciencia es generada por un subconjunto de fórmulas atómicas primitivas, C_3 se refiere a la reflexividad de la conciencia y C_4 al conocimiento de la propia conciencia (monotonía de la función conciencia en ambas direcciones). Supongamos que S es cualquiera de los sistemas modales K, T, S4 y S5 (para el operador de conocimiento implícito monoagente K), entonces podemos definir diferentes sistemas compuestos por S, A0 y los axiomas de C_i (para i = 1, ..., 4). Halpern demuestra la corrección y completitud de los distintos sistemas respecto de la clase de modelos correspondiente. Estas clases atienden a las propiedades de la relación

de accesibilidad características de los modelos de S y a las restricciones de la función de conciencia mencionadas anteriormente. Señalemos que los modelos de K no poseen ninguna condición especial respecto de la relación de accesibilidad, los de T son reflexivos, los de S4 son reflexivos y transitivos, y los de S5 son modelos donde dicha relación es de equivalencia. Por ejemplo, el sistema compuesto por S5, A0 y C_2 es correcto y completo respecto de la clase de modelos donde la relación de accesibilidad es de equivalencia y la conciencia es generada por un subconjunto de fórmulas atómicas primitivas.

Halpern da otras axiomatizaciones eliminando el conocimiento implícito y con versiones modificadas de **K**, **T**, **4** y **5** (para K^e). Por ejemplo, las versiones de **4** y **5** son respectivamente:

4' $(K^e\varphi \,\&\, AK^e\varphi) \rightarrow K^eK^e\varphi$
5' $(\text{-}K^e\varphi \,\&\, A\text{-}K^e\varphi) \rightarrow K^e\text{-}K^e\varphi$

Informalmente, estos axiomas indican que para que el agente sepa que sabe (resp. desconoce) algo, se requiere que sea consciente de que sabe (resp. desconoce) dicha cuestión. Para más detalles, remito al lector a (Halpern, 2000).

VI. Algunos aspectos del planteamiento de la conciencia

Ya hemos comentado la relación entre el planteamiento de la conciencia y la solución a OL. Señalaremos, además, otras cuestiones relacionadas con este planteamiento:

1.- Consideremos los axiomas típicos de introspección positiva y negativa con referencia a la creencia explícita o bien el conocimiento explícito. Tomemos, por ejemplo, el operador de conocimiento explícito K^e_i:

$K^e_i\varphi \rightarrow K^e_iK^e_i\varphi$ (si *i* sabe que φ, sabe que lo sabe)
$\text{-}K^e_i\varphi \rightarrow K^e_i\ \text{-}K^e_i\varphi$ (si *i* no sabe que φ, sabe que no lo sabe)

Estos axiomas no se cumplen con la semántica de la conciencia general (ni siquiera suponiendo que R_i sea una relación de equivalencia). Para recuperar estos axiomas hemos de suponer -además de que R_i sea una relación de equivalencia- que la función A_i tenga la propiedad de la monotonía en ambas direcciones: si eR_ie', entonces $A_i(e) = A_i(e')$. En

este mismo caso también se cumplen las modificaciones de estos dos axiomas teniendo en cuenta la conciencia, es decir los ya mencionados **4'** y **5'** en la sección anterior.

2- Una propiedad interesante del planteamiento de la conciencia es su *potencia expresiva* comparado con otros planteamientos que tratan la OL. La generalización de la LCG efectuada por Thijsse (1992 y 1993) posee la misma expresividad que planteamientos tan expresivos como el de «mundos imposibles» o las «estructuras sintácticas» (ver también Fagin et al. (1995)). Intuitivamente, esto significa que cualquier situación descrita por uno de estos planteamientos puede ser capturado por los otros.

3- La incorporación de un operador de conciencia al formalismo es un modo de tratar las creencias inconsistentes de un agente sin caer por ello en ciertas idealizaciones. Un agente en el mundo real puede creer (implícitamente) que φ y que -φ sin caer en la cuenta. Bastaría con que creyera que ψ, siendo ψ <-> -φ, y que el agente no creyera que se da dicha equivalencia. Desde un punto de vista semántico, si intentáramos modelar dicha conducta dentro del planteamiento kripkeano, nos encontraríamos con que un agente que creyera tanto que φ como que -φ se hallaría en un estado «terminal», es decir, no podría concebir ninguna alternativa posible al estado en el que se encuentra, o técnicamente hablando, si el agente i se encuentra en un estado e tendríamos que el conjunto $\{e'/eR_ie'\}$ es vacío. Este planteamiento es poco satisfactorio para tratar la cuestión planteada, pues en un estado terminal el agente creería en toda fórmula. En los modelos de conciencia podemos solventar esta cuestión, pero no con el planteamiento de Fagin y Halpern (1988), debido a la propiedad de serialidad de la relación de accesibilidad. La inconsistencia se puede modelar estableciendo de nuevo que el conjunto $\{e'/eR_ie'\}$ sea vacío pero, además, que dada cualquier fórmula φ, φ pertenezca a $A_i(e)$ si y sólo si $B^e{}_i\varphi$ es verdadera en e; o sea, que la conciencia coincida con las creencias explícitas del agente. Esto no conduce a que el agente crea explícitamente en todo. Basta con manipular convenientemente el conjunto $A_i(e)$.

VII. Una alternativa a la LCG

Una alternativa a lo planteado por Fagin y Halpern para tratar también sintácticamente la conciencia aparece en Huang y Kwast (1991). Pero en este planteamiento, la definición de creencia explícita es menos

general que en la LCG. La intención de estos autores es capturar una noción de creencia explícita más estricta que la de Fagin y Halpern. Para ello, Huang y Kwast proponen la definición siguiente:

$$Be_i\varphi <-> (B_i\varphi \ \& \ A_iB_i\varphi)$$

Esta definición permite capturar la nueva noción de creencia explícita en términos de la «antigua» como sigue:

$$He_i\varphi <-> A_iBe_i\varphi$$

(denotamos mediante He a la conectiva de creencia explícita de Huang y Kwast para distinguirla de la de Fagin y Halpern). Es decir, la idea de este nuevo planteamiento es que el agente cree explícitamente algo (tiene una creencia real de algo) precisamente cuando es consciente de tal creencia. La definición de Huang y Kwast es el resultado de imponer a la función de conciencia condiciones como las siguientes:

- propiedades de cierre de la función A_i respecto de – y &:
  A_i -$\varphi \in A_i(e)$ si y sólo si $A_i\varphi \in A_i(e)$
  $A_i(\varphi \ \& \ \psi) \in A_i(e)$ si y sólo si $A_i\varphi \in A_i(e)$ y $A_i\psi \in A_i(e)$
- si $A_i\varphi \in A_i(e)$, entonces $\varphi \in A_i(e)$
- $A_i\varphi \in A_i(e)$ si y sólo si $B_i\varphi \in A_i(e)$

Huang y Kwast ven razonables e intuitivas las dos primeras propiedades, mientras que reconocen que la tercera es un tanto especial. La primera propiedad ya ha sido comentada, nos remite a la generación de la conciencia a partir de un conjunto de fórmulas atómicas primitivas, pero sólo con la ayuda de conectivas booleanas. La segunda propiedad conduce a aceptar $A_iA_i\varphi$ -> $A_i\varphi$, que Huang y Kwast justifican entendiendo la conciencia como percepción (si uno percibe que percibe alguna cosa, percibe esa cosa). La tercera propiedad trae como consecuencia aceptar como ley: $A_iB_i\varphi <-> A_iA_i\varphi$ («interpretación de la conciencia de la creencia»). En Thijsse (1993) se discute su supuesta intuición, especialmente la dirección ->. Comparto las dudas de Thijsse respecto de dicha dirección, pero tampoco veo una intuición clara respecto de la otra. Además, dicha propiedad, en combinación con $A_iA_i\varphi$ -> $A_i\varphi$, tiene otras consecuencias poco intuitivas, como, por ejemplo, $B_iB_i\varphi$ -> $A_i\varphi$.

VIII. Conciencia y tiempo

Con la inclusión del tiempo se abren nuevas perspectivas para el tratamiento de la conciencia en el campo de la IA desde la perspectiva lógica considerada. La incorporación de operadores temporales permite aumentar la potencia expresiva de los lenguajes modales que contienen el operador de conciencia y tratar situaciones más complejas. Un buen ejemplo son las creencias inconsistentes de los agentes humanos. Por ejemplo, se puede modelar la situación de un agente humano que descubre una inconsistencia en su base de creencias y modifica ésta con intención de eliminarla. Esta operación se desarrolla de modo natural en el tiempo, así que la incorporación de operadores temporales está más que justificada. Si usamos, por ejemplo, el operador O de tiempo discreto (cuya lectura informal es «mañana» o «en el instante siguiente»), en el caso discutido tendríamos que aceptar como axioma:

$$(Be_i\varphi \;\&\; Be_i\text{-}\varphi \;\&\; A_i(Be_i\varphi \;\&\; Be_i\text{-}\varphi)) \rightarrow O\text{-}(Be_i\varphi \;\&\; Be_i\text{-}\varphi)$$

Esto es, si el agente i cree (explícitamente) que φ y también cree en su negación y además es consciente de ambas cosas, entonces inmediatamente resuelve la inconsistencia dejando de creer en una de las dos.

También podemos expresar otras propiedades que requieren tener en cuenta aspectos temporales. Usemos los operadores de Prior G (será siempre el caso que) y F (será alguna vez el caso que). Ofrezco un botón de muestra:

$A_i\varphi \rightarrow GA_i\varphi$ (conservación de la conciencia)

Si el agente i es consciente de que φ, entonces siempre será consciente de que φ.

$\text{-}A_i\varphi \rightarrow FA_i\varphi$ (incremento de la conciencia)

Si el agente i no es consciente de que φ, entonces algún día lo será.

$A_i\varphi \rightarrow F\text{-}A_i\varphi$ (pérdida de la conciencia)

Si el agente i es consciente de que φ, entonces algún día dejará de serlo.

$(\varphi \;\&\; A_i\varphi \;\&\; \text{-}Ke_i\varphi) \rightarrow FKe_i\varphi$ (optimismo cognitivo)

Si el agente *i* es consciente de que φ pero no sabe que φ es verdadera, tarde o temprano lo sabrá.

IX. Interpretaciones del operador de conciencia en la IA

Hay varios campos dentro de la investigación lógica en IA donde pueden encontrarse definiciones del operador de conciencia introducido por Fagin y Halpern. Tocaremos dos: (i) el *cómputo de conocimiento*, que ofrece un modelo computacional de la conciencia y (ii) la *dependencia de creencias*, que ofrece una visión más psicológica.

El problema de computar conocimiento está relacionado con la cuestión de la OL. Una razón por la cual los agentes pueden no ser omniscientes es su limitación de recursos computacionales. En las aplicaciones en las que los agentes necesitan manipular su propio conocimiento (e.g. *bases de conocimiento*), el conocimiento implícito resulta insuficiente, pues es frecuente que los agentes realicen sus acciones basándose en el conocimiento explícito. En esta sección vamos a caracterizar computacionalmente la noción de conocimiento y esto requiere contemplar sistemas donde los agentes dispongan de algoritmos para contestar a cuestiones relativas a su propio conocimiento. Tales sistemas, en los que pueden interactuar diversos agentes, se denominan *sistemas algorítmicos*. Cuando se habla de sistemas algorítmicos, los estados locales de un agente consisten tanto en los algoritmos a los que el agente puede acceder como datos o información. Hay que tener en cuenta, además, para una perspectiva realista, los recursos de estos algoritmos, como el tiempo disponible, etc. En este contexto, los estados locales se denominan *estados algorítmicos* y el conocimiento explícito se denomina *conocimiento algorítmico* (Moses (1988)).

Dicho informalmente, un agente *i sabe algorítmicamente* que φ si puede computar que sabe que φ (es decir, que tiene un conocimiento implícito o racional de que es el caso que φ). En otras palabras, el agente tiene que poder «explicitar» mediante cómputo que posee dicho conocimiento ideal. Para ser algo más precisos hemos de relativizar tal saber a los estados algorítmicos del agente *i* a lo largo de una serie de momentos del tiempo. En un estado semejante, el agente *i* dispone de un algoritmo A y de datos *l*. Entonces pone en marcha A con los datos *l* y la fórmula φ como entrada y se trata de ver si la salida es SI. Estamos considerando algoritmos que pueden dar la respuesta SI, NO o ? (bajo el supuesto de que siempre terminan). Lo que esto significa es que el agente manipula su conocimiento racional o implícito para comprobar si tiene tal conocimiento de que es el caso que φ aplicando A con los datos *l*. En caso de

que el agente sea capaz de dar la respuesta SI, entonces es que posee conocimiento de la veracidad de φ, lo que desemboca en un conocimiento real o explícito de su veracidad; si da la respuesta NO, es que tiene conocimiento ideal de la falsedad de φ (esto no significa que el agente sepa explícitamente que φ es falsa, simplemente que no sabe explícitamente que φ). La respuesta ? indica que el agente es incapaz de computar su conocimiento ideal acerca de la cuestión φ. Así pues, tanto si da la respuesta NO como la respuesta?, ello indica que el agente no tiene conocimiento explícito de que φ sea el caso. De una manera un tanto más formal, el agente *i* tiene conocimiento algorítmico de (sabe explícitamente) que φ (i.e. $K^e_i\varphi$) cuando es cierto que $K_i\varphi$ (el algoritmo dice «SI»). Carece de conocimiento algorítmico de (no sabe explícitamente) que φ (i.e. $-K^e_i\varphi$) cuando o bien es cierto que K_i-φ (el algoritmo contesta «NO») o bien no es cierto ni $K_i\varphi$ ni K_i-φ (el algoritmo contesta ?).

Los algoritmos pueden cometer errores, luego podría ser el caso que un algoritmo contestara SI (luego sería cierto que $K^e_i\varphi$) aunque no se diera que φ, con lo cual se estaría recogiendo más bien una noción de creencia que de conocimiento, así que interesan especialmente algoritmos carentes de ellos (*correctos*). Un agente que maneja siempre algoritmos correctos puede interpretarse como un «agente racional» (no utiliza procedimientos defectuosos). Si además, el agente está dotado de un algoritmo que nunca da la respuesta ?, su algoritmo es *completo*. La completitud se relaciona con la «experiencia» del agente: un agente más experto que otro dará menos respuestas del tipo ? ante las mismas preguntas.

El carácter sintáctico del conocimiento algorítmico establece una estrecha conexión con el tratamiento igualmente sintáctico de la conciencia. En el contexto de sistemas algorítmicos una expresión como $A_i\varphi$ (*i* es consciente de que φ) significa que el algoritmo local de *i* da la respuesta SI o NO cuando se le presenta φ. La respuesta ?, que excluimos, significaría que el agente carece de conciencia de que φ. Esto quiere decir que la conciencia significa que el agente posee la capacidad de decidir su conocimiento acerca de la fórmula φ, sea cual sea φ. El no ser consciente de algo, significaría entonces un fallo o incapacidad en el cómputo de dicho conocimiento.

En este contexto, un agente sabe implícitamente si tiene o no conocimiento algorítmico; o sea, tenemos como fórmulas válidas: $K^e_i\varphi \rightarrow K_iK^e_i\varphi$ y $-K^e_i\varphi \rightarrow K_i$-$K^e_i\varphi$ (la relación usada en los modelos para tratar el operador K_i es de equivalencia lo que corresponde al sistema S5). Sin embargo, no valen en general axiomas análogos para el operador de conciencia, esto es,

$Ke_i\varphi \rightarrow A_iKe_i\varphi$
$-Ke_i\varphi \rightarrow A_i-Ke_i\varphi$.

La validez de estos axiomas requeriría presuponer que el agente es capaz de dar una respuesta definida (SI o NO) ante la pregunta de si sabe si tiene conocimiento algorítmico de la fórmula φ. Además, si los algoritmos que maneja el agente son correctos, entonces el operador Ke_i satisface: $Ke_i\varphi \rightarrow \varphi$ y $Ke_i\varphi \rightarrow K_i\varphi$. Más aún, la corrección del algoritmo es garantía de que el operador de conciencia satisfaga la definición ya conocida: $Ke_i\varphi \leftrightarrow (K_i\varphi \ \& \ A_i\varphi)$. En cambio, si el algoritmo usado además de correcto es completo, entonces se satisface el axioma $Ke_i\varphi \leftrightarrow K_i\varphi$ (el conocimiento implícito se reduce al explícito) (ver Fagin et al.(1995)).

Es frecuente en la vida real que las personas consulten la opinión de otras acerca de diversos temas, de manera que lo que sabe o cree un agente depende en buena medida de lo que aprende de otros (este fenómeno se llama «dependencia de creencias»). Un área de aplicación de este concepto son los *sistemas distribuidos*. Estos consisten en una colección de agentes (humanos, procesadores o robots) que, generalmente, cuentan con recursos limitados y se hallan conectados por una red de comunicación. En este tipo de entornos, el razonamiento acerca del conocimiento y la creencia ha producido variadas aplicaciones (en *las bases distribuidas de conocimiento*, en *la comunicación y cooperación para planificación multiagente en IA* o *en la ingeniería del conocimiento*).

En entornos distribuidos con dependencia de creencias encontramos en Huang (1990) un *modelo general de dependencia de creencias* para sistemas multiagentes basado en mundos posibles o estados. El lenguaje usado, que denotaremos L^D_n (para n agentes), contiene las conectivas clásicas y los operadores B_i (para la creencia implícita) y $D_{i,j}$ (para la dependencia de creencias). Este último operador se denomina *operador de dependencia*. Una expresión como $D_{i,j}\varphi$ se lee «el agente i depende de j para creer que φ». Otras lecturas posibles son las siguientes: «el agente j es el mejor consejero del agente i acerca de la fórmula φ», «el agente i pregunta al agente j acerca de la fórmula φ» o, en redes de procesos distribuidos, «el procesador i puede obtener conocimiento acerca de la fórmula φ del procesador j» o «el procesador i recibe una respuesta a un mensaje acerca de la fórmula φ del procesador j».

Para dotar de una semántica a L^D_n consideremos, a modo de explicación informal, un sistema distribuido con n agentes. Si el agente i toma como creencia propia a la fórmula φ influenciado por el hecho de que cree en ella el agente j, decimos entonces que i considera a j como

su *consejero* o *consultor* («adviser») respecto de dicha fórmula. El agente i puede tener otros consejeros aparte de j respecto a la fórmula φ, incluido él mismo (en el sentido de que i se basa igualmente en sus propias consideraciones para creer que φ). Además, i puede tomar a j como consejero para otras creencias aparte de la fórmula φ. Por esta razón se clasifican las fórmulas en diversos campos de conocimiento $\Psi_1, ..., \Psi_m$. Más aún, puede también que el agente i tome más en cuenta a un consejero que a otro dependiendo de tales campos de conocimiento. Por ello se impone el establecer una estructura jerárquica para los agentes respecto de sus consejeros y creencias. Aquí la naturaleza del problema es determinante. Es claro que, si j es matemático y k es abogado, i tendrá más en cuenta a j para un tema de matemáticas y a k para una cuestión legal.

Además, podemos establecer una función D_i que determine para cada fórmula cuál es el consejero más creíble para i respecto de dicha fórmula en el estado e. Así, $D_i(\varphi, e) = j$ significará que el agente j es el consejero más creíble (con mayor credibilidad) para i respecto de la fórmula j en el estado e. Admitiremos también un símbolo especial, «λ» con el significado de «nadie»; de forma que $D_i(\varphi, e) = \lambda$ quiere decir que el agente i no tiene agentes con credibilidad para tratar j en el estado e (ni siquiera confía en sí mismo).

Formalmente, un modelo de *dependencia de creencias* (para n agentes) es una tupla de la forma $(E, R_1, ..., R_n, D_1, ..., D_n, V)$, donde E es un conjunto no vacío de estados, cada R_i es una relación de accesibilidad (del agente i), cuyas propiedades variarán según modelemos conocimiento o creencia, cada D_i es una función de $E \times L^D{}_n$ en $\{1, ..., n, \lambda\}$ (cuyas propiedades pueden variar igualmente) y V es una función de valoración definida como es usual. La semántica de las conectivas clásicas booleanas y del operador de creencia implícita o conocimiento implícito son las acostumbradas. Respecto del operador $D_{i,j}$ tenemos:

$V(D_{i,j}\varphi, e) = 1$ si y sólo si $D_i(\varphi, e) = j$

¿Qué decir en este contexto de la conciencia? Podemos incorporar el operador de conciencia A_i definiendo:

$V(A_i\varphi, e) = 1$ si y sólo si $D_i(\varphi, e) \neq \lambda$

En efecto, $D_i(\varphi, e) \neq \lambda$, que equivale a $D_i(\varphi, e) = j_1$, o ..., o $D_i(\varphi, e) = j_n$, quiere decir que o bien el agente i se considera a sí mismo como su mejor consejero respecto de la fórmula φ o bien para algún agente j (diferente de i y de λ) éste es su consejero más creíble respecto de dicha fórmula. En el

$Ke_i\varphi \rightarrow A_iKe_i\varphi$
$-Ke_i\varphi \rightarrow A_i-Ke_i\varphi$.

La validez de estos axiomas requeriría presuponer que el agente es capaz de dar una respuesta definida (SI o NO) ante la pregunta de si sabe si tiene conocimiento algorítmico de la fórmula φ. Además, si los algoritmos que maneja el agente son correctos, entonces el operador Ke_i satisface: $Ke_i\varphi \rightarrow \varphi$ y $Ke_i\varphi \rightarrow K_i\varphi$. Más aún, la corrección del algoritmo es garantía de que el operador de conciencia satisfaga la definición ya conocida: $Ke_i\varphi \leftrightarrow (K_i\varphi \ \& \ A_i\varphi)$. En cambio, si el algoritmo usado además de correcto es completo, entonces se satisface el axioma $Ke_i\varphi \leftrightarrow K_i\varphi$ (el conocimiento implícito se reduce al explícito) (ver Fagin et al.(1995)).

Es frecuente en la vida real que las personas consulten la opinión de otras acerca de diversos temas, de manera que lo que sabe o cree un agente depende en buena medida de lo que aprende de otros (este fenómeno se llama «dependencia de creencias»). Un área de aplicación de este concepto son los *sistemas distribuidos*. Estos consisten en una colección de agentes (humanos, procesadores o robots) que, generalmente, cuentan con recursos limitados y se hallan conectados por una red de comunicación. En este tipo de entornos, el razonamiento acerca del conocimiento y la creencia ha producido variadas aplicaciones (en *las bases distribuidas de conocimiento*, en *la comunicación y cooperación para planificación multiagente en IA* o *en la ingeniería del conocimiento*).

En entornos distribuidos con dependencia de creencias encontramos en Huang (1990) un *modelo general de dependencia de creencias* para sistemas multiagentes basado en mundos posibles o estados. El lenguaje usado, que denotaremos L^D_n (para *n* agentes), contiene las conectivas clásicas y los operadores B_i (para la creencia implícita) y $D_{i,j}$ (para la dependencia de creencias). Este último operador se denomina *operador de dependencia*. Una expresión como $D_{i,j}\varphi$ se lee «el agente *i* depende de *j* para creer que φ». Otras lecturas posibles son las siguientes: «el agente *j* es el mejor consejero del agente *i* acerca de la fórmula φ», «el agente *i* pregunta al agente *j* acerca de la fórmula φ» o, en redes de procesos distribuidos, «el procesador *i* puede obtener conocimiento acerca de la fórmula φ del procesador *j*» o «el procesador *i* recibe una respuesta a un mensaje acerca de la fórmula φ del procesador *j*».

Para dotar de una semántica a L^D_n consideremos, a modo de explicación informal, un sistema distribuido con *n* agentes. Si el agente *i* toma como creencia propia a la fórmula φ influenciado por el hecho de que cree en ella el agente *j*, decimos entonces que *i* considera a *j* como

su *consejero* o *consultor* («adviser») respecto de dicha fórmula. El agente i puede tener otros consejeros aparte de j respecto a la fórmula φ, incluido él mismo (en el sentido de que i se basa igualmente en sus propias consideraciones para creer que φ). Además, i puede tomar a j como consejero para otras creencias aparte de la fórmula φ. Por esta razón se clasifican las fórmulas en diversos campos de conocimiento $\Psi_1, ..., \Psi_m$. Más aún, puede también que el agente i tome más en cuenta a un consejero que a otro dependiendo de tales campos de conocimiento. Por ello se impone el establecer una estructura jerárquica para los agentes respecto de sus consejeros y creencias. Aquí la naturaleza del problema es determinante. Es claro que, si j es matemático y k es abogado, i tendrá más en cuenta a j para un tema de matemáticas y a k para una cuestión legal.

Además, podemos establecer una función D_i que determine para cada fórmula cuál es el consejero más creíble para i respecto de dicha fórmula en el estado e. Así, $D_i(\varphi, e) = j$ significará que el agente j es el consejero más creíble (con mayor credibilidad) para i respecto de la fórmula j en el estado e. Admitiremos también un símbolo especial, «λ» con el significado de «nadie»; de forma que $D_i(\varphi, e) = \lambda$ quiere decir que el agente i no tiene agentes con credibilidad para tratar j en el estado e (ni siquiera confía en sí mismo).

Formalmente, un modelo de *dependencia de creencias* (para n agentes) es una tupla de la forma $(E, R_1, ..., R_n, D_1, ..., D_n, V)$, donde E es un conjunto no vacío de estados, cada R_i es una relación de accesibilidad (del agente i), cuyas propiedades variarán según modelemos conocimiento o creencia, cada D_i es una función de $E \times L^D{}_n$ en $\{1, ..., n, \lambda\}$ (cuyas propiedades pueden variar igualmente) y V es una función de valoración definida como es usual. La semántica de las conectivas clásicas booleanas y del operador de creencia implícita o conocimiento implícito son las acostumbradas. Respecto del operador $D_{i,j}$ tenemos:

$V(D_{i,j}\varphi, e) = 1$ si y sólo si $D_i(\varphi, e) = j$

¿Qué decir en este contexto de la conciencia? Podemos incorporar el operador de conciencia A_i definiendo:

$V(A_i\varphi, e) = 1$ si y sólo si $D_i(\varphi, e) \neq \lambda$

En efecto, $D_i(\varphi, e) \neq \lambda$, que equivale a $D_i(\varphi, e) = j_1$, o ..., o $D_i(\varphi, e) = j_n$, quiere decir que o bien el agente i se considera a sí mismo como su mejor consejero respecto de la fórmula φ o bien para algún agente j (diferente de i y de λ) éste es su consejero más creíble respecto de dicha fórmula. En el

primer caso, el agente es consciente de que j porque tiene una noción directa del asunto, en el segundo caso se refleja la situación de un agente que tiene noticia de que j a través de otro agente y por este motivo ya tiene conciencia de la cuestión. Por tanto, para *n* agentes se cumple lo siguiente:

$$A_i\varphi <-> (D_{i,j1}\varphi \text{ v } ... \text{ v } D_{i,jn}\varphi)$$

Asimismo, podemos introducir también el operador de creencia explícita (o del conocimiento explícito) y de nuevo se cumple que $Be_i\varphi <-> (B_i\varphi \& A_i\varphi)$.

Las propiedades de la conciencia varían según modelemos la función D_i. En Huang (1990) se presenta un sistema axiomático para tratar conocimiento (el sistema $S5D_n$) [NOTA 13], adecuado para tratar dependencia de creencias (de «conocimientos» más exactamente) en *entornos distribuidos sincronizados*. En los modelos de este sistema, la relación R_i es de equivalencia y la función D_i cumple una serie de propiedades que expresaremos como sigue [NOTA 14]:

1. Si $D_i(\varphi, e) = j$, entonces $D_j(\varphi, e) = j$.
2. D_i cumple las siguientes propiedades respecto de las conectivas booleanas – y &, y los operadores modales K_i y $D_{i,j}$:

   $D_i(\varphi, e) = D_i(-\varphi, e)$
   $D_i(\varphi \& \psi, e) = j$ si y sólo si $D_i(\varphi, e) = D_i(\psi, e) = j$
   $D_i(\varphi, e) = D_i(D_{i,j}\varphi, e)$, para un único j
   $D_i(\varphi, e) = D_i(K_j\varphi, e)$, para un único j.

3. Si $D_i(\varphi, e) = j$, entonces $D_i(\varphi, e') = j$ y $D_i(\varphi, e'') = j$, para estados cualesquiera e', e'' tales que eR_ie' y eR_je''.

La propiedad 1 induce una relación *secundariamente reflexiva* [NOTA 15], la condición 2 establece propiedades de cierre y sus direcciones opuestas, por ejemplo, la dependencia de creencias satisface el cierre bajo conjunción, $(D_{i,j}\varphi \& D_{i,j}\psi) -> D_{i,j}(\varphi \& \psi)$, así como la distribución en conjunción, $D_{i,j}(\varphi \& \psi) -> (D_{i,j}\varphi \& D_{i,j}\psi)$. Por último, la propiedad 3 indica una especie de «permanencia del consejero» a través de todos los estados que el consultado y el que consulta consideran como posibles.

Dadas estas propiedades, la conciencia satisface condiciones como las siguientes: la conciencia es generada por un conjunto dado de fór-

mulas atómicas primitivas (justificado por la condición 2 de la función D_i). También se cumplen:

$(A_i\varphi \,\&\, A_i(\varphi \rightarrow \psi)) \rightarrow A_i\psi$
$A_i\varphi \rightarrow K_iA_i\varphi$
$\neg A_i\varphi \rightarrow K_i\neg A_i\varphi$

La primera de estas fórmulas (cierre bajo implicación material) se justifica por la condición 2 y la definición de A_i en términos de $D_{i,j}$, las fórmulas restantes (propiedades de monotonía o conocimiento de lo que hay en la conciencia) se justifican fácilmente teniendo en cuenta la tercera propiedad de la función D_i.

En Huang y Kwast (1991) se comenta que este estilo de conciencia (definida en términos de dependencia de creencias) es una «conciencia indirecta», por oposición a la noción corriente de conciencia de Fagin y Halpern, a la que denominan «conciencia directa». Huang y Kwast lo plantean como una extensión de esta noción común de conciencia. Tomando un ejemplo de estos autores, si uno lee en un libro de zoología que «el conejo es un *oryctolagus caniculus*», uno puede creer tal cosa porque confía en lo que el autor dice, aunque no sea consciente de ello (porque no tiene noción de lo que significa *oryctolagus caniculus*). Huang y Kwast pretenden justificar intuitivamente que no ser consciente de algo no conduce necesariamente a carecer de una creencia explícita de ese asunto. Basta con tener noticia de alguien que tenga conciencia del asunto y nos lo comunique. La misión de los sistemas de dependencia de creencias sería, entonces, convertir la conciencia indirecta en directa.

Se da una situación en este planteamiento que merece un comentario. La definición del operador A_i en la semántica de dependencia de creencias señala que φ pertenece a $A_i(e)$ si y sólo si $D_i(\varphi, e) \neq \lambda$, con lo cual, en última instancia, $A_i(e)$ proporciona un conjunto de fórmulas de las que el agente puede ser finalmente consciente, aunque i no sea consciente de todo lo que hay en $A_i(e)$. La función A_i, en este caso, contrariamente al estilo de la LCG, proporciona fórmulas de las que el agente i puede no tener conciencia aunque sí la tengan otros agentes relacionados con i. Esto es, la función de conciencia definida para cada agente i proporciona, en cada estado e, un conjunto que podemos particionar en dos subconjuntos: el tradicional –al estilo de los modelos generales de conciencia- y otro, disjunto con el primero, que podemos interpretar como lo que i puede consultar a otros agentes del sistema en dicho estado aunque i no es consciente expresamente de los elementos de dicho conjunto.

X. Un operador dual del operador de conciencia

El planteamiento de Fagin y Halpern (1988) concibe la conciencia como un filtro que tiene como efecto eliminar ciertas fórmulas con objeto de evitar la OL. El procedimiento es simple: se define la función de conciencia de forma que dichas fórmulas no pertenezcan al conjunto de expresiones que forman parte de la conciencia del agente. Pero igualmente podríamos seguir una dirección opuesta a ésta, en el sentido de añadir fórmulas al conjunto de creencias ideales en vez de eliminarlas mediante un filtro. La intención de este proceder podría ser la de modelar un agente con una «confianza ciega» en sus creencias aunque sean inconsistentes. Este es precisamente el planteamiento de van der Hoek y Meyer (1989), quienes tratan esta cuestión sintácticamente, al estilo de la LCG. El operador introducido con este fin se denomina «principios» (*principles*), P_i, de forma que $P_i\varphi$ se puede leer «forma parte de los principios de i que φ» o simplemente «es un principio para el agente i que φ». El hecho de que algo sea un principio para un agente significa que se trata de una cuestión irrenunciable para él y esto trae como consecuencia una noción de creencia implícita más amplia que la considerada hasta ahora. De acuerdo con esta nueva noción, las creencias implícitas comprenden tanto las creencias implícitas según la antigua noción (la creencia racional) como creencias situadas «más allá de toda razón» o «fuera de toda discusión» (aunque no sea racional el sostenerlas). La irracionalidad de este tipo de creencias se comprueba cuando chocan con creencias a las que el agente se ve abocado a sostener racionalmente porque éstas se deducen lógicamente de lo que cree. Para esta nueva noción de creencia implícita usaremos el operador B^I_i, como hacen los autores mencionados.

El lenguaje usado en este planteamiento consta de las conectivas booleanas clásicas y los operadores B_i, B^I_i y P_i. En la semántica se utilizan modelos de la forma $(E, R_1, ..., R_n, P_1, ..., P_n, V)$, donde E es un conjunto de estados, R_i es la relación de accesibilidad del agente i (serial, transitiva y euclídea) y cada P_i es una función que asocia a cada estado de E un conjunto de fórmulas del lenguaje.

Las cláusulas semánticas para los operadores modales son las siguientes:

$V(B_i\varphi, e) = 1$ si y sólo si para todo e' tal que eR_ie', $V(\varphi, e') = 1$
$V(P_i\varphi, e) = 1$ si y sólo si $\varphi \in P_i(e)$
$V(B^I_i\varphi, e) = 1$ si y sólo si $\varphi \in P_i(e)$ o para todo e' tal que eR_ie', $V(\varphi, e') = 1$

Tenemos, pues, la definición:

$B^I_i\varphi <-> (B_i\varphi \vee P_i\varphi)$

(i.e. creer implícitamente que φ según la nueva noción significa que φ es una creencia racional o bien un principio).

Los agentes pueden tener opiniones contradictorias. Si consideramos expresiones como

$B^l_i(\varphi \ \& \ -\varphi)$
$B^l_i\varphi \ \& \ B^l_i{-}\varphi$

resulta que son satisfacibles, como lo es el conjunto

$\{B^l_i\varphi, B^l_i{-}\varphi\}$

No se cumplen las variadas formas de la OL para B^l_i. Pero las diferencias con el planteamiento de la conciencia son notorias. Por ejemplo, mientras que el planteamiento de la conciencia puede evitar que el agente crea que φ aunque φ sea una fórmula válida, el presente planteamiento puede hacer que el agente crea incluso que -φ en tal caso.

XI. Consideraciones finales

Hemos examinado el tema de la conciencia, una noción con tintes psicológicos, tal y como se trata en los formalismos lógicos usados en el razonamiento sobre conocimiento y creencia. En este sentido, el planteamiento inicial de Fagin y Halpern (1988), estableciendo la conciencia como una noción sintáctica, es muy flexible y su generalización por Thijsse (1992 y 1993) es de gran potencia expresiva. Este planteamiento es una forma de tratar el problema de la OL modelando agentes con recursos limitados, pero igualmente permite rescatar diversas formas de OL con la imposición de restricciones a la función conciencia. Además, la noción de conciencia posee diversas interpretaciones en distintas áreas del razonamiento sobre conocimiento y creencia, como el conocimiento algorítmico y la dependencia de creencias, que poseen variadas aplicaciones en el campo de la computación.

Notas

NOTA 1*

Una discusión de carácter filosófico y científico acerca de la conciencia y sus tipos puede encontrarse en Martínez-Freire 2000.

NOTA 2*

Usaremos las siguientes conectivas booleanas: - (negación), & (conjunción), v (disyunción inclusiva), -> (implicación material), < > (equivalencia material). El símbolo |= denota que la fórmula que sigue a su derecha es válida (en cierta clase de modelos previamente especificada). No es necesario que hagamos más precisiones sobre este símbolo.

NOTA 3*

La semántica de Kripke es estándar en el tratamiento de la lógica epistémica y doxástica. La idea fundamental en la que se basa es la de «mundo posible» que -en el presente contexto- podemos entender como un mundo o estado de cosas que el agente considera como posible. De lo que se trata con este planteamiento es de dar cuenta de expresiones como «el agente *i* sabe (cree) que φ». Intuitivamente podemos aclarar este punto como sigue. Un agente puede tener dudas acerca de la naturaleza del mundo real, de forma que considera alternativas diversas (distintos mundos al suyo). Si en todos los mundos que el agente considera posibles no tiene dudas acerca de que es el caso que φ, entonces sabe efectivamente que φ.

NOTA 4*

Para una introducción a las nociones de lógica modal, como la noción de validez o la distinción entre marco y modelo, puede consultarse Hughes y Cresswell 1996. Una fórmula es válida en un marco kripkeano si es verdadera en todo mundo posible. Hay marcos en los que la noción de válido se halla restringida a cierto tipo de mundos y que intentan así superar el problema de la OL. Véanse los planteamientos de *mundos no clásicos* de Cresswell 70, 72, 73, los *mundos imposibles* de Hintikka 1975 y Rantala 1982 o los *mundos no estándar* de Rescher y Brandon 1979.

NOTA 5*

Levesque distingue entre *situaciones parciales*, donde alguna fórmula atómica primitiva del lenguaje no es ni verdadera ni falsa (una descripción parcial de un mundo posible), *situaciones incoherentes* (en las que alguna fórmula atómica primitiva es simultáneamente verdadera y falsa) y *situaciones completas* (mundos posibles), que vienen definidas por la verdad o falsedad de cada fórmula atómica primitiva del lenguaje y, además, no son incoherentes.

NOTA 6*

En Konolige 1986 se critica las dos últimas lecturas de $A_i\varphi$ aduciendo que la noción de «verdadero» no parece muy apropiada en bases de conocimiento, por cuanto pueden contener información incorrecta o incompleta acerca del mundo que modelan.

NOTA 7*

Usaremos variantes de la notación original. Igualmente, en la formulación de los modelos a lo largo del artículo introduzco ciertas variaciones en la notación de los distintos autores, pero conservando una uniformidad.

NOTA 8*

Se dice que una relación R es *serial* si para todo x existe un y tal que xRy. R es *transitiva* si para cualesquiera x, y, z se tiene que si xRy e yRz, entonces xRz. R es *euclídea* si para cualesquiera x, y, z se tiene que si xRy y xRz,, entonces yRz.

NOTA 9*

Se dice que una relación R es de *equivalencia* si es reflexiva, simétrica y transitiva (alternativamente, si es reflexiva y euclídea). Las propiedades transitiva y euclídea ya se han formulado en la nota 5. En cuanto a las otras propiedades mencionadas, decimos que R es *reflexiva* si cumple que para cualquier x resulta xRx. R es *simétrica* si para cualesquiera x, y se tiene que si xRy, entonces yRx.

NOTA 10*

En *sistemas multiagentes* (una de las áreas más importantes de aplicación del razonamiento sobre el conocimiento y la creencia) se distingue entre el «estado local» de un agente y el «estado global» (del sistema) -el cual comprende los estados locales de una serie de agentes que interactúan y de lo que se llama «el entorno». Intuitivamente, si imaginamos un juego de cartas, el «estado local» de un jugador podría ser las cartas que posee, las que pueda ver arrojadas por los otros jugadores así como cualquier información sobre la estrategia de juego de éstos. El «entorno» es un concepto ambiguo, a veces es un agente. Se considera cualquier aspecto que pueda ser relevante no comprendido por los estados locales de los agentes (en sistemas de comunicación podría contener los mensajes en tránsito, las líneas de comunicación abiertas o cerradas, etc.). Hay un número dado de estados globales iniciales posibles (en el juego de cartas los posibles repartos de manos y cada estado local es la primera mano de cada jugador).

NOTA 11*

Si en lugar de un operador de creencia, usamos uno de conocimiento, la regla se denomina de «generalización epistémica». Por otro lado, el símbolo |- denota que la fórmula que sigue a su derecha es un teorema del sistema.

NOTA 12*

El sistema K se define mediante **Prop**, **K**, Generalización epistémica y *modus ponens*. Extensiones interesantes de K para el conocimiento y creencia son los sistemas KT, KT4 y KT5 (como el propio nombre indica, KT resulta de añadir el axioma **T** al sistema K, KT4 añade **T** y **4** a K, KT5 añade **T** y **5** a K). También podemos denominar a estos sistemas respectivamente T, S4 y S5 simplemente. Si deseamos resaltar que consideramos n agentes (y no un solo agente) podemos añadir a los nombres índices suscritos y denominamos a los sistemas mencionados K_n, T_n, $S4_n$ y $S5_n$.

NOTA 13*

El sistema $S5D_n$ es una extensión de S5 (para el operador K_i) junto con los axiomas siguientes:

$D_{i,j}\varphi <-> D_{i,j}\text{-}\varphi$
$D_{i,j}(\varphi \,\&\, \psi) <-> D_{i,j}\varphi \,\&\, D_{i,j}\psi$
$D_{i,j}\varphi <-> D_{i,j}D_{i,j}\varphi$
$D_{i,j}\varphi <-> D_{i,j}K_j\varphi$
$D_{i,j}\varphi -> K_iD_{i,j}\varphi$
$D_{i,j}\varphi -> K_jD_{i,j}\varphi$
$D_{i,j}\varphi -> D_{j,j}\varphi$
$D_{i,j}\varphi -> \text{-}D_{i,k}\varphi$ (con $k \neq j$)

Este sistema es correcto y completo respecto de su semántica (ver Huang 1990).

NOTA 14*

Aunque la forma en que se exponen aquí las condiciones de la función D_i difiere de la de Huang y Kwast, sin embargo, en el fondo viene a ser lo mismo.

NOTA 15*

Se dice que una relación R es *secundariamente reflexiva* si para cualesquiera x e y se tiene que si xRy, entonces yRy.

Bibliografía

CRESSWELL, M. J. 1970: «Classical intensional logics», en *Theoria 36*, pp. 347-372.

CRESSWELL,M. J. 1972: «Intensional logics and logical truth», en *Journal of Philosophical logic 1*, pp.2-15.

CRESSWELL, M. J. 1973: *Logics and Languages*, Methuen and Co., London.

FAGIN, R. y HALPERN, J.Y. 1988: «Belief, awareness and Limited Reasoning», en *Artificial Intelligence 34*, pp. 39-76.

FAGIN, R., HALPERN, J.Y., MOSES, Y, y VARDI, M. Y. 1995: *Reasoning about Knowledge*, The MIT Press.

HALPERN, J.Y. 1987: «Using reasoning about knowledge to analyze distributed systems», en J.F. Traub, B.J. Grosz, B.W. Lampson y N.J. Nilsson (Eds.), *Annual Review of Computer Science*, vol. 2, Annual Reviews Inc., Palo Alto, California, pp. 37-68.

HALPERN, J.Y. 2000: «Alternative semantics for unawareness», documento en http://www.es.cornell.edu/home/halpern.

HINTIKKA, J.1962: *Knowledege and Belief*, Cornell University. Hay trad.esp. de Juan José Acero: *Saber y creer*, Tecnos, Madrid, 1979.

HINTIKKA, J. 1975: «Impossible possible worlds vindicated», en *Journal of Philosophical Logic 4*, pp. 475-484.

HUANG, Z.1990: «Dependency of belief in distributed systems», en M. Stokhof y L. Torenvliet (Eds.), *Proceedings of the 7th Amsterdam Colloquium*, ITLI, University of Amsterdam, pp. 637-662.

HUANG, Z.y KWAST, K. 1991: «Awareness negation and logical omniscience», en J. Van Eijck (Ed.), Logics in AI (*Proceedings* JELIA'90), *Lectures Notes in Computer Science* 478, Springer, pp. 282-300.

HUGHES, G.E. y CRESSWELL, M. J. 1996: *A New Introduction to Modal Logic*, Routledge, London y New York.

KONOLIGE, K. 1984: *A Deduction Model of Belief and its Logics*, Tesis doctoral, Universidad de Stanford, Departamento de Ciencias de la computación, Stanford, California.

KONOLIGE, K. 1985: «Belief and incompleteness», en J.R. Hobbes and R.C. Moore (Eds.), *Formal Theories of the Commonsense World*, Ablex Publishing Company, pp. 359-404.

KONOLIGE, K. 1986: «What awareness isn't: a sentential view of implicit and explicit belief», en J. Y. Halpern (Ed.) *Theoretical Aspects of Reasoning about Knowledge: proceedings of the 1986 Conference*, Morgan Kaufmann, Los Altos, California, pp. 241-250

LEVESQUE, H.J. 1984: «A logic of implicit and explicit belief», en *Proceedings of National Conference on Artificial Intelligence (AAAI'84)*, pp. 198-202.

MARTÍNEZ-FREIRE, P.F. 2000: «Aproximaciones científicas al problema de la conciencia», en Ildefonso Murillo (Ed.), Fronteras de la filosofía de cara al siglo XXI, *Diálogo filosófico*, pp. 181-197.

MOSES, Y. 1988: «Resource-bounded knowledge», en M. Vardi (Ed.), *Proceedings of the Second Conference on Theoretical Aspects of Reasoning about Knowledge*, Morgan Kaufmann, San Francisco, California, pp. 261-276.

RANTALA, V. 1982: «Impossible worlds semantics and logical omniscience», en *Acta Philosophica Fennica 35*, pp. 18-24.

RESCHER, N. 1969: *Topics on Philosophical Logic*, Dordrecht.

RESCHER, N. y BRANDOM, R. 1979: *The Logic of Inconsistency*, Totowa, N.J.: Rowman and Littlefield.

THIJSSE, E. 1992: *On Partial Logic and Knowledge Representation*. Netherlands: Eburon, Delft (Tesis doctoral, Universidad de Tilburg).

THIJSSE, E. 1993: «On total awareness logics», en M. De Rijke (Ed.), Diamonds and Defaults, pp. 309-347, Dordrecht, Kluwer.

ULE, A. 2000: «Awareness as a logical operator», documento en http://ciiiweb.ijs.si/dialogues/r-ule.htm.

VARDI, M.Y. 1986: «On epistemic logic and logical omniscience», en J.Y. Halpern (Ed.), *Theoretical Aspects of Reasoning about Knowledge: Proceedings of the 1986 Conference*, Morgan Kaufmann, Los Altos, California, pp. 293-306.

VAN DER HOEK, W. y MEYER, J. J. Ch. 1989: «Possible logics for belief», en *Logique et Analyse* 127-128, pp.177-194.

WRIGHT, G.H.von. 1951: *An Essay in Modal Logic*, North-Holland, Amsterdam.