UNIVERSIDAD NACIONAL AUTÓNOMA DE MÉXICO

POSGRADO EN FILOSOFÍA DE LA CIENCIA

INSTITUTO DE INVESTIGACIONES FILOSÓFICAS

FACULTAD DE FILOSOFÍA Y LETRAS

FACULTAD DE CIENCIAS

DIRECCIÓN GENERAL DE DIVULGACIÓN DE LA CIENCIA

INTENCIONES: REVISIÓN Y NO-MONOTONICIDAD

TESIS

QUE PARA OPTAR POR EL GRADO DE:

DOCTOR EN FILOSOFÍA DE LA CIENCIA (FILOSOFÍA DE LAS CIENCIAS COGNITIVAS)

PRESENTA:

JOSÉ MARTÍN CASTRO MANZANO

TUTOR: DR. AXEL ARTURO BARCELÓ ASPEITIA INSTITUTO DE INVESTIGACIONES FILOSÓFICAS
COMITÉ: DR. JESÚS RAYMUNDO MORADO ESTRADA INSTITUTO DE INVESTIGACIONES FILOSÓFICAS
DR. FRANCISCO HERNÁNDEZ QUIROZ FACULTAD DE CIENCIAS
DR. ALEJANDRO GUERRA HERNÁNDEZ POSGRADO EN FILOSOFÍA DE LA CIENCIA
DR. MAURICIO OSORIO GALINDO POSGRADO EN FILOSOFÍA DE LA CIENCIA

MÉXICO, D.F. JUNIO, 2014.

But Mousie, thou art no thy lane,
In proving foresight may be vain:
The best-laid schemes o' mice an' men
Gang aft agley,
An' lea'e us nought but grief an' pain,
For promis'd joy!
Robert Burns, To a mouse

Filosofía exacta: La filosofía construida con la ayuda de herramientas lógicas y matemáticas
Mario Bunge, Diccionario de filosofía

En aquel Imperio, el Arte de la Cartografía logró tal Perfección que el Mapa de una sola Provincia ocupaba toda una Ciudad, y el Mapa del Imperio, toda una Provincia. Con el tiempo, estos Mapas Desmesurados no satisficieron y los Colegios de Cartógrafos levantaron un Mapa del Imperio, que tenía el Tamaño del Imperio y coincidía puntualmente con él. Menos Adictas al Estudio de la Cartografía, las Generaciones Siguientes entendieron que ese dilatado Mapa era Inútil y no sin Impiedad lo entregaron a las Inclemencias del Sol y los Inviernos. En los Desiertos del Oeste perduran despedazadas Ruinas del Mapa, habitadas por Animales y por Mendigos; en todo el País no hay otra reliquia de las Disciplinas Geográficas
Jorge Luis Borges, Del rigor en la ciencia

There is no mathematical substitute for philosophy
Saul Kripke, Is there a problem about substitutional quantification?

Índice general

Índice de figuras

Índice de cuadros

Índice de algoritmos

Prefacio

Utrum sit logicae philosophicae, nec ne

El extracto del poema de Burns con el que hemos abierto este trabajo resume, de manera exquisita, la motivación profunda de la cual brota esta investigación. Siguiendo la interpretación que hace McGown en su *Primer* [114], con algunas reservas,[1] podríamos explicar por qué hemos elegido iniciar con ese fragmento a costa de eliminar su dulzura.

Los ratones no están solos, pero la simpatía hacia ellos no proviene de una simpatía general hacia la naturaleza y sus creaturas, sino del hecho de que algunos de sus problemas también son nuestros problemas. Del hecho, más universal, de que no son los únicos seres cuyos proyectos pueden fallar, pues resulta que en este mundo hasta los planes más detallados y cuidados, más meditados y calculados pueden, por cualquier circunstancia, tener resultados diferentes de los esperados, dejándonos en la incertidumbre y el dolor.

Por supuesto, que la angustia generada por esta realidad es directamente proporcional a la diferencia de los resultados finales menos los resultados obtenidos no es una revelación. Afirmar que nuestros mejores planes pueden fallar es una perogrullada; lo que probablemente ya no es tan obvio ni evidente es *cómo* es que somos, quizás más que los propios ratones, *persistentes* ante los fallos y las circunstancias adversas.

[1] McGown, hablando de este poema, hace una serie de comparaciones con otros poetas bajo la hipótesis de que éste tiene como motivo la simpatía hacia los animales y la naturaleza dada una anécdota que dice que, en 1841 en Kilmarnock, vivía un sirviente de Burns llamado John Blane. McGown escribe que John cuenta que 55 años antes—esto es noviembre de 1785, año en que el poema está fechado—un niño perseguía un ratón por el campo, armado con un arado (*plough*). Burns, que estaba arando cerca de ahí increpó al muchacho para que dejara al ratón en paz. Blane relata que el resto del día Burns estuvo pensativo. Luego de que el sirviente se había preparado para dormir el poeta se quedó meditando. Poco tiempo después, despertó a su compañero para mostrarle su nueva producción, como Minerva de la cabeza de Zeus.

El problema de la persistencia, por tanto, no es nuevo ni exclusivo *De Ratones y Hombres*. Es, además, el problema que subyace en preguntas típicas como las siguientes: ¿Cómo es posible que un adulto recuerde su infancia si, con el paso del tiempo, ha perdido tantas células cerebrales? ¿Cómo es posible que un correo electrónico llegue en tiempo y forma a su destino si de México va a Japón, de Japón va a Australia y de Australia llega a su destino en México?

Estas preguntas se pueden precisar un poco más para proveer la ejemplaridad adecuada: ¿Cómo es que el funcionamiento del cerebro es persistente a pesar de la constante pérdida de neuronas? ¿Cómo es que un paquete de información, en el protocolo ITP, es persistente a pesar de la pérdida de datos en la red mundial? La filosofía y la inteligencia artificial no han sido ajenas a este tipo de pesquisas. Centenas de trabajos examinan, desde un punto de vista lógico, la persistencia de nuestras creencias ante información adversa al responder a la pregunta: ¿Cómo es que algunas de nuestras creencias permanecen a pesar de cambios súbitos de información?

Algunos grupos de investigadores, tanto de filosofía como de inteligencia artificial, de la segunda mitad del siglo XX en adelante, se han enfrentado a semejante dificultad a través de estudios de carácter primariamente lógico. Esto ha sido así porque la lógica es un tópico central tanto de la filosofía como de la inteligencia artificial, puesto que el comportamiento lógico es una fracción fundamental de la inteligencia y la inteligencia constituye un problema central, histórica y tópicamente, para estas dos disciplinas.

Aquí es necesario hacer una aclaración. Si bien el término *lógica* suele evocar el uso y estudio de sistemas lógicos que representan creencias o conocimiento, el comportamiento lógico no está conformado sólo por creencias y conocimiento, dado que no sólo razonamos usando creencias: también lo hacemos usando intenciones—incluso deseos—de manera cotidiana. La pregunta obligada en este momento es, entonces, ¿por qué este tipo de razonamiento, que aquí llamamos razonamiento intencional, no ha recibido la misma atención, desde el punto de vista de la lógica, que el razonamiento sobre creencias?

Aunque en su momento nos aventuraremos a sugerir una hipótesis para explicar esta situación, lo que nos interesa no es la investigación sobre el porqué de tal estado de cosas: aunque, sin duda, sería muy interesante, no haremos un estudio de perfil historiográfico. Tampoco discutiremos, por cierto, la naturaleza de las intenciones, aunque ciertamente dedicaremos buena parte del trabajo a hablar de ellas: no haremos filosofía de la mente o de la acción. Lo que realmente nos atrae es la búsqueda de ciertos mecanismos y patrones lógicos subyacentes a un tipo de razonamiento persistente, robusto, que conlleva creencias e intenciones

al interior de un modelo cognitivo. Lo que pretendemos es, por tanto, mostrar que el razonamiento intencional es un tipo de razonamiento lógico legítimo mediante una exploración de carácter primariamente lógico. En este sentido es muy pertinente la cita de Bunge con la que también hemos iniciado.

Para lograr nuestro propósito hacemos algo relativamente simple. Buscamos razones para justificar que el razonamiento intencional es razonamiento lógico *bona fide*. Usamos convenientemente ciertas relaciones entre filosofía e inteligencia artificial mediante aproximaciones lógicas y conceptuales. Desarrollamos un modelo para que tenga alguna repercusión práctica. Y si acaso logramos algo más allá de esto, estaremos complacidos; pero si logramos nuestros objetivos, sin duda nos sentiremos afortunados.

Antes de pasar a la descripción de los detalles procedimentales sobre la estructura expositiva de este trabajo nos gustaría hacer una aclaración y una distinción. En primer lugar, como cualquier otra investigación, la nuestra no sólo tiene un sesgo particular, sino que además constituye un modelo, con todo lo que ello implica. En consecuencia, preservamos cierto grado de coherencia interna con todo y la pérdida de la riqueza total del fenómeno a modelar: aquí la cita de Borges resulta más que iluminadora. A cambio, obtenemos una investigación extensible y aplicable: extensible, porque algunos parámetros de investigación tendrán que desarrollarse en trabajos posteriores; aplicable, porque el modelo mismo está vinculado con algunas aplicaciones y, en este aspecto, la cita de Kripke es alentadora.

Finalmente, antes de seguir la tradición de responsabilizar al autor principal de todo crimen intelectual, me gustaría desligarme un momento del *pluralis modestiae* para agradecer a mis asesores—y a la larga lista de árbitros anónimos—por tratar de convencerme y educarme para no cometer tales delitos. No sólo soy el responsable de cualquier crimen, también soy culpable de ser el responsable.

Características del trabajo

Tratando de presentar nuestros resultados de un modo autocontenido hemos diseñado la exposición de tal manera que sea, en la medida de lo posible, amigable. Así, cada capítulo viene acompañado de una presentación y de un resumen; además hacemos uso constante de cuadros, algoritmos, figuras y notas a pie de página cuando consideramos que es pertinente. Y a riesgo de parecer monótonos, repetimos constantemente algunas ideas específicas a lo largo de todo el trabajo.

Presentación

La presentación que acompaña a cada capítulo pretende anticipar los contenidos del mismo, de tal modo que se sepa qué esperar de él. En particular, la presentación indica cuál es el tema principal del capítulo, para qué se ha desarrollado y cómo es que se organiza; en otros términos, su tema, su motivación y su estructura.

Cuadros, algoritmos, figuras y notas

Además hemos incluido cuadros, algoritmos, figuras y notas a pie de página con el objetivo de facilitar la difusión del contenido. Buena parte de este trabajo es visual y divulgativa.

Resumen

Al finalizar cada capítulo presentamos un resumen con la intención de facilitar la transición al siguiente y preservar cierto grado de coherencia narrativa.

Agradecimientos

Nos gustaría agradecer a todos los que hicieron posible el inicio, desarrollo y fin de este trabajo. Desde aquellos que nos ofrecieron alojo durante ratos de emergencia hasta aquellos que nos facilitaron breves momentos de amistad en el transporte público. En especial queremos agradecer a los que nos aguantaron amorosamente tantos años: mi familia, Irma, Numa, Omar, Verónica, Yolanda y Zandra. Al comité constituido por el Dr. Axel Barceló Aspeitia,[2] el Dr. Alejandro Guerra Hernández,[3] el Dr. Raymundo Morado Estrada,[4] el Dr. Mauricio Osorio Galindo[5] y el Dr. Francisco Hernández Quiroz.[6] A la *Triple A* (AAA), presente desde la licenciatura, constituida por el Mtro. Ariel Campirán Salazar,[7] el Dr. Alejandro Guerra Hernández y el Dr. Axel Barceló Aspeitia. A la Coordinación del Posgrado en Filosofía de la Ciencia: Noemí y Elizabeth. A la Beca CONACyT

[2]Instituto de Investigaciones Filosóficas, Universidad Nacional Autónoma de México.
[3]Departamento de Inteligencia Artificial, Universidad Veracruzana.
[4]Instituto de Investigaciones Filosóficas, Universidad Nacional Autónoma de México.
[5]Departamento de Actuaría, Física y Matemáticas, Universidad de las Américas Puebla.
[6]Facultad de Ciencias, Universidad Nacional Autónoma de México.
[7]Facultad de Filosofía, Universidad Veracruzana.

y a la excelente organización del Área de Becas CONACyT de la Facultad de Filosofía y Letras: Gabriel Ramos. A la larga lista de inefables árbitros anónimos. Y dado que este trabajo fue elaborado usando LaTeX 2_ε, agradecemos también a Donald Knuth y Leslie Lamport; a Paul Taylor y a Christophe Fiorio por los paquetes `diagrams.sty` y `algorithm2e.sty`, respectivamente.

Capítulo 1

Introducción

1.1. Presentación

Las relaciones entre filosofía e inteligencia artificial son profundas en el espacio y amplias en el tiempo, y sobre todo, son de un carácter especial y, hasta cierto punto, único.[1] No sólo es así porque estas disciplinas comparten un origen común compuesto, entre otras cosas, de motivos históricos—desde el *Órganon* de Aristóteles[2]—y anécdotas interesantes—como la correspondencia entre Simon y Russell[3]—, sino porque del constante intercambio entre éstas ganamos hipótesis

[1]Esta relación es de una naturaleza *sui generis* por dos razones sustantivas. Por un lado, es posible utilizar técnicas, modelos y otros derivados de las ciencias computacionales y la inteligencia artificial para tratar problemas filosóficos de manera legítima [75]; por otro, las mismas ciencias computacionales generan problemas que son filosóficos porque se conectan directamente con la ontología, la ética, la epistemología y, por supuesto, la lógica [60, 61, 158].

[2]El rastreo de estos rasgos genéticos se hace más interesante si consideramos referencias en otros lugares y tiempos, por ejemplo, con Ramon Llull y su *Ars magna* [99]; Descartes y su carta a Beeckman para hablar de la necesidad de un lenguaje universal [99]; y por supuesto, con Leibniz y su célebre *Dissertatio de arte combinatoria* [44, 102].

[3]Herbert Simon—uno de los padres de la inteligencia artificial junto con Shaw y Newell por un lado, y McCarthy y Minsky por otro—le informa en una carta a Bertrand Russell, fechada el 21 de septiembre de 1957, el éxito de lo que se convertiría en el *GPS—General Problem Solver* [121]—en la tarea de demostrar teoremas de *Principia Mathematica*. La respuesta de Russell es más que interesante [57]:

> "Muchas gracias por su carta y por el trabajo adjunto. Me encanta el ejemplo de que su máquina es superior a Whitehead y a mí. Entiendo muy bien las razones por las cuales piensa que estos datos no deben llegar a oídos de los escolares. ¿Cómo podríamos pretender que aprendan a sumar si supieran que las máquinas pueden ejecutar la operación mejor que ellos? También me fascina que haya demostrado con tanta precisión la veracidad del viejo adagio que dice que no es lo mismo la sabiduría que erudición".

útiles, métodos formales y análisis funcionales que pueden arrojar luz sobre diferentes aspectos de la naturaleza de la conducta humana, particularmente bajo esquemas cognitivos.

Partiendo de la relación entre estas dos disciplinas el esquema cognitivo que asumimos es el modelo BDI (por *Beliefs*, *Desires* e *Intentions*) como fue originalmente expuesto por Michael Bratman en *Intention, plans, and practical reason* [27] y formalmente desarrollado por la escuela australiana de Anand Rao y Michael Georgeff [134, 136].

Bajo este modelo, el fenómeno que nos interesa es el del razonamiento intencional. De este tipo de razonamiento nos interesan dos atributos: su no-monotonicidad y su capacidad de revisión.

Así, aunque probablemente en este punto no es del todo claro, diremos, como la tradición aristotélica manda, cuál es la materia y el objetivo de este trabajo [8]: la materia es un tipo de razonamiento que nosotros llamamos *intencional*; la meta es la comprensión, con miras a la reproducción, de los atributos lógicos de no-monotonicidad y revisión de tal tipo de razonamiento.

Qué sea el razonamiento intencional, o al menos, qué entendemos por *razonamiento intencional* en esta investigación es algo que definiremos con precisión más adelante. Por el momento basta decir que es un tipo de razonamiento que todos usamos, que todos asumimos y, aunque es importante para la filosofía—y para las ciencias computacionales—, no ha sido estudiado con la misma profundidad que el razonamiento con creencias.

Antes de proporcionar algún ejemplo que ilustre este tipo de razonamiento diremos algo acerca de la situación descrita en el párrafo anterior y que se relaciona con la relevancia de nuestro problema y nuestra investigación: si es verdad que el razonamiento intencional—nuestro objeto de investigación—está tan presente en la vida diaria, es tan común y todos lo usamos, ¿por qué no ha sido estudiado con la misma profundidad que el razonamiento con creencias? Tal vez lo más sencillo y menos prudente sea apuntar y disparar a Platón,[4] pero quizás haya una

---

[4]Si, por un momento, interpretamos el *Elephants don't play chess* de Brooks [33] como un fragmento de epistemología y no como un manifiesto de la *nouvelle AI* contra la *Good Old Fashioned AI* [85], podríamos argumentar que si la inteligencia está en la interacción directa con el mundo y no sólo en las representaciones simbólicas de él, al desenterrar el cuerpo que Platón había enterrado—y con él las cuestiones ajenas al conocimiento—podríamos hallar una hipótesis explicativa de por qué el razonamiento intencional no ha sido estudiado como una forma de razonamiento *bona fide* en términos puramente lógicos. Una hipótesis similar puede ser extraída de la crítica de Dreyfuss a la inteligencia artificial simbólica [53].

explicación más simple, menos comprometedora y con tanto o más sentido: el extraño caso de la visión por computadora.

Cuando la inteligencia artificial (IA, de ahora en adelante) empezó a constituirse como disciplina científica, 1956 *circa* [109], se ocupaba de problemas que parecían, por su dificultad, requerir altos grados de inteligencia. Así, siguiendo este paradigma, en un principio la IA se encargaba fundamentalmente del razonamiento simbólico[5] y de la capacidad para jugar ajedrez con maestría, de cómo ganar en las damas o demostrar teoremas de geometría [34, 43, 123, 141], pero poco se hacía por estudiar un aspecto que posibilitaba la comprensión de todas estas actividades: la visión.

La explicación de este estado de cosas no parece ser tan misteriosa. Para la IA la visión era tan cercana, que no aparecía como un problema en el horizonte de investigación, pues era realmente la condición que posibilitaba el horizonte mismo. No fue sino hasta hace poco, en términos de años científicos, que la visión por computadora se convirtió en una disciplina autónoma de la IA, con sus propios avances y derrotas [94].

El relato anterior nos sirve para explicar, de manera análoga, por qué el fragmento I (por *Intentions*) del modelo BDI no es tan estudiado como lo es, en este caso, el B (por *Beliefs*). Sin duda, no nos faltan ganas de culpar a Platón de nuevo—alguien debe ser culpable. Pero la explicación puede ser más simple: el razonamiento intencional lo llevamos a cabo de manera diaria, tan naturalmente, que no se nos presenta como un problema más en el horizonte y es, en el mejor de los casos, una condición para problematizar otros aspectos del razonamiento. Sin embargo, como ha sucedido con la visión por computadora, el estudio del razonamiento intencional se está haciendo presente con cierta autonomía, y por tanto, se las está viendo con sus propios problemas y desarrollos, con sus propios éxitos y fracasos.

Habiendo explicado el estado de cosas del que partimos, en este momento un ejemplo es pertinente para explicar lo que pretendemos hacer con más detalle. Imaginemos que un individuo α tiene la intención de obtener su doctorado en

[5] El mayor exponente de este programa de investigación, posteriormente conocido como *GOFAI* [85], se encuentra en el modelo teórico construido bajo la hipótesis del sistema físico de símbolos de Newell y Simon que sugiere que un sistema físico de símbolos tiene los medios necesarios y suficientes para la acción inteligente [122]:

> "A physical symbol system has the necessary and sufficient means for general intelligent action. By necessary, we mean that any system that exhibits general intelligence will prove upon analysis to be a physical symbol system. By sufficient we mean that any physical symbol system of sufficient size can be organized further to exhibit general intelligence".

filosofía, ϕ. En principio podemos identificar tres problemas y formar, en consecuencia, tres preguntas:

1. ¿Qué significa que α intente ϕ?
2. ¿Qué razones puede tener α para intentar ϕ?
3. ¿Bajo qué condiciones diríamos que es racional para α intentar ϕ?

Estas tres preguntas, así formuladas, definen tres programas de investigación con respecto a tres problemas [105]:

1. El *problema de la acción* que requiere la caracterización precisa de la transformación de una intención en una acción; esto es, precisa un modelo que explique la transición de un estado mental, en este caso la intención, a una acción.
2. El *problema de la generación* que necesita proveer una explicación de cómo es que se generan las intenciones.
3. El *problema de la persistencia* que necesita un modelo de las condiciones bajo las cuales las intenciones se mantienen o se modifican durante el tiempo.

Esta investigación tiene referencias a los problemas 1 y 2 pero se centra en el problema 3, puesto que estamos interesados en los atributos lógicos de no-monotonicidad y revisión del razonamiento intencional.

La relación entre el problema de la persistencia y los atributos lógicos de no-monotonicidad y revisión del razonamiento intencional pueden caracterizarse con más precisión en dos subproblemas del programa de investigación 3 a los que dedicaremos nuestra atención.

Sabemos que el problema de la persistencia no es nuevo, pero se ha concentrado en el fragmento B y no ha puesto atención suficiente en otras estructuras cognitivas que también son importantes para la comprensión y la reproducción del comportamiento inteligente. Esto lo podemos justificar observando la diferencia de sistemas lógicos desarrollados en torno a B frente a la cantidad de sistemas lógicos en torno a I. No obstante, es razonable suponer que el fragmento I es lógicamente interesante porque el comportamiento lógico no está restringido a las creencias y además, como discutiremos más adeltante, lo incluye.

La discusión generada por el Dilema de Jörgensen[6], para empezar, constituye un argumento *knock-down* para justificar la razonabilidad de la suposición de que el razonamiento lógico no está restringido al fragmento B. Pero no sólo no está restringido a B, sino que incluye a I, por ejemplo, cuando llevamos a cabo razonamientos prácticos.

Para apoyar la suposición de que lo incluye podemos considerar las teorías filosóficas y formales existentes que tratan la naturaleza de las intenciones y el razonamiento intencional [27, 30, 41, 52, 100, 137]—incluso con modelos experimentales [97, 164, 143]—; sin embargo, de estas únicamente algunas consideran la posibilidad de una exploración lógica legítima y explícita [72, 86, 126]. Por tanto, aunque el estudio del razonamiento intencional bajo el problema de la persistencia no es completamente nuevo, la novedad aparece cuando nos preguntamos por los atributos de no-monotonicidad y revisión del razonamiento intencional: si las hay, ¿cuáles son y cómo se comportan las características de las intenciones que permiten un mecanismo de inferencia? Análogamente, si las hay, ¿cuáles son y cómo se comportan las características de las intenciones que permiten un mecanismo de revisión?

Buscando responder preguntas similares a éstas se han desarrollado interpretaciones del mecanismo de revisión [86, 126] y sistemas lógicos [136, 165] para capturar y entender la naturaleza o aspecto material del razonamiento intencional. Los resultados, como veremos, son impresionantes; sin embargo, el problema con estos enfoques es que el razonamiento intencional no es monotónico, y por tanto, los modelos lógicos deberían respetar algunas cláusulas de no-monotonicidad.

Por otro lado tenemos el caso de las lógicas derrotables y las representaciones de inferencias no-monotónicas [133] que tratan de sujetar y explicar el estatus o aspecto formal del razonamiento derrotable. De manera similar, los resultados han sido más que sólo prometedores; no obstante, el problema aquí es que los sistemas que son no-monotónicos no son intencionales.

Esta situación conforma un problema fundamental, puesto que los sistemas formales tradicionales para estudiar el razonamiento intencional pretenden estar

[6]El Dilema de Jörgensen puede formularse así: supongamos que las proposiciones sobre normas u obligaciones no son proposiciones legítimas porque no se les puede adscribir un valor de verdad. Si esto es así, no pueden existir lógicas de proposiciones normativas, a saber, lógicas deónticas o imperativas. Sin embargo, si la lógica es acerca de relaciones inferenciales y podemos tener tales relaciones entre normas, las lógicas deónticas y las imperativas no sólo resultan posibles, sino efectivas. Por tanto, o bien las normas no son proposiciones legítimas, y entonces no hay lógicas deónticas; o bien hay lógicas deónticas, y entonces la adscripción de verdad no es un criterio de suficiencia para definir a las proposiciones [95].

basados en lo que denominamos *modelo bratmaniano*. Un modelo bratmaniano, como explicaremos más adelante, implica ciertas características de intencionalidad, temporalidad y no-monotonicidad simultáneamente. Pero los sistemas previamente mencionados no consideran estos aspectos simultáneamente. De manera más específica, estos sistemas, o bien definen procesos no-monotónicos pero no son intencional-temporales o bien son intencional-temporales pero monotónicos. Preciso es, después de explorar las características del fragmento I, desarrollar un modelo formal que represente, en la medida de lo posible, los atributos de no-monotonicidad y revisión propios de un modelo bratmaniano.

Para resolver este problema hacemos una reducción a dos conjuntos de problemas elementales. El primer conjunto puede identificarse mediante el siguiente par de preguntas:

- ¿Hay características en las intenciones que permitan un mecanismo de revisión?
- Si la respuesta es afirmativa, ¿cómo se comporta este mecanismo?

El segundo conjunto puede identificarse con el siguiente par de preguntas:

- ¿Hay características en las intenciones que permitan un mecanismo de inferencia?
- Si la respuesta es afirmativa, ¿cómo se comporta este mecanismo?

Para asegurar cierta precisión en el uso de las preguntas anteriores hemos clasificado estos problemas de la siguiente manera:

- *Problema externo*: el estudio sobre el cambio de estructuras de datos se ha concentrado tradicionalmente en el cambio racional de creencias (el fragmento B) pero no se ha fijado en otras estructuras que también padecen modificaciones y que también son importantes para la comprensión y la reproducción de la conducta inteligente. Ante esta situación, en el marco de un *modelo bratmaniano*, nos preguntamos si las intenciones poseen propiedades que justifican una noción mínima de revisión; y si la respuesta es afirmativa, cómo podemos modelar el comportamiento de estos mecanismos de revisión.

- *Problema interno*: los sistemas formales tradicionales para estudiar el razonamiento intencional están basados en un *modelo bratmaniano*. Un modelo bratmaniano implica ciertas características de intencionalidad, temporalidad y no-monotonicidad. Pero los sistemas tradicionales no consideran los aspectos temporales y no-monotónicos de dicho modelo: o bien son no-monotónicos pero no intencional-temporales o bien son intencional-temporales pero monotónicos. Así, nos preguntamos si las intenciones tienen propiedades que definan un mecanismo de inferencia tomando en cuenta estas propiedades; y si la respuesta es afirmativa, cómo podemos modelar el comportamiento de estos patrones de inferencia.

Veremos que, respecto al *problema externo*, las intenciones son estructuras de datos que poseen propiedades que justifican una noción de revisión; y respecto al *problema interno*, que el razonamiento intencional tiene una noción de inferencia lógica legítima.

Para mostrar la relevancia del problema externo podemos proceder por comparación en términos históricos. Dentro del modelo BDI no cabe duda que el fragmento B ha sido históricamente el más estudiado, por lo que la revisión de creencias se ha convertido en un programa de investigación paradigmático, relativamente nuevo y que une las dos disciplinas con las que hemos comenzado esta investigación: la filosofía y la IA.

Desde que los programadores se las vieron con estructuras de información, como las bases de datos, se enfrentaron con el problema de la actualización [96]. Por otro lado, los filósofos trataron el cambio de información dentro de estructuras epistémicas. Por esto hoy podemos identificar, respectivamente, dos momentos importantes en la historia de este programa de investigación: uno en el trabajo de Fagin, Ullman y Vardi [56] y el otro en las aportaciones de Harper [83] y Levi [103]; si bien la teoría que se encuentra en el trabajo fundamental de Alchourrón, Gärdenfors y Makinson (AGM) [1] constituye el núcleo de cualquier programa de revisión de creencias. En consecuencia, es fácil ver que el cambio racional de creencias bajo nueva información ha sido ampliamente estudiado durante los últimos 30 años, mientras que el proceso dinámico de otros estados mentales ha recibido menos atención y, en particular, hablamos de las intenciones [86], lo cual tiene explicación histórica, pero no está justificado.

Con respecto al problema interno podemos observar, por un lado, el caso de la lógicas BDI [136, 165] para capturar y entender la naturaleza o aspecto material del razonamiento intencional; y por otro lado, el caso de las lógicas derrotables [133] para tratar de sujetar y explicar el estatus o aspecto formal

del razonamiento derrotable. El problema con estos enfoques, no obstante, es que, en primer lugar, el razonamiento intencional no es monotónico [125], y por tanto, los modelos lógicos deberían ser no-monotónicos, pero las técnicas BDI son monotónicas; y en segundo lugar, los sistemas que son no-monotónicos no son intencional-temporales.

Por ello, ante este par de problemas proponemos dos cosas:

- Una *interpretación de funciones de aprendizaje automático* para responder al *problema externo*.
- El *desarrollo de un sistema lógico* para responder al *problema interno*.

Comencemos por precisar la primera propuesta. Si bien las intenciones han recibido atención desde los puntos de vista filosófico y computacional, sus características dinámicas no han sido estudiadas completamente. Y mientras la adaptación de los postulados de revisión de creencias para analizar los cambios intencionales, como veremos, provee una especificación abstracta y útil, no está comprometida con ningún mecanismo fijo o una implementación.

Así, la actual teoría de revisión, por su generalidad, no analiza explícitamente los eventos que producen cambios intencionales ni los mecanismos que definen tales cambios, mientras se enfoca únicamente en las tres operaciones de expansión, contracción y revisión, cuya completud como repertorio de acciones aún está abierta [16]. Sin embargo, argumentaremos que existe un procedimiento de aprendizaje automático para agentes BDI [77, 79] que puede ser usado, dentro de una interpretación bratmaniana, para fijar un mecanismo particular de revisión de intenciones que nos permite visualizar el *mecanismo* lógico de la revisión de intenciones como un proceso de aprendizaje.

Esta propuesta está inspirada por el modelo de revisión de creencias clásico AGM [1] y los estudios sobre el cambio racional de intenciones de Wiebe van der Hoek *et al* [86]. De manera particular, la idea de considerar el papel del aprendizaje intencional automático proviene del trabajo de Guerra *et al* [79].

En lo que concierne a la segunda propuesta nuestra principal contribución consiste en un estudio de la naturaleza y el estatus del razonamiento intencional sugiriendo la hipótesis de que este razonamiento es un razonamiento lógico *sui generis* por su doble naturaleza (aspecto material) temporal y derrotable y que puede ser considerado (aspecto formal) como razonamiento lógico legítimo dado que se comporta, *mutatis mutandis*, como un sistema lógico usando el concepto de modelo bratmaniano.

Esta parte de la investigación es importante de suyo porque el razonamiento derrotable tiene ciertos patrones de inferencia y por tanto el reto usual consiste en proveer una descripción razonable de tales patrones. La idea es que si la monotonía no es una propiedad del razonamiento intencional y queremos ofrecer una descripción adecuada de inferencia, entonces debemos estudiar las propiedades metalógicas de la inferencia intencional que ocurren en lugar de la monotonía, porque una vez que ésta se abandona: ¿Por qué deberíamos considerar la inferencia intencional como una instancia de lógica legítima? Argumentaremos, entonces, que la inferencia intencional tiene un *mecanismo* lógico.

La motivación de esta segunda propuesta proviene de los trabajos de Nute [125] y Governatori *et al* [72] en torno a la no-monotonicidad. En particular, la idea de considerar la adecuación formal y material tiene sus orígenes en las investigaciones de Antonelli [5].

A continuación resumimos de manera breve y protocolar lo que hemos dicho en esta presentación.

1.2. Planteamiento del problema

- ¿Es el razonamiento intencional razonamiento lógico *bona fide*?

1.3. Tesis

- El razonamiento intencional es razonamiento lógico *bona fide*.

1.4. Método

Responderemos el problema usando como marco global de discusión el problema de la persistencia. Asumiendo el modelo BDI como marco teórico identificamos dos subproblemas, uno externo, asociado al atributo de revisión; y otro interno, asociado al atributo de no-monotonicidad, cuyas soluciones servirán para sustentar nuestra tesis. Los subproblemas son los siguientes:

- ¿Cómo se comporta el mecanismo de revisión de intenciones?
- ¿Cómo se comporta el mecanismo de inferencia intencional?

Para responder a estos subproblemas seguimos una estrategia de tres pasos:

1. En primer lugar hacemos una lectura de los aspectos descriptivos y normativos de las intenciones y el razonamiento intencional según el modelo BDI de Bratman en *Intention, plans, and practical reason*.

Con este primer paso afirmaremos que, en efecto, hay características de las intenciones que permiten un mecanismo de razonamiento que tiene ciertas propiedades. Al estudiar estas propiedades notaremos que se puede abstraer un modelo que denominaremos *bratmaniano* y que justifica los atributos lógicos de revisión y no-monotonicidad, lo que nos permite ofrecer argumentos para responder a los problemas externo e interno.

2. Posteriormente exploramos cuáles son los mecanismos de revisión de intenciones y sugerimos cómo podemos fijar el mecanismo de la revisión de intenciones mediante aprendizaje automático.

La idea de este segundo paso consiste en asegurar que, en efecto, las propiedades de las intenciones permiten un mecanismo de revisión.

3. Finalmente, en un tercer momento, estudiamos los mecanismos de inferencia intencional y consideramos los aspectos material y formal de este tipo de inferencia definiendo un sistema lógico.

Con este tercer y último paso argumentamos que el razonamiento intencional tiene un mecanismo de inferencia.

Con estos tres pasos ofrecemos una respuesta afirmativa a la pregunta de si el razonamiento intencional es razonamiento lógico legítimo. El argumento es que si el razonamiento intencional tiene las características exigidas por los problemas externo e interno entonces es razonamiento lógico *bona fide*.

1.5. Objetivos

Los objetivos principales que pretendemos alcanzar son dos:

1. Al responder al problema externo argumentamos que las intenciones son objetos legítimos para una teoría mínima de revisión de intenciones y que esta teoría puede implementarse mediante aprendizaje automático.

2. Al responder al problema interno argumentamos que el razonamiento intencional es razonamiento lógico legítimo y que puede, en principio, mecanizarse.

1.6. Justificación

Por un lado, como decíamos, aunque las intenciones han recibido atención, sus características dinámicas no han sido estudiadas completamente. Considerando un conjunto de procedimientos de aprendizaje automático es posible visualizar cómo podemos entender la revisión de intenciones mediante el aprendizaje de intenciones, es decir, exploramos cómo algunos procedimientos de aprendizaje automático pueden usarse como mecanismos de revisión de intenciones. La relevancia de este punto consiste en visualizar, con miras a la implementación, cómo es que cambian esas estructuras de datos llamadas intenciones puesto que nos interesa un problema que es cognitivo y computacional de alto nivel: cómo se ven estos cambios y cuáles son las pautas para reproducir este tipo de comportamiento. La idea es, por tanto, que:

- Las intenciones son tan revisables como las creencias.

Bajo el problema interno que hemos descrito nuestra principal contribución consiste en un estudio de la naturaleza y el estatus del razonamiento intencional para mostrar que el razonamiento intencional es razonamiento lógico dado que se comporta, *mutatis mutandis*, lógicamente. La relevancia de este segundo punto consiste en visualizar cómo es el razonamiento intencional porque nos interesa un problema que es cognitivo y computacional de alto nivel: cómo se ve este tipo de razonamiento y cuáles son las direcciones para reproducir este tipo de comportamiento. La idea es, por tanto, que:

- La inferencia intencional es tan lógica como cualquier otro tipo de inferencia.

En consecuencia, *grosso modo*, nos interesa estudiar la cuestión del razonamiento intencional en el marco del problema de la persistencia porque queremos comprender de manera más clara cómo es que se comporta el razonamiento intencional. Y queremos entender esto último para determinar dos cosas: que el razonamiento intencional es tan lógico como cualquier otro tipo de razonamiento y que por tanto, hasta cierto punto, no sólo puede reproducirse mecánicamente sino que merece el mismo tipo de atención.

1.7. Trasfondo

Las principales obras que asumimos giran en torno a la teoría BDI de agencia racional de Bratman [27], puesto que asumimos tal modelo de manera explícita.

Además usamos el lenguaje formal $AgentSpeak(L)$ [136] para llevar a cabo nuestra formalización y la metodología de Bordini *et al* [18] para acompañarla.

Para el problema externo seguimos la propuesta de aprendizaje de Guerra *et al* [77] con el propósito de lograr nuestra interpretación del aprendizaje intencional y la teoría del cambio racional de intenciones de Wiebe van der Hoek *et al* [86] como motivación para la noción revisión de intenciones acompañada de la generalización clásica del modelo AGM [1].

Para el problema interno la motivación de la no-monotonicidad nos viene principalmente de las ideas de Nute [125] y Governatori *et al* [72]. Además seguimos muy de cerca la distinción entre adecuación formal y material de Antonelli [5].

1.8. Antecedentes del problema y estado del arte

Hay cuatro antecedentes bien definidos que muestran el estado de la cuestión. Aunque todos están vinculados, los dividiremos en dos por motivos de exposición. Para el problema externo tenemos como antecedentes *i)* el problema del aprendizaje en el modelo BDI y *ii)* el problema de la revisión de creencias. Para el problema interno podemos identificar *iii)* el problema de la laguna entre la teoría y la práctica y *iv)* el problema de la inferencia.

1.8.1. El problema del aprendizaje en el modelo BDI

El modelo computacional de agencia BDI tiene un problema: carece de un modelo explícito de aprendizaje.[7] A partir de esta carencia el problema del aprendizaje en Sistemas Multi-Agente bajo una arquitectura BDI ha sido estudiado individual y socialmente [76]. Actualmente existe un protocolo de aprendizaje social automático basado en la adopción cooperativa de metas en donde los agentes deciden cooperar intencionalmente para actualizar el contexto de sus planes cuya ejecución ha fallado. Cada agente, al interior de una comunidad de agentes, es capaz de recordar, hasta cierto punto, las creencias que soportan la adopción de planes como intenciones, así como el resultado de su ejecución: éxito o fallo. De este modo cada agente dispone de un conjunto de ejemplos de entrenamiento para aprender el contexto de éxito o fracaso al ejecutar planes. De esta manera

[7] En este contexto el aprendizaje automático ocurre cuando un agente captura su interacción con el mundo y su propio proceso de decisión mediante un diseño de retroalimentación y representacion [141].

los contextos de los planes se pueden interpretar como un conocimiento común que los agentes deben mantener actualizado y consistente.

En este modelo el protocolo de aprendizaje social se describe, *grosso modo*, así: cada vez que un agente falla en la ejecución de un plan intenta aprender el contexto del plan individualmente. Si esta intención falla, busca ayuda con otros agentes que comparten el mismo plan fallido. Estos agentes también comparten la intención de aprender un nuevo contexto para sus planes fallidos, de tal modo que envían de vuelta al agente inicial sus ejemplos de entrenamiento relevantes. Una vez que un nuevo y consistente contexto es aprendido el resultado es compartido por todos los agentes que participaron en este proceso de aprendizaje social. Este modelo de aprendizaje social permite resolver el problema del aprendizaje en el modelo BDI, aunque no ha sido usado explícitamente para fijar mecanismos de revisión de intenciones.[8]

1.8.2. El problema de la revisión de creencias

Las principales teorías de la intención han sugerido las ideas o problemas que una teoría sobre las intenciones debe tomar en cuenta; sin embargo, tales teorías no tratan con la dinámica de las intenciones. La dinámica de las intenciones debe tratar con el problema de cómo un agente adopta o abandona una intención y qué cambios producen estos procesos en los otros componentes de la arquitectura del agente [161, 162]. La dinámica de las intenciones, así, requiere de una teoría del cambio de intenciones o, más específicamente, una teoría de revisión de intenciones.

Si bien el cambio de creencias sobre la base de nueva información, como hemos dicho, ha sido ampliamente estudiado con cierto éxito, la revisión de otros estados mentales ha recibido poca atención. Aunque existen algunas teorías formales de la intención, la lógica de la revisión de intenciones ha sido escasamente considerada. Así, tenemos la propuesta de una teoría de revisión de intenciones [86, 126] y algunos experimentos con agentes programados en ambientes adecuadamente estructurados [164, 143], incluso en ambientes como el Tileworld [97].

[8] Al momento de escribir esto una librería de planes para llevar a cabo este proceso de aprendizaje está disponible en `jildt.sourceforge.net` con la versión 1.3.6a de la plataforma *Jason*.

1.8.3. El problema entre la teoría y práctica

La teoría de razonamiento práctico propuesta por Bratman expone los fundamentos filosóficos de los enfoques computacionales de agencia racional conocidos bajo la etiqueta BDI. La teoría de Bratman en su momento resultó innovadora porque no reducía las intenciones a una combinación de creencias y deseos; antes bien, asumía que las intenciones eran elementos particulares compuestos de planes parciales y jerárquicos. Bajo esta teoría diferentes sistemas lógicos BDI han sido propuestos para caracterizar formalmente la conducta racional de los agentes en términos de las propiedades de los operadores BDI y sus mutuas relaciones.

Debido a su expresividad estos sistemas han sido usados para razonar acerca de las propiedades racionales de los agentes, pero dado su alto costo computacional, no son usados para programarlos. Inversamente, lenguajes de programación de agentes, tales como $AgentSpeak(L)$, han sido propuestos para reducir la laguna entre la teoría (especificación lógica) y la práctica (implementación) de agentes racionales. Y aunque este lenguaje de programación, como veremos, tiene una semántica operacional bien definida, la verificación de propiedades racionales no es evidente, pues en éste las modalidades intencionales y temporales son abandonadas por mor de la eficiencia computacional. Este es un problema técnico que necesita ser solventado.

1.8.4. El problema de la inferencia

La lógica trata del proceso inferencial. En sentido abstracto trata con un conjunto de objetos y alguna relación especial entre ellos. En términos generales ésta suele ser una relación binaria entre morfismos de signaturas. Cuando dicha relación cumple las propiedades especificadas, por ejemplo, por Meseguer [115] y Tarski [156] y los objetos son sentencias o proposiciones, decimos que la relación es de consecuencia lógica deductiva. Esta noción de inferencia clásica, denotada por el signo $\vdash$, es una relación binaria que junto con cierto conjunto definido C de proposiciones constituye un par $\langle C, \vdash \rangle$ que configura un orden parcial y que además satisface las propiedades estructurales de monotonía, permutación, contracción y corte.

Cuando una inferencia no se comporta con las propiedades arriba mencionadas el reto usual consiste en proveer una descripción razonable de tal noción de inferencia, ya que si la monotonía no es una propiedad del razonamiento intencional y queremos ofrecer una descripción adecuada de su inferencia, entonces debemos estudiar las propiedades metalógicas de la inferencia intencional que

ocurren en lugar de la monotonía.

1.9. Exposición y método

Hemos dividido el trabajo en tres partes. La Primera, de carácter introductorio, expone los elementos preliminares que consideramos necesarios para entender el entramado conceptual de la siguiente. La Segunda Parte, de carácter propositivo, expone los avances y resultados de esta investigación. Así, mientras la Primera introduce la terminología y el marco teórico del que partimos, la Segunda desarrolla las propuestas de la investigación. En la última Parte hemos añadido algunos complementos.

La Primera Parte consta de 4 capítulos. En el Capítulo 2 exponemos información preliminar que sirve de sustento conceptual para seguir con los siguientes capítulos. Allí detallamos el trasfondo general y el marco teórico que seguiremos como hilo conductor durante esta investigación.

En el Capítulo 3 argumentamos que tanto la revisión de intenciones como la no-monotonicidad del razonamiento intencional son atributos lógicos legítimos que pueden estudiarse filosófica y lógicamente. Puede pensarse que esto es una reivindicación del fragmento I dentro del modelo BDI. La idea de este capítulo puede resumirse diciendo que un modelo BDI bratmaniano de razonamiento intencional debe considerar los atributos lógicos de revisión de intenciones y la no-monotonicidad del razonamiento intencional.

En el Capítulo 4 pasamos revista a los trabajos fundacionales de Cohen-Levesque y de Rao-Georgeff para formalizar el razonamiento intencional. Veremos que estos modelos no respetan las propiedades fundamentales de un modelo bratmaniano y que por ello es necesario otro modelo.

En el Capítulo 5 notamos que las implementaciones exitosas de la arquitectura BDI asumen una simplificación: modelan las actitudes intencionales como estructuras de datos y mejoran su desempeño computacional a cambio de una falta de fundamentos teóricos. El problema, entonces, es que los sistemas lógicos multimodales BDI, empleados en la especificación de estas arquitecturas poco ofrecen en relación con los problemas prácticos de su programación. Veremos que $AgentSpeak(L)$ es un lenguaje de programación para solventar este problema. Expondremos su sintaxis y su semántica, así como su teoría de prueba. Posteriormente revisaremos la sintaxis y la semántica de *Jason*, el intérprete de $AgentSpeak(L)$.

Con esto entraremos a la Segunda Parte del trabajo y empezaremos con los

pasos 2 y 3 de nuestra estrategia. Observaremos que en el desarrollo habrá una aparente separación entre el tratamiento del problema externo (Capítulos 7, 8, 9) y el tratamiento del problema interno (Capítulos 6, 10, 11) (Figura 1.1).

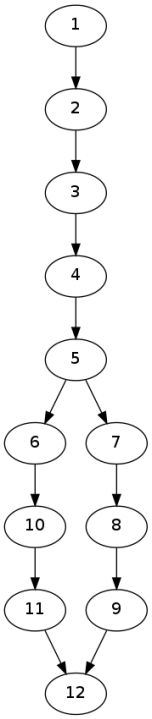


Figura 1.1: Estructura del trabajo

Para el Capítulo 6 habremos visto que diferentes sistemas lógicos han sido propuestos para caracterizar la conducta racional de los agentes BDI a partir de un modelo bratmaniano. Habremos apreciado también los problemas técnicos y profundos de estos sistemas. Además tendremos un dominio de $AgentSpeak(L)$ junto su intérprete $Jason$. Con esto podremos entender que el costo de la excelente semántica de $AgentSpeak(L)$ es la exclusión del uso de las modalidades que hacen a las lógicas BDI lenguajes altamente expresivos. Propondremos, entonces, para razonar acerca de los agentes BDI dentro de un modelo bratmaniano y resolver el problema técnico, el formalismo $BDI^{CTL}_{AgentSpeak(L)}$ como una herramienta lógica para la especificación y verificación de agentes BDI representados vía $AgentSpeak(L)$. Usaremos $BDI^{CTL}_{AgentSpeak(L)}$ como el sistema lógico para especificar formalmente a los agentes BDI.

En el Capítulo 7 nos separaremos un momento de esa problemática para discutir la noción de aprendizaje intencional y enfatizar su importancia, filosófica y técnicamente, en el comportamiento adecuado—racional—de los agentes BDI en su ambiente. A partir de este capítulo comenzaremos a enfrentarnos al problema externo; en particular, en él mostraremos unos resultados experimentales sobre aprendizaje automático de intenciones.

En el Capítulo 8 nos enfrentaremos de manera teórica a la conjetura que inferimos de los resultados experimentales. Observaremos que aunque la revisión de intenciones está justificada en un modelo BDI bratmaniano en términos declarativos no lo está en términos procedimentales.

Posteriormente, en el Capítulo 9 veremos que mientras la adaptación del modelo AGM para analizar los cambios intencionales provee una especificación abstracta, compatible y útil, no está comprometida o relacionada con ningún mecanismo fijo o una implementación de agentes BDI dentro de un modelo bratmaniano. Lo que haremos será interpretar la revisión de intenciones mediante el aprendizaje de intenciones. En este capítulo, entonces, mostraremos cómo podemos usar una implementación de aprendizaje BDI como un mecanismo para la revisión de intenciones, respondiendo, en consecuencia, al problema externo.

Así, hasta el Capítulo 10 habremos desarrollado un aparato analítico y crítico desde dos frentes para responder al problema externo. A partir de aquí volveremos nuestra atención al problema interno relacionado con el atributo de no-monotonicidad. El objetivo de este capítulo consistirá en acercarnos al desarrollo de un marco no-monotónico para representar intenciones considerando la cuestión del aspecto material.

El Capítulo 11, de carácter más formal, tratará el estatus del modelo expuesto en el capítulo anterior. Siguiendo la hipótesis de que el razonamiento intencional es un modo de razonamiento *sui generis* por su naturaleza derrotable y temporal, sugeriremos que también tiene el derecho a ser llamado *razonamiento lógico* por su comportamiento.

Finalmente, para cerrar el trabajo terminaremos en el Capítulo 12 con un resumen y una narración para integrar los resultados de la investigación. Por supuesto, como es costumbre, también tendremos un espacio dedicado para mencionar el trabajo que nos queda por hacer.

En la Tercera Parte hemos añadido un par de apéndices: en el Apéndice A mostraremos las pruebas y bosquejos de demostración de algunos de los resultados; en el Apéndice B describiremos las referencias de los productos que han sido fruto de esta investigación. Agregaremos, además, un Glosario que funciona como índice analítico y un apartado de Nomeclatura.

Parte I

Preliminares

Capítulo 2

El modelo de agencia BDI

2.1. Introducción

La idea que motiva este capítulo es sencilla: para empezar algo hay que empezar desde el principio. Estamos interesados en modelar cierto aspecto de la conducta inteligente, de manera precisa, nos interesa el razonamiento intencional. Pero como nos interesa este tipo de razonamiento desde un punto de vista cognitivo, para abordarlo seguimos, entonces, el modelo BDI instanciado a través de agentes BDI, por lo que necesitaremos hablar tanto de la noción de agencia que asumiremos como del modelo BDI.

Además, puesto que uno de nuestros objetivos conlleva la generación de un modelo formal a partir del modelo de agencia BDI describimos algunos detalles del lenguaje formal $AgentSpeak(L)$ que ha sido utilizado para representar este tipo de agencia.

Toda esta información es preliminar y sirve de sustento conceptual para seguir con los siguientes capítulos. Así pues, en este capítulo exponemos el trasfondo general y el marco teórico que seguiremos como hilo conductor durante esta investigación. En la Sección 2.2 argumentamos, a grandes rasgos, que nuestro enfoque tiene un perfil que podemos clasificar como cognitivo. Posteriormente describimos el concepto de *agencia* que asumimos para lograr nuestros propósitos (§ 2.3), las nociones básicas del modelo BDI de agencia racional (§ 2.4) junto con una ejemplificación de la arquitectura de agencia intencional (§ 2.5) y ciertos detalles del lenguaje $AgentSpeak(L)$ (§ 2.6) a los que haremos referencia constantemente en próximos capítulos.

2.2. Un enfoque cognitivo

Después de la Segunda Guerra Mundial, en especial en la década de 1970,[1] los programas de investigación alrededor del procesamiento de información comenzaron a construir una nueva ciencia: la ciencia cognitiva, entendiendo por tal a la empresa de responder, con los conocimientos y tecnologías disponibles, cuestiones epistemológicas que habían estado presentes, históricamente, desde los orígenes de la filosofía; sin embargo, esta nueva ciencia tenía características novedosas gracias a tres supuestos:

- Que para estudiar los procesos cognitivos hay que servirse de una metodología científica basada en estudios empíricos y formales, pero sobre todo, interdisciplinarios.
- Que, además, para explicar la organización y el funcionamiento de los procesos cognitivos es preciso suponer representaciones (conceptos, imágenes, esquemas, modelos, etc.) susceptibles de ser analizadas a un nivel distinto del biológico o neurológico, por un lado; y del sociológico o cultural por otro; y que también es diferente del introspectivo o fenomenológico.
- Y que hay distintos niveles de explicación que deberían interactuar:[2]

[1] Es interesante destacar que el primer número de la revista *Cognitive Science* es de 1977. En la editorial del primer número Allan Collins afirma [42]:

> "Recently there has begun to grow a community of people from different disciplines, who find themselves tackling a common set of problems in natural and artificial intelligence. The particular disciplines from which they come are cognitive and social psychology, artificial intelligence, computational linguistics, educational technology, and even epistemology. The work of these researchers is converging toward a coherent point of view that is different from the focus of any of the current journals. This view has recently begun to produce a spate of books and conferences, which are the first trappings of an emerging discipline. This discipline might have been called applied epistemology or intelligence theory, but someone on high declared it should be cognitive science and so it shall".

[2] Al respecto, en el mismo lugar Collins detalla algunas de las "normas" que los autores deben seguir para someter un artículo a la revista [42]:

> "It will perhaps be helpful to define what kinds of articles are *not* appropriate for a journal in cognitive science. I will list examples below of articles which the editors will be biased against no matter how excellent they may be:
>
> 1. Descriptions of intelligent systems that do not provide any insight into human processing.
> 2. Detailed system descriptions that do not focus on theoretical issues.
> 3. Descriptions of efficient algorithms for dealing with problems in artificial intelligence.
> 4. Descriptions of experimental work that do not provide any insight into the construction of computer models of intelligent behavior.

- Un *nivel computacional* que especifica la naturaleza de la información.
- Un *nivel algorítmico* que especifica de manera abstracta las tareas a cumplir por parte del sistema.
- Un *nivel físico* que especifica la realización de estos procesos.

Entendiendo el término *cognición* en un sentido amplio, incluyendo la adquisición, elaboración y uso de la información, un grupo particular de disciplinas se ha concentrado para estudiar dichos procesos siguiendo los supuestos anteriores. De esta manera, no existe propiamente *la* ciencia cognitiva, sino más bien una familia de ciencias que estudian fenómenos cognitivos. Así, tenemos a la lingüística de origen chomskyano, la inteligencia artificial, las neurociencias, la psicología, la antropología cultural y, por supuesto, la filosofía. Estas disciplinas constituyen el clásico hexágono cognitivo de Gardner [67] (Figura 2.1).

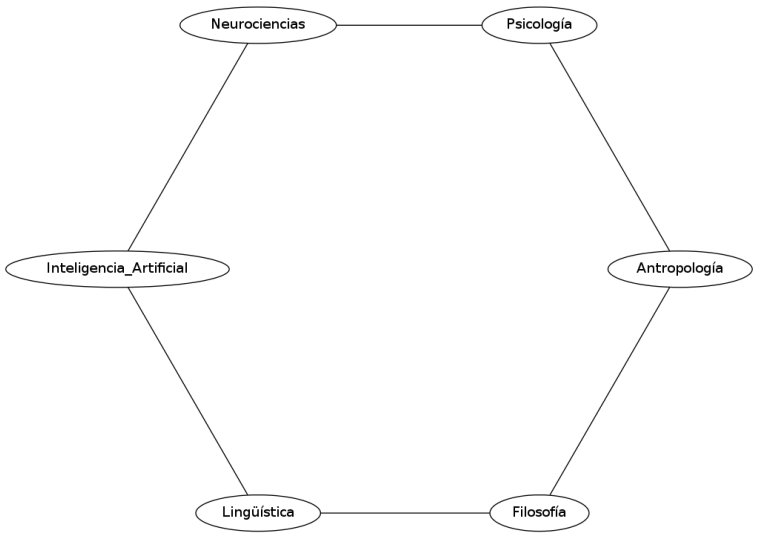


Figura 2.1: Hexágono cognitivo

5. Experimental work on cognitive processing in artificial tasks rather than tasks which strive for a natural experimental setting.
6. Work on theories where the representation of knowledge is not of concern, such as content-free, mathematical theories.
7. Work on peripheral processing such as buffer storage or short-term memory, except as part of a larger theory.
8. Work on small-scale models, particularly those developed for a specific psychological paradigm, rather than on complex interactive theories".

En este trabajo se ven claramente dos disciplinas incluidas en el hexágono: la filosofía y la inteligencia artificial. A partir de éstas podemos definir una triple perspectiva generada por las ciencias computacionales, la lógica y la filosofía; y que configura un marco teórico lo suficientemente rico como para contener nuestros propósitos dentro de un enfoque cognitivo (Figura 2.2). Ciertamente, los límites entre estas áreas del conocimiento pueden ser a veces muy sutiles y otras veces muy bruscos, pero creemos que esta triple perspectiva no sólo es interesante, sino útil.

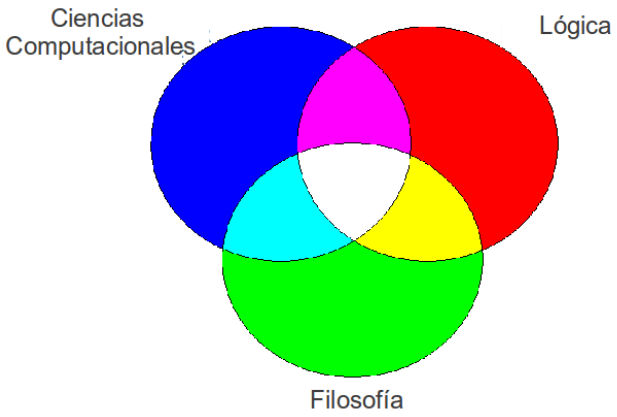


Figura 2.2: Enfoque cognitivo I

Dado que los problemas que estamos considerando en este trabajo incluyen, entre otras cosas, la especificación, formalización y, hasta cierto punto, la verificación de algunas propiedades de agentes racionales, podemos afirmar que existen ciertas intersecciones importantes entre estas ciencias.

Del encuentro entre filosofía y ciencias computacionales obtenemos la *especificación* de agentes por medio de arquitecturas abstractas que describen a los agentes, sus elementos y sus mutuas relaciones. En nuestro caso usamos las nociones básicas del modelo BDI como ha sido expuesto por Bratman porque, más allá de ser un modelo filosófico, ha sido usado, implementado y enriquecido por las ciencias de la computación en términos de arquitecturas computacionales.

A partir de la intersección entre lógica y filosofía obtenemos la *formalización* de los agentes BDI con la ayuda de la lógica y los sistemas formales. En nuestro caso proponemos una formalización de agentes BDI cercana a $AgentSpeak(L)$.

Finalmente, de la intersección entre lógica y ciencias de la computación obtenemos la *verificación* formal de propiedades de agentes BDI usando ciertas

formalizaciones y técnicas (Figura 2.3).

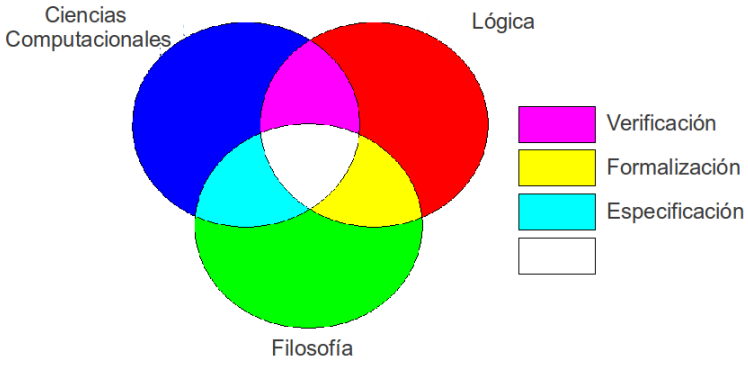


Figura 2.3: Intersecciones

Como puede notarse, el área realmente importante aquí es la intersección que ocurre entre todas éstas. Dicha intersección define el marco de referencia en el cual trabajamos: arquitecturas BDI, sistemas formales y *model checking*. Esa pequeña área de color blanco representa los medios y los fines de nuestro trabajo, pues nos muestra una suerte de enfoque cognitivo tal y como lo hemos descrito previamente (Figura 2.4).

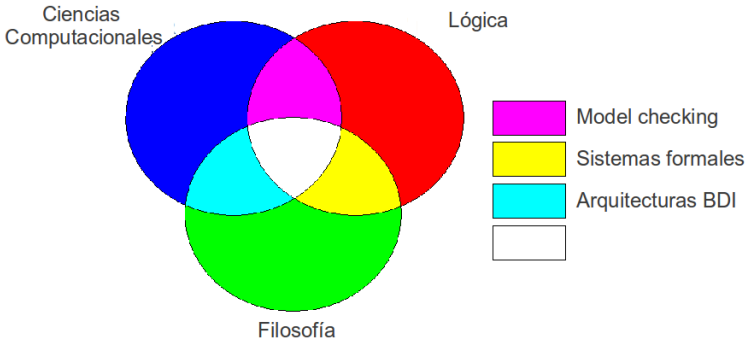


Figura 2.4: Enfoque cognitivo II

2.3. Agencia

Desde un punto de vista cognitivo un agente BDI es un sistema agente descrito en términos del modelo BDI. Por supuesto, en este momento esta caracte-

rización nos dice poco o nada. Lo que haremos en esta sección será explicar el sentido de la expresión *agente* usando una aproximación informal y otra formal.

2.3.1. Aproximación informal

Es un lugar común aseverar que entre los propósitos compartidos por la filosofía y la IA se encuentra, fundamentalmente, el estudio sistemático del comportamiento inteligente. Este estudio, desde un enfoque cognitivo, tiene un carácter, *prima facie*, doble: científico e ingenieril [68]. Científico, porque se encarga del estudio de los procesos cognitivos de manera general; ingenieril, porque se las ve con la construcción de sistemas inteligentes.

Así, es usual asumir que la filosofía y la IA tienen como uno de sus objetivos principales lograr el entendimiento, con miras a la reproducción, del comportamiento inteligente. Bajo esta idea la filosofía y la IA pueden verse como dos diciplinas del enfoque cognitivo que, entre otras cosas, se dedican a estudiar agentes de manera descriptiva y prescriptiva.[3]

Sin embargo, qué sea un agente no es algo evidente. En un sentido muy general se dice que un agente es cualquier cosa capaz de percibir y actuar. Bajo los términos de esta definición casi toda entidad puede clasificarse como agente. Una definición de carácter propiamente filosófico y con más tradición puede ser rastreada, con los matices pertinentes, en el *nous poietikós* aristotélico interpretado como un entendimiento que ejecuta cierto tipo acciones [7]. Esta idea, quizá muy moderna para su tiempo, hace que la definición sea más precisa al introducir un elemento cognitivo.

Esta definición clásica tiene una traducción contemporánea en términos de las ciencias computacionales: un agente, se dice, es un sistema de información situado en un ambiente y que es capaz de actuar autónomamente sobre él para alcanzar sus objetivos (adaptada de [166]). Mientras el ambiente, a su vez, se entiende como el entorno físico [33] o virtual [55] que provee el *locus* donde las propiedades del agente se enfatizan.

Formalmente, como explicaremos más adelante, un sistema agente es modelado como un par $\langle ag, Env \rangle$ donde ag es propiamente la especificación formal de un agente y Env la de un ambiente [166]. Esta caracterización de una arquitectura agente abstracta muestra algo que es muy relevante para nuestra investigación: que la relación entre un agente y su ambiente es tan estrecha que

[3] Este es, en particular, el enfoque o paradigma de la IA contemporánea propuesto principalmente por Russell y Norvig [141] y Nilsson [123].

al hablar de un agente no se puede hablar de él sin su ambiente [166].

Bajo esta arquitectura abstracta e informal se dice que un agente tiene ciertas propiedades que también se especifican informalmente: la *autonomía*, la *reactividad*, la *proactividad* y las *habilidades sociales* [69, 166] suelen ser características predicables de un agente *bona fide*. A estas propiedades generales es usual añadir ciertos predicados cognitivos como *conocimiento*, *creencia*, *deseo*, *compromiso* e *intención* [152], predicados que son descritos, con suficiente detalle, en el modelo BDI. Kiss [98], Shoham y Cousins [153] clasifican estas propiedades bajo el nombre de *actitudes intencionales* (Caudro 2.1).

Actitud	**Kiss**	**Shoham y Cousins**
Creencia	Cognitiva	Informativa
Intención, compromiso, plan	Conativa	Motivacional
Meta, deseo, preferencia	Afectiva	Motivacional
Obligación, permisión	—	Social

Cuadro 2.1: Clasificación de actitudes intencionales

Esta caracterización informal, si bien sigue siendo general, provee un nivel de abstracción que presenta las siguientes ventajas:

- Permite enumerar las facultades cognitivas de los agentes al realizar sus acciones.
- Permite considerar diferentes tipos de agentes, incluyendo aquellos que no se supone tengan tales facultades cognitivas.
- Permite considerar diferentes especificaciones sobre los sub-sistemas que componen a los agentes.
- Muestra que un agente tiene que especificarse con su ambiente correspondiente.

El ambiente, por su parte, también puede caracterizarse. Brooks, por ejemplo, considera que el medio ambiente por antonomasia es el mundo real, que el mundo es el mejor modelo del mundo, por lo que un agente debe tener una implementación física (en el caso de Brooks, una implementación robótica) [32]. Por otro lado, Etzioni argumenta que no es necesario que los agentes tengan implementaciones físicas dado que los ambientes virtuales, como los sistemas operativos y la web, son tan reales como el mundo real [55]. En este trabajo

estamos de acuerdo con que no es necesario tener tales implementaciones físicas porque un enfoque cognitivo garantiza y justifica, sin ser excluyentes, que lo importante para este tipo de estudio es que la interacción del agente con su ambiente se dé en términos de arquitecturas bien descritas.

Para describir un ambiente Russell y Norvig [141] sugieren las siguientes categorías:

- Si un agente puede percibir a través de sus sensores los estados completos del ambiente donde se encuentra, se dice que el ambiente es *accesible*; de otro modo, es *inaccesible*. Esta propiedad depende no sólo del ambiente, sino de las capacidades de percepción del agente. Como puede notarse, mientras más accesible sea el ambiente, más sencillo será de construir.

- Si el siguiente estado del ambiente está determinado por la acción del agente, se dice que el ambiente es *determinista*. Si otros factores influyen en el próximo estado del ambiente, se dice que éste es *no-determinista*. El no-determinismo implica dos nociones importantes: *i)* que los agentes tienen un control parcial sobre el ambiente, y *ii)* que las acciones del agente pueden fallar.

- Si la experiencia del agente puede evaluarse a través de episodios o rondas, decimos que el ambiente es *episódico*. Las acciones se evalúan en cada episodio. Dada la persistencia temporal de los agentes, estos tienen que hacer continuamente decisiones locales que tienen consecuencias globales. Los episodios reducen el impacto de estas consecuencias, y por lo tanto es más sencillo construir agentes en ambientes episódicos.

- Si el ambiente puede cambiar mientras el agente se encuentra deliberando, se dice que el ambiente es *dinámico*; de otro modo, es *estático*. Si el ambiente no cambia con el paso del tiempo, pero la evaluación de las acciones del agente si lo hace, se dice que el ambiente es *semi-dinámico*.

- Si hay un conjunto finito numerable de posibles estados del ambiente, distintos y claramente definidos, se dice que el ambiente es *discreto*; de otro modo, se dice que es *continuo*.

Esta categorización sugiere que es posible encontrar diferentes clases de ambientes. Russell y Norvig [141] presentan algunos ejemplos de ambientes bien estudiados en IA y sus propiedades (Cuadro 2.2). Cada ambiente, o clase de ambientes, requiere de alguna forma agentes diferentes para que estos tengan éxito.

La clase más compleja de ambientes corresponde a aquellos que son inaccesibles, no-episódicos, dinámicos y continuos.

Ambiente	Accesible	Determinista	Episódico	Estático	Discreto
Ajedrez sin reloj	sí	sí	no	sí	sí
Ajedrez con reloj	sí	sí	no	semi	sí
Análisis de imágenes	sí	sí	sí	semi	no
Backgammon	sí	no	no	sí	sí
Póker	no	no	no	sí	sí
Tutor de inglés	no	no	no	no	sí
Robot toma piezas	no	no	sí	no	no
Controlador de refinería	no	no	no	no	no
Robot navegador	no	no	no	no	no
Conductor de autos	no	no	no	no	no
Diagnóstico médico	no	no	no	no	no

Cuadro 2.2: Ejemplos de ambientes estudiados en IA y sus propiedades

Con estos parámetros, por ejemplo, aunque es discutible concebir a un *daemon* de sistema operativo, como *xbiff*, como un agente, podemos mostrar cómo esta aproximación informal permite una caracterización: *xbiff* de algún modo se las arregla para identificar a su usuario, encontrar su buzón electrónico en la red (su ambiente accesible, determinista, episódico, semi-dinámico y discreto), buscar mensajes nuevos y comunicar al usuario la presencia de éstos.[4] El resultado de hacer esto es que podemos aproximar la definición de *xbiff* de una manera más comprensible.

2.3.2. Agencia débil

Asumiendo, entonces, que un agente es un sistema de información situado en un ambiente y que es capaz de actuar autónomamente sobre él para alcanzar sus objetivos, podemos distinguir algunas propiedades:

- La *autonomía* en este contexto significa que, una vez activo, un agente opera en su ambiente sin intervención directa externa y tiene cierto control sobre sus acciones y su estado interno [166].[5]

[4] El agente *xbiff* es un *daemon* del sistema X Windows situado en un ambiente UNIX y que vigila constantemente el buzón del usuario para avisarle cuándo llegan mensajes nuevos a través de una interfaz gráfica.

[5] Esta es posiblemente la propiedad que genera más discusiones, pues aunque la investigación reciente sobre el tema de la autonomía en agentes tiene aspectos interesantes, es superada por su falta de sistematicidad [159].

- Al estar situado en un ambiente responde de manera consecuente. A esto se le llama *reactividad* [166].
- Además se dice que el agente es *proactivo* en tanto que exhibe una conducta orientada a metas u objetivos [166].
- Y como puede interactuar con otros sistemas a través de cierto lenguaje se dice que posee *habilidades sociales* [69].

Estas propiedades básicas determinan la noción de agencia en sentido *débil*.

2.3.3. Agencia fuerte

Cuando además incluimos una serie de propiedades mentalistas para describir a los agentes obtenemos el sentido de agencia *fuerte*. Así, bajo esta aproximación fuerte es común caracterizar a los agentes usando nociones o predicados cognitivos como *conocimento*, *creencia*, *intención*, *obligación* [152] y *emoción* [15].

Como será fácil de apreciar, es en el marco conceptual de la agencia fuerte que el concepto de agencia BDI adquiere significado teórico y práctico. Teórico, porque el modelo BDI así lo exige; práctico, porque el modelo BDI es suficientemente general como para extender sus resultados a diferentes clases de agentes que pueden entederse en términos BDI.

2.3.4. Otros atributos

Finalmente, algunos autores sugieren otras propiedades que están implícitas en la especificación de los agentes:

- Se dice que un agente tiene la propiedad de *movilidad* cuando tiene la habilidad para moverse en su ambiente [160].
- Cuando se requiere que un agente no comunique información falsa se dice que es *veraz* [65].
- Se dice que es *benevolente* y que tratará de alcanzar sus metas [140].
- Y que en tanto que actúa para alcanzar sus objetivos es *racional* [65].

Una discusión sobre varios atributos de agencia aparece en [71]. Una clasificación más nueva se la debemos a Franklin y Graesser [63] (Cuadro 2.3).

Propiedad	Significado
Reactivo	El agente responde de manera rápida a los cambios en el ambiente
Autónomo	El agente ejerce control sobre sus propias acciones
Orientado-a-metas	El agente es proactivo, no meramente reactivo
Temporal	El agente es un proceso continuo
Comunicativo	El agente es capaz de comunicarse con otros agentes
Aprendiz	El agente muestra cambios adaptativos en su conducta dadas experiencias previas
Móvil	El agente es capaz de transportarse a sí mismo
Flexible	Las acciones del agente no están totalmente determinadas (*scripted*)
Carácter	El agente tiene personalidad creíble y estado emocional

Cuadro 2.3: Una taxonomía de agentes

2.3.5. Aproximación formal

La aproximación informal que acabamos de relatar tiene la fortuna de que puede definirse formalmente en una arquitectura abstracta [166]. Para hacer esto comenzamos introduciendo:

Definición 1 *(Conjunto de estados) Un conjunto finito de estados* $E = \{e_0, ..., e_n\}$ *donde* e_i *es una configuración agente.*

Por el momento basta aclarar que un estado e_i es una configuración de un agente en un momento dado. Además, según la clasificación de ambientes que observamos previamente, los estados no necesariamente han de ser discretos, pero en esta aproximación asumiremos, sin pérdida de generalidad, que E es discreto.

Posteriormente definimos:

Definición 2 *(Conjunto de acciones) Un conjunto de acciones* $A = \{\alpha_0, ..., \alpha_n\}$ *donde* α_i *es una acción que ejecuta el agente.*

Donde una acción α_i es una función $a_i : E \longrightarrow E$ que va de un estado a otro. Con esto se define una corrida:

Definición 3 *(Corrida) Dado un conjunto de estados* $E = \{e_0, ..., e_n\}$ *y un conjunto de acciones* $A = \{\alpha_0, ..., \alpha_n\}$*, una corrida se define como una secuencia* c*:*

$$c = e_0 \xrightarrow{\alpha_0} e_1, ..., \xrightarrow{\alpha_{n-1}} e_n$$

donde $e_k \xrightarrow{\alpha_m} e_{k+1}$ *expresa que un estado se transforma en el estado siguiente mediante la acción* α_m*.*

Decimos que $C = \cup_{i=1}^{n} c_i$ es el conjunto de todas las corridas. Así, C_{A_i} será el subconjunto de las corridas que terminan en una acción i y C_{E_n} el subconjunto de C que termina en un estado n del ambiente.

Con esto podemos definir la acción de un agente en el ambiente:

Definición 4 *(Acción en el ambiente) Dado el subconjunto de las corridas que terminan en una acción i, C_{A_i}, y un conjunto de estados $E = \{e_0, ..., e_n\}$, la acción en un ambiente es una función τ que va de las corridas que terminan en una acción a todos los estados posibles, $\tau : C_{A_i} \longrightarrow \wp(E)$.*

De este modo el ambiente es sensible a su historia, por lo que las acciones ejecutadas por el agente en el pasado también afectan la transición a estados futuros.

Con esto estamos en condiciones de definir formalmente las caracterizaciones informales de ambiente y agente:

Definición 5 *(Ambiente) Un ambiente es una tripleta $Env = \langle E, e_0, \tau \rangle$ donde E es un conjunto de estados, e_0 es un estado inicial y τ es una función de acción sobre el ambiente.*

Definición 6 *(Agente) Un agente es una función $ag : C_{E_n} \longrightarrow A$ donde C_{E_n} es el subconjunto de las corridas que terminan en un estado n del ambiente y A es un conjunto de acciones.*

De este modo:

Definición 7 *(Sistema agente) Un sistema agente es un par $\langle ag, Env \rangle$ donde ag es un agente y Env es un ambiente.*

Esta arquitectura abstracta de un sistema agente nos muestra la relación que hay entre el agente y su ambiente (Figura 2.5).

El Algoritmo 1 muestra la función que implementa un agente de este tipo, donde p es una percepción del ambiente y $mapeo$ es un conjunto de acciones predefinido.

Algoritmo 1: Agente basado en un mapeo ideal

Datos : p
$percepciones \leftarrow percepciones \cup p$
$acción \leftarrow busca(percepciones, mapeo)$
return $acción$

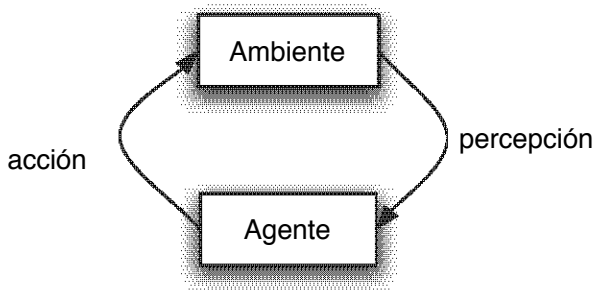


Figura 2.5: Abstracción de un agente a partir de su interacción con el ambiente

Por supuesto, también es posible implementar un programa básico de ambiente que ilustre la relación entre éste y los agentes situados en él. El Algoritmo 2 muestra el programa *Ambiente*. La historicidad del ambiente queda oculta en las percepciones del agente.

Algoritmo 2: Ambiente

Datos : e_0, τ, ags
mientras $\top$ **hacer**
  para $ag \in ags$ **hacer**
    $p(ag) \leftarrow percibir(ag, e_0)$
  fin
  para $ag \in ags$ **hacer**
    $acción(ag) \leftarrow ag(p(ag))$
  fin
  $e \leftarrow \tau(\bigcup_{ag \in ags} acción(ag))$
fin

Donde ags es un conjunto de agentes, e_0 es el estado inicial del ambiente y τ es una función de transición del ambiente.

2.4. Agentes como sistemas intencionales

Hasta este punto tenemos una descripción de lo que entendemos por un sistema agente. Ahora necesitamos detallar los componentes de agencia fuerte que se describen en el modelo de agencia BDI. Antes de pasar a exponer directa-

mente las generalidades de dicho modelo nos permitimos discutir, brevemente, la interpretación de los agentes como sistemas intencionales, pues la noción de *intencionalidad* no es unívoca ni carece de historia.

Así, por ejemplo, la escolástica entendió una noción de intencionalidad en términos de la suposición de que todos los hechos de consciencia poseen y manifiestan una dirección u orientación hacia un objeto. Esta orientación, que se afirmaba de todo pensamiento, volición o representación consciente en general consistía, por un lado, en la presencia o existencia mental del objeto que se piensa, quiere o representa y, por otro, en la referencia de este hecho mental a un objeto, en principio, real.

Franz Brentano, por ejemplo, desarrolló la tesis de que la intencionalidad es la característica propia de todos los fenómenos mentales [31] y sus estudios influyeron particularmente en el desarrollo de la fenomenología de Husserl, para quien la intencionalidad es como una teleología de la consciencia, de modo que toda consciencia es siempre consciencia de algo.

Husserl estipula la conocida distinción entre *nóesis* y *nóema*, entendiendo por *nóesis* al hecho mental o la vivencia y por *nóema* el contenido de tal hecho o vivencia: la intencionalidad, a su vez, es la correlación entre estos dos componentes. Estas tesis husserlianas tienen importancia general en cuanto pretenden ser una superación de la distinción que existía entre sujeto y objeto en las teorías clásicas y modernas del conocimiento, siendo la intencionalidad de la conciencia el punto de fusión entre ambos componentes [91].

Aquí, sin embargo, manejaremos *intencionalidad* desde otra tradición y, por ello, en otros sentidos: uno relacionado con el lenguaje, otro con la interpretación de la conducta y otro relacionado con el razonamiento.

2.4.1. Intencionalidad y lenguaje

El uso de *intencionalidad* como aquí será entendido tiene sus raíces en la tradición analítica iniciada por Frege y Russell. Esta tradición abstrae la intencionalidad de la conciencia y la sitúa en las actitudes proposicionales [119]. Cuando a esta postura se le añaden algunos supuestos generales del conductismo clásico resulta un concepto de intencionalidad que se define como un comportamiento lingüístico y que tiene su mejor explicación en la teoría de los actos de habla de Austin [12] y Searle [144].

El término *actitud proposicional* es usado para denotar un estado mental intencional. De alguna forma este término enfatiza el hecho de que no todos los estados mentales son intencionales: las creencias, los deseos y las intenciones son

intencionales, pero algunas formas de nerviosismo y ansiedad no lo son. A su vez, el término *estado* es usado para expresar ciertas características globales de un sistema o que el sistema se encuentra en una situación bien identificada.

Tradicionalmente tres actitudes proposicionales son consideradas para modelar agentes racionales: creencias, deseos e intenciones, las actitudes que le dan el nombre al modelo BDI que explicaremos más adelante. Ferber [58], por poner un ejemplo, presenta una clasificación de estas actitudes a través de sus usos:

- Con el *uso interactivo*: percepción, información, comando, petición, norma.
- Con el *uso representacional*: creencias, hipótesis.
- Con el *uso conativo*: deseos, metas, impulsos, demandas, intenciones, compromisos.
- Con el *uso organizacional*: métodos, tareas.

2.4.2. Intencionalidad y conducta

Hay otro sentido relacionado con la intencionalidad y la interpretación de la conducta. Para mostrarlo resulta pertinente iniciar comentando que si bien McCarthy es uno de los primeros en argumentar a favor de la adscripción de estados intencionales a máquinas abstractas [110], es el enfoque de Dennett el que ha sido utilizado con más frecuencia como el fundamento filosófico para describir a los agentes como entidades a las que se les pueden predicar creencias, deseos y otras actitudes intencionales. A esta posición se le conoce como la postura intencional [51] y se puede describir usando la *escalera de personalidad* [50].[6]

De acuerdo a Dennett los sistemas intencionales son, por definición, todas y sólo aquellas entidades cuyo comportamiento puede ser explicado o predicho desde una posición intencional, la cual consiste en la interpretación del comportamiento de tal entidad (persona, animal, artefacto o lo que sea) como si fuera un agente racional que gobierna su selección de acción considerando sus actitudes intencionales.

A modo de ejemplo consideremos el caso del apagador de luz de Yoav Shoham [152]: es perfectamente coherente interpretar a un apagador como un

[6]La escalera de personalidad tiene la siguiente estructura: racionalidad, intencionalidad, postura (*stance*), reciprocidad, comunicación, conciencia. Los dos primeros componentes describen un *sistema intencional básico* [50], concepto que reaparecerá formalmente interpretado en el Capítulo 4, §4.5.

agente con la capacidad de transmitir corriente eléctrica cuando queremos el cuarto iluminado y de no hacerlo en cualquier otra circunstancia. Oprimir el botón del apagador es una forma de comunicarle nuestros deseos. En este ejemplo la descripción intencional es, en efecto, simplista, porque no es necesaria; pero hay casos de mayor complejidad que la requieren, por ejemplo, al explicar conductas como la siguiente: *Juan leyó el* Ulises *de Joyce porque* creía *que era una buena obra* [120].

Dennett propone otras dos aproximaciones posibles para interpretar la conducta: la física, donde usamos nuestro conocimiento de la física junto con el estado físico del agente para explicar su conducta; y la de diseño, que se basa en el conocimiento de las funciones de los componentes del agente y que, por ejemplo, se adapta perfectamente al caso del apagador.

Lo que podemos inferir de estos niveles de explicación es que para sistemas más complejos y sofisticados, como los agentes inteligentes—o BDI—como nosotros, las aproximaciones física y de diseño no siempre están disponibles o no es práctico usarlas. Por lo general, cuanto más complejo resulta el sistema que queremos describir, más necesitamos del enfoque intencional.

Dennett propuso, así, el término *intencional* para referirse a las entidades cuya conducta pudiera ser explicada a través de la atribución de tales actitudes intencionales pertenecientes a la psicología *folk* [51] y observó, además, que hay grados de intencionalidad:

- Cuando tenemos sistemas intencionales con creencias, deseos y otras actitudes proposicionales pero sin creencias ni deseos acerca de sus propias creencias y deseos (sin actitudes proposicionales anidadas) tenemos *sistemas intencionales de primer orden*.

- Cuando tenemos sistemas intencionales con creencias, deseos y otras actitudes proposicionales, más creencias y deseos acerca de sus propias creencias y deseos (con actitudes proposicionales anidadas) estamos ante la presencia de *sistemas intencionales de segundo orden*.

- Como la jerarquía de intencionalidad puede extenderse tanto como sea necesario tenemos *sistemas intencionales de orden* $n > 2$.

A su vez, las actitudes intencionales de la psicología *folk* [62], a pesar de ciertas desventajas,[7] están lo suficientemente establecidas como para desarrollarlas formalmente y poseen las siguientes ventajas [150]:

[7] Paul Churchland es muy conocido por haber criticado duramente esta postura observando

- Nos son familiares a todos: diseñadores, analistas de sistemas, programadores y filósofos.
- Nos permiten descripciones sucintas del comportamiento de los sistemas complejos, por lo que ayuda a entenderlos y explicarlos.
- Su representación formal provee de ciertas regularidades y patrones que son independientes de la implementación física de los agentes.
- Además, una vez desarrollado, un agente puede razonar sobre sí mismo y sobre otros agentes usando este modelo *folk*.

2.4.3. Intencionalidad y razonamiento

Finalmente, el tercer sentido de intencionalidad que queremos referir aquí tiene su origen en la propuesta de razonamiento práctico desarrollada por Michael Bratman en *Intention, plans, and practical reason* [27].

Esta teoría define un marco de tipo cognitivo para entender a *otros*[8] y a nosotros mismos usando como eje una noción de intencionalidad. Su idea central es que nuestra concepción de las intenciones, de acuerdo al sentido común, está ligada al fenómeno de la planificación y a los planes. Esto provee las bases para entender, comunicar e incluso predecir lo que los demás harán, explicar porqué han hecho lo que han hecho e incluso para coordinar nuestras acciones con las suyas.

Esta noción de intencionalidad está relacionada con un tipo de razonamiento que comprende dos actividades:

- La *deliberación* que consiste en decidir qué metas debe llevar a cabo un agente.
- Y el *análisis medios-fines* para determinar cómo es que el agente va a satisfacer sus metas.

Ambas actividades pueden verse como procesos computacionales ejecutados por agentes racionales acotados [147]. La racionalidad acotada en este contexto consiste en una racionalidad con un par de limitaciones:

---

que las actitudes intencionales no tienen un carácter científico y por haber argumentado que la psicología *folk* sería desplazada por las neurociencias [39].

[8] El subrayado es nuestro.

- Una externa derivada del *dinamismo del ambiente*: en ambientes dinámicos un agente debe controlar su razonamiento eficientemente para tener un buen desempeño en términos de costos, beneficios y recursos disponibles.
- Una interna expresada por la noción de *no-retención infinita*.[9] Los agentes no pueden deliberar indefinidamente; deben en algún momento elegir las metas a atender y comprometerse a satisfacerlas.

Estos tres sentidos nos permiten interpretar la noción de intencionalidad como un fenómeno susceptible de análisis lógico: dado que las intenciones se pueden interpretar intencionalmente como planes, es posible representarlas lógicamente como actitudes proposicionales.

2.5. IRMA: una arquitectura intencional

Veremos, por tanto, que dado que las intenciones se pueden representar como actitudes proposicionales, es legítimo enriquecer al modelo de agencia inicial con un componente intencional que, para nuestros fines, tiene una modelación perspicua en el modelo de agencia BDI. Pero antes de explicar los detalles de este modelo nos permitimos ejemplificar los conceptos de la agencia intencional que hemos discutido previamente mostrando la arquitectura abstracta IRMA (por *Intelligent Resource-Bounded Machine*) [30]. Esta arquitectura tiene cinco componentes que nos permitirán entender sin problema el modelo BDI:

- Las *creencias* (*beliefs* denotadas por B) representan el estado del mundo. Se expresan simbólicamente como hechos en Prolog [22], es decir, son literales aterrizadas (i.e., *grounded*) de la lógica de primer orden [124].[10] Para efectos de esta arquitectura abstracta no es necesario especificar la representación exacta de las creencias.
- Los *deseos* (*desires* denotados por D) representan estados buscados por el agente. Al igual que con las creencias, los detalles completos de representación serán por el momento omitidos.

[9] *No-infinite deferral.*

[10] Una fórmula atómica cuyos argumentos son constantes se llama fórmula atómica aterrizada. Una literal es una fórmula atómica o su negación. Cuando no hay variables entre los argumentos de la fórmula atómica, se dice que la literal es aterrizada. Así, por ejemplo, $p(X, a)$ es una fórmula atómica y por tanto una literal, pero no es aterrizada porque, aunque a es una constante, X es una variable; similarmente, $NOTp(X, a)$ es una literal pero no es atómica ni aterrizada. Por el contrario, las expresiones $p(a, b)$ y $NOTp(a, b)$ son literales aterrizadas.

- La *librería de planes* es el conjunto de planes π que el agente tiene como recetas o procedimientos (*recipes*). La función $execute(\pi)$ toma un plan como argumento y lo ejecuta. La librería de planes está incluida en las creencias, puesto que los agentes creen sus planes como procedimientos, mientras que las intenciones no están incluidas en las creencias.

- Las *intenciones* (*intentions* denotadas por I) son planes tomados de la librería e instanciados por contexto.

- Las *percepciones* (denotadas por ρ) son datos provenientes del ambiente. La percepción es normalmente empaquetada en conjuntos discretos llamados *perceptos*. La función $nextPercept$ regresa la siguiente percepción disponible para el agente. Por simplicidad, sólo se muestra en la Figura 2.6 como una entrada al agente ligada directamente a las creencias.

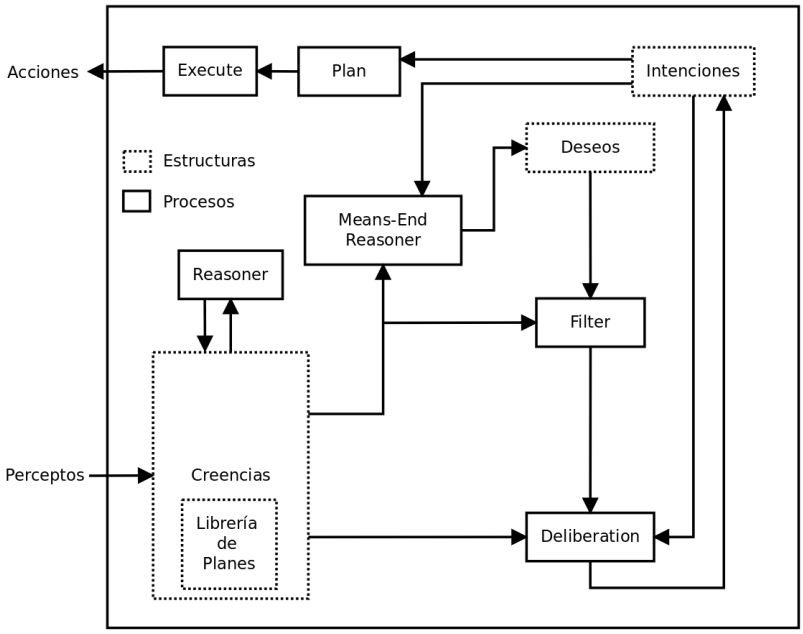


Figura 2.6: Una arquitectura para agentes basada en IRMA

En esta arquitectura el agente comienza su percepción con una función:

$$nextPercept : \wp(\rho) \longrightarrow \rho$$

y actualiza sus creencias con una función de revisión de creencias basada en su percepción y sus creencias actuales. Algunas veces esta función se conoce como razonador (*reasoner*) y tiene la siguiente forma:

$$reasoner : \wp(B) \times \rho \longrightarrow \wp(B)$$

El agente selecciona los planes relevantes usando un razonador de medios-fines (*meansEndsReasoner*) basado en las creencias y en las intenciones. Los planes seleccionados por esta función son los deseos del agente:

$$meansEndsReasoner : \wp(B) \times \wp(I) \longrightarrow \wp(D)$$

Los planes generados por el razonador de medios-fines son filtrados por el agente basado en sus creencias actuales, sus deseos y sus intenciones previas. La función filtro (*filter*) mantiene la consistencia y restringe las opciones que serán consideradas en la deliberación. El filtrado debe ser computacionalmente acotado, de forma que un proceso para detener el filtrado es necesario para equilibrar inercia y reconsideración:

$$filter : \wp(B) \times \wp(D) \times \wp(I) \longrightarrow \wp(D)$$

Finalmente una función de deliberación (*deliberation*) selecciona las opciones que serán incorporadas como intenciones, basada en razones creencia-deseo y la librería de planes:

$$deliberation : \wp(B) \times \wp(D) \times Planes \longrightarrow \wp(I)$$

Y dado que las intenciones están estructuradas como planes, una función *plan* es usada para seleccionar la función ejecutada:

$$plan : \wp(I) \longrightarrow P$$

Esta arquitectura fue la primera implementación de un sistema computacional de razonamiento práctico (Algoritmo 3).

Algoritmo 3: Agente intencional basado en IRMA

Datos : $creencias, intenciones, planes$
mientras $\top$ **hacer**
  $\rho \leftarrow nextPercept()$
  $creencias \leftarrow reasoner(creencias, \rho)$
  $deseos \leftarrow meansEndsReasoner(creencias, intenciones)$
  $deseos \leftarrow filter(creencias, deseos, intenciones)$
  $intenciones \leftarrow deliberation(creencias, deseos, planes)$
  $\pi \leftarrow plan(intenciones)$
  $execute(\pi)$
fin

2.5.1. El modelo BDI

Con estos elementos podemos dar una caracterización más simple del modelo BDI.[11] De manera general, una arquitectura BDI es un modelo, una idealización, cuyos componentes son las creencias, los deseos y las intenciones. Por tanto, un agente BDI es un sistema agente que puede describirse usando creencias, deseos e intenciones con sus interacciones (la arquitectura general BDI es la que se muestra en la Figura 2.7):

- Las *creencias* constituyen el fragmento *informativo* de un agente y representan la información que éste tiene del estado del ambiente y de su propio estado. Por lo general cada creencia es representada como una literal aterrizada (i.e., *grounded*) de primer orden *à la* Prolog. Las literales que no están aterrizadas son usadas en las definiciones de los planes. Además, las creencias son actualizadas por la percepción del agente y la ejecución de intenciones.

- Los *deseos* conforman la parte *motivacional* y representan sus metas. Los deseos incluyen alcanzar una meta o probar que una proposición es verdadera.

- Los *eventos* son mapeos de la percepción del agente que se mantienen en una cola. Los eventos incluyen la adquisición y eliminación de creencias, la recepción de mensajes en un sistema multi-agente y la adquisición de nuevas metas.

- Los *planes* están compuestos por un evento disparador que especifica cuándo un plan tiene que ser ejecutado, un contexto que determina si el plan puede ser ejecutado y un cuerpo que dicta los cursos de acción a tomar por dicho plan. Todo agente BDI tiene una librería de planes.

- Las *intenciones* conforman la porción *deliberativa* del agente y representan los cursos de acción con los que se ha comprometido un agente.

[11] Nos parece importante mencionar que si bien hay varios modelos de agencia racional, nos enfocamos en el modelo BDI porque además de sus ventajas técnicas [70], tiene una tradición filosófica que podemos rastrear con Searle [144], Dennett [51, 50] y, sobre todo, con Bratman [27], que es el principal referente en nuestro trabajo; posee también una sólida tradición formal en el campo de la ciencias de la computación y la inteligencia artificial como veremos más adelante [134] y hasta dentro del reino de los lenguajes de programación [19], siendo así, un excelente modelo cognitivo.

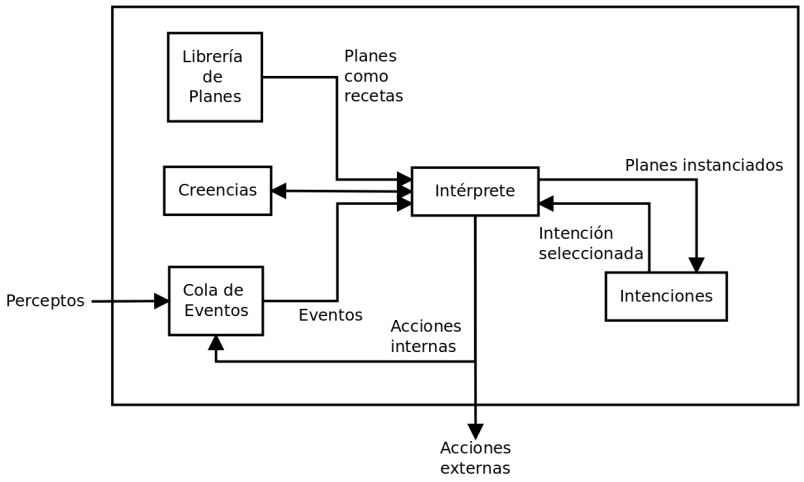


Figura 2.7: Arquitectura BDI

Un intérprete BDI (Algoritmo 4) ejecuta ciertas operaciones sobre estas estructuras de datos mediante la semántica de *pop*, *top* y *push*, que son operaciones usuales de pila; las definiciones de *aplicable* y *relevante* pueden encontrarse en [19], si bien las mencionaremos cuando sea el momento adecuado en el Capítulo 5; la función *sel* denota una función de selección.

Algoritmo 4: Algoritmo general BDI

Datos : planes,creencias,deseos
mientras $\top$ **hacer**
  $eventos \leftarrow percepción()$
  mientras $eventos \neq \emptyset$ **hacer**
    $ev \leftarrow pop(eventos)$
    $deseos \leftarrow relevante(aplicable(planes, creencias, ev)))$
    $int \leftarrow sel(deseos)$
    $intenciones \leftarrow push(int, intenciones)$
  fin
  $execute(top(\sigma_{int}(intenciones)))$
fin

Con esto tenemos la caracterización de un agente BDI como un sistema agente descrito en términos del modelo BDI. Pero además, como nos interesa desarrollar un modelo formal a partir de este modelo de agencia BDI nos falta ver qué podemos hacer para hablar formalmente de la agencia BDI.

ag	$::=$	$bs \quad ps$	h	$::=$	$h_1; \top \mid \top$
bs	$::=$	$b_1 \ldots b_n \ (n \geq 0)$	h_1	$::=$	$a \mid g \mid u \mid h_1; h_1$
ps	$::=$	$p_1 \ldots p_n \ (n \geq 1)$	at	$::=$	$P(t_1, \ldots, t_n) \ (n \geq 0)$
p	$::=$	$te : ctx \leftarrow h$	a	$::=$	$A(t_1, \ldots, t_n) \ (n \geq 0)$
te	$::=$	$+at \mid -at \mid +g \mid -g$	g	$::=$	$!at \mid ?at$
ctx	$::=$	$ctx_1 \mid \top$	u	$::=$	$+b \mid -b$
ctx_1	$::=$	$at \mid \neg at \mid ctx_1 \wedge ctx_1$			

Cuadro 2.4: Sintaxis de $AgentSpeak(L)$

2.6. $AgentSpeak(L)$

$AgentSpeak(L)$ [134, 136, 137] es un lenguaje formal usado para razonar sobre agentes BDI. Aquí lo usamos por varias razones. Por el momento basta decir que lo utilizamos de manera inercial porque es un lenguaje que ha sido utilizado con éxito para representar este tipo de agencia. Ahora exponemos brevemente su sintaxis y su semántica porque volveremos a ellas recurrentemente.

2.6.1. Sintaxis de $AgentSpeak(L)$

En $AgentSpeak(L)$ un agente ag está formado por un conjunto de planes ps y creencias bs (literales aterrizadas). Cada plan tiene la forma $te : ctx \leftarrow h$. El contexto ctx de un plan es una literal o conjunción de literales. Un cuerpo no vacío h de un plan es una secuencia finita de acciones $A(t_1, \ldots, t_n)$, metas g (de tipo *achieve* ! o *test* ? una fórmula atómica $P(t_1, \ldots, t_n)$), o actualizaciones de creencias u (adición $+$, borrado $-$). $\top$ denota elementos vacíos, por ejemplo, planes vacíos, contextos vacíos. Los eventos disparadores te son actualizaciones (adición o borrado) de creencias o metas. La sintaxis se muestra en el Cuadro 2.4.

2.6.2. Semántica de $AgentSpeak(L)$

La semántica operacional de $AgentSpeak(L)$ está definida por un sistema de transición, como se muestra en la Figura 2.8, entre estados o configuraciones agente $\langle ag, C, M, T, s \rangle$ donde:

- ag es un programa agente formado por creencias bs y planes ps.
- Una circunstancia C es una tripleta $\langle I, E, A \rangle$ donde I es el conjunto de intenciones $\{i, i', \ldots, n\}$ t.q. $i \in I$ es una pila de planes parcialmente instanciados $p \in ps$; E es un conjunto de eventos $\{\langle te, i \rangle, \langle te', i' \rangle, \ldots, n\}$, t.q. te es un evento disparador y cada i es una intención (evento interno)

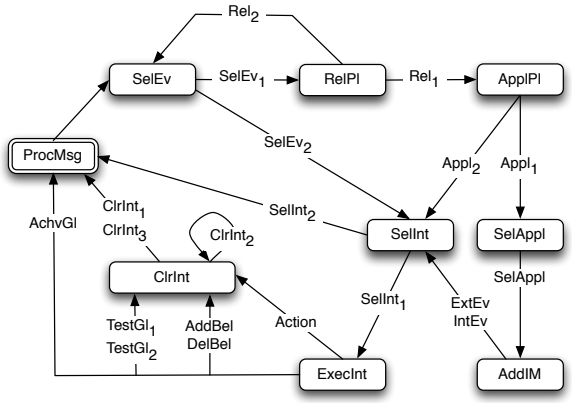


Figura 2.8: El intérprete para $AgentSpeak(L)$ como un sistema de transición

o una intención vacía $\top$ (evento externo); y A es un conjunto de acciones para ser ejecutadas por el agente en su ambiente.

- M es una tripleta $\langle In, Out, SI\rangle$ que funciona, para el caso multi-agente, como un buzón donde In es el buzón de entrada de un agente, Out es una lista de mensajes a entregarse y SI es un registro de intenciones suspendidas (intenciones que esperan por un mensaje de respuesta). Aunque este elemento es parte de la configuración de un agente en $AgentSpeak(L)$ no lo usaremos en el desarrollo del modelo y, por tanto, lo omitiremos para no complicar las nociones y las notaciones; lo mencionamos aquí, sin embargo, porque es parte de la estructura original de $AgentSpeak(L)$.

- T es una quíntupla $\langle R, Ap, \iota, \epsilon, \rho\rangle$ que registra información temporal: R es el conjunto de planes relevantes dado cierto evento disparador; Ap es el conjunto de planes aplicables (el subconjunto de R t.q. $bs \models ctx$); ι, ϵ y ρ registran, respectivamente, la intención, el evento y plan actual durante una ejecución.

- La etiqueta $s \in \{SelEv, RelPl, AppPl, SelAppl, SelInt, AddIM, ExecInt, ClrInt, ProcMsg\}$ indica el paso que está llevando a cabo el agente en un ciclo de razonamiento.[12]

[12]Por el momento simplemente señalaremos que estas funciones modelan el ciclo de razonamiento de los agentes. La elección de los nombres es más bien mnemotécnica: $SelEv$ es una función que selecciona un evento, $RelPl$ es una función que selecciona un plan relevante,

Usando esta semántica una corrida de un agente es:

$$Corrida = \{(\sigma_i, \sigma_j) | \Gamma \vdash \sigma_i \rightarrow \sigma_j\}$$

donde Γ es el sistema de transición definido por la semántica operacional de $AgentSpeak(L)$ y σ_i, σ_j son configuraciones agente.

2.7. Resumen

En este capítulo hemos expuesto el trasfondo y el marco teórico que seguimos como hilo conductor durante esta investigación argumentando que tiene un enfoque cognitivo. Bajo este enfoque argumentamos que un agente BDI es un sistema agente descrito en términos del modelo BDI. Esto es interesante porque el modelo BDI y la postura intencional nos permiten hacer generalizaciones sobre la conducta de diversos sistemas inteligentes—como nosotros—en términos BDI. Además, puesto que uno de nuestros objetivos conlleva la generación de un modelo formal en agentes BDI hemos mostrado la sintaxis y la semántica de $AgentSpeak(L)$.

$AppPl$ selecciona un plan aplicable, $SelInt$ selecciona una intención, $AddIM$ añade un mensaje, $ExecInt$ ejecuta una intención, $ClrInt$ limpia intenciones y $ProcMsg$ procesa mensajes. Esta semántica operacional será explicada a detalle en § 5.5.2.

Capítulo 3

Intenciones: revisión y no-monotonicidad

3.1. Introducción

En el capítulo anterior decíamos que estábamos interesados en modelar cierto aspecto de la conducta inteligente: el razonamiento intencional. En lugar de adentrarnos en ese tema comenzamos por mostrar una postura y una serie de proposiciones básicas.

Al presentar tales proposiciones tratamos de caracterizar sistemas agentes descritos en términos generales del modelo BDI desarrollado a partir de la teoría de Bratman. En este capítulo presentamos los detalles de ésta llevando a cabo el primer paso de nuestra estrategia: la lectura de los aspectos descriptivos y normativos de las intenciones, del razonamiento intencional y del cambio racional de intenciones según el modelo BDI en *Intention, plans, and practical reason* para observar las tesis, conceptos y argumentos en torno a la revisión de intenciones y la no-monotonicidad intencional.[1]

[1]Ciertamente la obra bratmaniana alrededor de este par de tópicos no es exclusiva de *Intention, plans, and practical reason*, ya que podemos encontrar argumentaciones similares en obras de semejante grado de importancia como *Faces of intention* [28] y *Structures of agency* [29]. Más todavía, podemos rastrear los orígenes de esta propuesta teórica en una serie de artículos presentados entre 1981 y 1985: *Intention and Means-End Reasoning* [23], *Taking plans seriously* [24], *Two faces of intention* [25] y *Davidson's Theory of Intention* [26]. Nosotros nos centramos en la primera obra principalmente por dos razones: primero, porque dicha obra es, históricamente, la fundamental para entender la generación y el desarrollo de las arquitecturas BDI tanto en filosofía como en IA; y segundo, porque en ella se exponen,

En particular argumentamos que tanto la revisión de intenciones como la no-monotonicidad del razonamiento intencional son atributos lógicos legítimos que pueden estudiarse filosófica y lógicamente. Puede pensarse que esto es una reivindicación lógica del fragmento I dentro del modelo BDI.

El capítulo está organizado de la siguiente manera. En § 3.2 hacemos un breve recuento de lo que parece ser el estado de la cuestión. En particular, esta sección es interesante porque queremos convencernos de que un modo diferente de racionalidad también incluye diferentes componentes de racionalidad. A partir de § 3.3 presentamos los detalles normativos y descriptivos de la postura de Bratman identificando cuatro supuestos.

Posteriormente desarrollamos la argumentación bratmaniana de la independencia del componente I dentro del modelo BDI y tratamos de mostrar que las intenciones, entendidas desde este modelo, permiten los atributos de no-monotonía y revisión. Finalmente, concluimos con un resumen de lo que entendemos informalmente por *modelo bratmaniano*, un concepto que juega un papel importante en todo el trabajo (§ 3.8).

3.2. Detalles preliminares

Cuando culpábamos a Platón en el Capítulo 1 lo hacíamos en tono de broma, pero la acusación no carecía de cierta solemnidad: la larga tradición filosófica que sugiere que para ser racional hay que ser deductivamente válido es bien conocida. Aunque podríamos encontrar críticas a esta posición mucho antes, es a partir de la década de 1960 que comenzó a ser cuestionada de manera no sólo explícita sino sistemática.[2]

3.2.1. Diferentes tipos de razones

Lo que causó este cuestionamiento sistemático fue el creciente reconocimiento de que muchos tipos de razonamiento que son coherentes, familiares y que tienen sentido no son deductivamente válidos. Para ejemplificar esto podemos considerar cuatro casos arquetípicos de razonamiento de sentido común: *i)* el

de manera unitaria y sistemática, las distinciones que nosotros consideramos pertinentes para definir lo que hemos denominado *modelo bratmaniano*.

[2]En efecto, a muy poca distancia cronológica del nacimiento y emergencia de la IA alrededor de 1956 [109] y de las ciencias cognitivas en 1977 [42], como lo hemos descrito previamente.

caso de la información perceptual, *ii)* la inducción cotidiana, *iii)* el razonamiento probabilístico y *iv)* el razonamiento temporal.

El primer caso puede ejemplificarse de forma sencilla: las conclusiones extraídas a partir de informaciones perceptuales pueden fallar. Para ilustrar esto podemos considerar el ejemplo, evidentemente de perfil cartesiano, de que podemos creer que la luna es de una tonalidad roja dado que la noche anterior la *vimos* de color rojo. Pero en realidad podría ser de un color diferente. Es, pues, muy sencillo concluir que la luna es roja, pero nueva evidencia nos podría conminar a retractar nuestra primera conclusión sin necesidad de caer en la irracionalidad. Contrario a lo que se podría pensar, en este caso lo irracional sería tener la nueva evidencia y no retractar la primera inferencia.

El caso de la inducción cotidiana es similar: frecuentemente concluimos argumentos con generalizaciones irrestrictas a partir de muestras restringidas. La proposición "todos los políticos son mentirosos" es una conclusión de este tipo. De manera similar al caso anterior, no es difícil notar que tal conclusión, ante nueva información, puede modificarse.[3] Así, como en el caso anterior, lo irracional no consistiría en concluir, con toda proporción guardada, generalizaciones; lo irracional sería tener nueva evidencia y no retractarse.

Ahora, cuando la inducción es tratada con más cuidado entramos al caso del razonamiento probabilístico. Este tipo de razonamiento es un tipo de razonamiento inductivo pero basado en generalizaciones probabilistas.[4] Así, por ejemplo, podemos concluir, con muy alta probabilidad, que nuestros hijos no serán altos, pero no lo creemos invariablemente. Podrían, tal vez por razones desconocidas, ser más altos de lo esperado (en [132] se discuten este tipo de ejemplos con más detalle).

El caso del razonamiento temporal es de especial importancia para nuestros fines. Cuando el ambiente es dinámico nuestro acceso a él es discrecional durante el tiempo, por lo que tenemos que hacer, constantemente, suposiciones *derrotables* acerca de la estabilidad del entorno. Esta idea de estabilidad tiene que ver,

[3] Creemos, con Pollock [132], que las críticas de origen humeano sobre el problema de la inducción [89], por poner un ejemplo paradigmático, nos debieron haber preparado para ser robustos y tolerantes ante nuevas formas legítimas de razonamiento y no para rechazarlas como irracionales o inválidas *tout court*. Consideramos además, con Aristóteles, que para este tipo de ejemplos es preciso recordar la distinción entre juicios *kathólou* y juicios *hòs epì tò polù* [9].

[4] Un ejemplo más formal de este tipo de razonamiento puede expresarse a través de *silogismos estadísticos* de la forma: si ϕ implica ψ con factor de certeza c (i.e., $p(\psi|\phi) = c$) y ϕ tiene factor de certeza c' ($p(\phi) = c'$) entonces ψ tiene factor de certeza $M(c, c')$, ($p(\psi) = M(c, c')$), donde M es una función apropiada como, por ejemplo, las definidas por las t-normas [82].

en principio, con la idea de que podemos, razonablemente, creer que el ambiente permanece en cierto estado durante los procesos de razonamiento.

Lo que podemos extraer de estos casos típicos es que los razonamientos de sentido común asumen de manera generalizada un mecanismo de tipo derrotable. Y por tanto, un sistema cognitivo sofisticado, como un agente BDI, no puede funcionar racionalmente sin considerar mecanismos de esta clase. Por supuesto, el problema filosófico-profundo y el técnico-ingenieril, haciendo uso de una clásica dicotomía de la que somos herederos por trabajar a caballo entre filosofía e IA, consisten en comprender, respectivamente, cómo son y cómo trabajan estos mecanismos.

3.2.2. Antecedentes históricos y teóricos

Siguiendo esta dicotomía como guía podemos ver, por un lado, que el estudio de la *derrotabilidad* comenzó en la filosofía, de manera muy clara y explícita, con el estudio de H.L.A. Hart sobre el razonamiento legal.[5] La adaptación e introducción de este concepto a la epistemología se lo debemos a Chisholm [36, 37] que fue seguido por Toulmin [157], Pollock [129] y Rescher [139].

Estas investigaciones de perfil filosófico tendían a estudiar los mecanismos de este tipo de razonamiento de manera muy esporádica y se centraban en el uso de ejemplos y observaciones para analizar argumentos filosóficos sofisticados, por lo que producían modelos abstractos muy profundos y que proveían cierta generalidad pero que estaban desligados de soluciones técnicas-ingenieriles.

Mientras tanto, casi de manera paralela, otro grupo de pensadores, del lado técnico e ingenieril, sin saber de la literatura filosófica redescubrieron esta problemática bajo la etiqueta de *no-monotonicidad*: desde el trabajo de McCarthy [111] hasta los aportes fundamentales de Reiter [138], McDermott y Doyle [112] en la década de 1980.

Este otro grupo de avances estuvo motivado, en buena medida, por mor de resolver el "problema del marco"[6] dado que el interés principal estaba en im-

[5] El término *defeasible*, cuya etimología viene del derecho medieval inglés sobre contratos, se refería a un contrato con una cláusula que podría deshacer el contrato en caso de que se cumplieran ciertas circunstancias. Fue introducido por H.L.A. Hart [84] en su sentido filosófico más general en el marco de la filosofía del derecho.

[6] McCarthy y Hayes definieron este problema conocido como el *frame problem* [111]. Posteriormente el término adquirió un sentido más amplio en filosofía, donde es formulado como el problema de limitar las creencias que se tienen que actualizar en respuesta a las acciones. El tratamiento de este problema desde dos áreas aparentemente dispares enriqueció el deba-

plementar soluciones a problemas típicos de razonamiento en sistemas de IA. Por esto mismo se preocuparon por los detalles más bien técnicos. Desafortunadamente, su falta de entrenamiento filosófico los condujo a generar teorías matemáticamente sofisticadas pero epistemológicamente subdesarrolladas pues, contraria o complementariamente al caso arriba descrito, no podían ser modelos generales de la cognición dado que no proveían suficiente generalidad.

Con estos antecedentes históricos y los ejemplos ilustrativos podemos identificar tres antecedentes teóricos claramente definidos. El primero tiene que ver con el cambio racional de información descrito por el interés en los cambios racionales en B. El segundo está vinculado con el problema de cómo relacionar especificaciones formales con implementaciones (el caso descrito por la relación histórica entre derrotabilidad y no-monotonicidad es relevante en este sentido). El tercero con la noción de inferencia y con la lógica, representado por los ejemplos típicos de razonamiento. Expondremos con más detalle en qué consisten estos antecedentes teóricos.

En primer lugar, no hay duda que dentro del modelo BDI el fragmento B ha sido históricamente el más estudiado: recientemente la revisión de creencias se ha convertido en un programa de investigación paradigmático, relativamente nuevo y que une las dos disciplinas con las que hemos comenzado este trabajo. Así, por ejemplo, desde que los programadores se las vieron con estructuras de información, como las bases de datos, se enfrentaron con el problema de la actualización [96]. Por otro lado, los filósofos se las vieron con el cambio de información dentro de estructuras epistémicas. De esta forma podemos identificar, respectivamente, dos momentos importantes en la historia de este programa de investigación: uno en el trabajo de Fagin, Ullman y Vardi [56] y el otro en las aportaciones de Harper [83] y Levi [103]; si bien la teoría que se encuentra en el trabajo fundamental de Alchourrón, Gärdenfors y Makinson (AGM) [1] constituye el núcleo de cualquier programa de revisión de creencias.

De esta forma, aunque el cambio racional de creencias bajo nueva información ha sido ampliamente estudiado durante los últimos 30 años, podemos apreciar que el proceso dinámico de otros estados mentales ha recibido menos atención y, en particular, hablamos de las intenciones [86]. Y, como habíamos mencionado, aunque hay teorías filosóficas y formales acerca de las intenciones, pocas de ellas consideran el cambio racional de intenciones como una posibilidad lógica

te durante las décadas de 1980 y 1990, pero dado que algunos problemas técnicos han sido resueltos, la discusión ha trasladado su importancia a otras discusiones. Así, por ejemplo, se ha argumentado que el problema del marco no es único del modelo *GOFAI*, sino que también está presente en modelos conexionistas.

legítima.

El siguiente antecedente es más palpable. La teoría de razonamiento práctico propuesta por Bratman expone los fundamentos filosóficos de los enfoques lógicos y computacionales de agencia racional conocidos bajo la etiqueta BDI. La teoría de Bratman en su momento resultó innovadora porque no reducía las intenciones a una combinación de creencias y deseos; antes bien, asumía que las intenciones eran elementos particulares compuestos de planes parciales y jerárquicos. Bajo esta teoría se desarrollaron diferentes sistemas lógicos BDI para caracterizar formalmente la conducta racional de los agentes en términos de las propiedades de los operadores BDI y sus mutuas relaciones.

Debido a su expresividad estos sistemas lógicos han sido usados para razonar acerca de las propiedades racionales de los agentes, pero dado su alto costo computacional, no son usados para programar agentes. Inversamente, lenguajes de programación de agentes, tales como $AgentSpeak(L)$, han sido propuestos para reducir la laguna entre la teoría (especificación lógica) y la práctica (implementación) de agentes racionales. Y aunque este lenguaje de programación, como veremos, tiene una semántica operacional bien definida, la verificación de propiedades racionales no es evidente, pues abandona las modalidades intencionales y temporales por mor de la eficiencia computacional. Este es un problema técnico que necesita ser solventado.

Finalmente, sabemos que la lógica trata del proceso inferencial de un conjunto de objetos a partir de alguna relación especial entre ellos. En términos generales ésta suele ser una relación binaria entre morfismos de signaturas. Cuando dicha relación cumple las propiedades especificadas, por ejemplo, por Meseguer [115] y Tarski [156] y los objetos son sentencias o proposiciones, decimos que la relación es de consecuencia lógica deductiva. Esta noción de derivación clásica, denotada por el signo $\vdash$, es una relación binaria que junto con un conjunto definido C de proposiciones constituye un par $\langle C, \vdash \rangle$ que configura un orden parcial y que además satisface las propiedades estructurales de monotonía, permutación, contracción y corte[7] (Cuadro 3.1).

Cuando una inferencia no sigue estas propiedades estructurales el reto usual consiste en proveer una descripción razonable de tal noción de inferencia y, como sucede en este caso, si la monotonía no es una propiedad del razonamiento

[7] *Monotonía* nos indica que el conjunto de consecuencias no decrece con el ingreso de nueva información: $\Gamma \vdash \phi \Rightarrow \Gamma \cup \{\psi\} \vdash \phi$. *Permutación* nos dice que el orden de las premisas no afecta la conclusión: $\Gamma \cup \Delta \vdash \phi \Rightarrow \Delta \cup \Gamma \vdash \phi$. *Contracción*, que información redundante puede obviarse sin perjudicar la conclusión: $\Gamma \cup \Delta \cap \Delta \vdash \phi \Rightarrow \Gamma \cup \Delta \vdash \phi$. *Corte* nos asegura que el tamaño de una deducción no importa: $\Gamma \vdash \Delta, \Gamma \cup \{\Delta\} \vdash \phi \Rightarrow \Gamma \vdash \phi$.

Propiedad	Nombre
$\frac{\Gamma \vdash \phi}{\Gamma, \psi \vdash \phi}$	Monotonía (atenuación)
$\frac{\Gamma, \Delta \vdash \phi}{\Delta, \Gamma \vdash \phi}$	Permutación
$\frac{\Gamma, \Delta, \Delta \vdash \phi}{\Gamma, \Delta \vdash \phi}$	Contracción
$\frac{\Gamma \vdash \Delta \quad \Gamma, \Delta \vdash \phi}{\Gamma \vdash \phi}$	Corte (transitividad cumulativa)

Cuadro 3.1: Propiedades estructurales de la relación de consecuencia

intencional, como veremos, y además queremos ofrecer una descripción adecuada de inferencia intencional, entonces debemos estudiar las propiedades metalógicas de la inferencia intencional que ocurren en lugar de la monotonía, porque una vez que ésta se abandona, ¿por qué deberíamos considerar la inferencia intencional como una instancia de una lógica *bona fide*?

3.2.3. Diversidad de componentes

Con esta exposición de antecedentes históricos y problemas teóricos es más sencillo hacer un diagnóstico del estado de la cuestión haciendo una paráfrasis de Morado en [116]: ciertamente, para ser racional hay que saber cuándo es pertinente identificar, evaluar y construir razones. Pero hay varias maneras de hacer todo esto.

En el modelo tradicional para que un agente haga todo esto y sea considerado como un agente racional debe *i)* inferir fuera de contexto, *ii)* tener recursos ilimitados, *iii)* ser lógicamente omnisciente y *iv)* ser infalible [116]. Sin embargo, la tesis de que un agente debe cumplir las cláusulas *i)-iv)* para ser racional no sólo deja a muchos agentes racionales fuera de una clasificación legítima, sino que es falsa: los antecedentes históricos, los casos típicos de razonamiento y los problemas teóricos lo confirman.

En efecto, se puede ser racional tomando en cuenta un poder inferencial contextual mientras mantengamos las reglas de inferencia contextuales bien definidas. Se puede ser racional con recursos limitados sin ser omniscientes, definiendo con claridad una racionalidad acotada [148]. Y finalmente, se puede ser

racional sin ser infalible mientras las inferencias gocen de ciertas condiciones de optimización.

Estas condiciones de optimización y racionalidad acotada permiten hallar máximos de plausibilidad con mínimos de incorrección para desarrollar un marco teórico que provea cierto rigor pues, contrariamente a lo que se pueda pensar, la pérdida de las cuatro características del modelo clásico no implica una pérdida de rigor total. Es posible desarrollar sistemas lógicos dentro de un marco teórico más generoso, pero no menos riguroso, de racionalidad, ya que es posible tener todo el rigor de un sistema clásico sin todos sus presupuestos idealizadores [116].

Para justificar la afirmación anterior basta considerar los ejemplos o casos típicos que hemos ilustrado en §3.2.1, los cuales nos muestran la necesidad de extender nuestros conceptos de comportamiento lógico con respecto a los criterios que usamos tradicionalmente para considerar a un agente como un agente racional. Los antecedentes históricos que hemos expuesto en §3.2.2 nos indican que la dirección que tomamos—la de *unir* filosofía e IA mediante relaciones lógicas y conceptuales—no carece de historia ni de sentido.

Podemos decir, además, que ésta es la línea de investigación oficial, poco novedosa; sin embargo, no hace falta ojo crítico para notar que, de nuevo, en todos estos ejemplos, en los antecedentes históricos y teóricos, e incluso en lo que va del resumen del estado de la cuestión, únicamente estamos hablando de una noción de racionalidad y comportamiento lógico alrededor de las creencias, es decir, otra vez en torno al fragmento B del modelo BDI.

No obstante, justamente, los problemas que hemos descrito nos sugieren que aún quedan problemas por resolver y que la tesis de que la noción de racionalidad es exclusiva del fragmento B es falsa. Admitimos, por tanto, que se puede ser racional sin cumplir las cuatro cláusulas del modelo clásico, pero añadimos que se puede serlo considerando otros componentes del modelo BDI. Proponemos, en consecuencia, desarrollar sistemas lógicos dentro de un marco BDI teóricamente más generoso pero no menos riguroso.

Podemos tener sistemas lógicos *stricto sensu* dentro de un modelo cognitivo de agencia BDI, como diría Morado, de "carne y hueso" [116]. Podemos seguir exigiendo la necesidad de un mecanismo lógico como un desiderátum al tiempo que tenemos una noción de racionalidad a nuestro alcance, pues la demanda de saber cuándo es pertinente identificar, evaluar y construir razones también se extiende a saber cuándo es pertinente identificar, evaluar y construir razones de carácter BDI.

3.3. Supuestos bratmanianos

Para ofrecer razones que den soporte a esta última tesis haremos una lectura de los aspectos descriptivos y normativos del modelo BDI a partir de cuatro supuestos bratmanianos [27]:

1. Las intenciones están ligadas al fenómeno de planificación.

2. Los agentes planifican a futuro.

3. Los agentes racionales planifican con cierta deliberación.

4. Una intención no es igual a un deseo.

De acuerdo al primer supuesto, uno de los elementos que permiten entender nuestra conducta y la de otros agentes es la noción de intención. Bratman sugiere que nuestra concepción natural de intención está íntimamente ligada al fenómeno de la planificación y a los planes mismos. Este supuesto, basta mencionar, tiene su origen en la crítica a la teoría de Davidson que no da lugar a las intenciones futuras para la coordinación interpersonal y para el uso de intenciones como entradas para razonamientos prácticos [26]. El desarrollo de este supuesto será importante porque nuestra descripción formal de las intenciones está estrechamente relacionada con ella.

La segunda suposición conlleva a una noción temporal. Como también argumentaremos, la naturaleza de las intenciones implica que los procesos de razonamiento poseen un carácter temporal. La justificación de este supuesto puede rastrearse en el hecho de que somos agentes extendidos durante el tiempo.

La tercera proposición nos servirá para justificar que hay una forma de razonamiento intencional. La justificación de este supuesto puede buscarse en nuestra necesidad de formar planes acerca del futuro y, dado que somos agentes racionales, para nosotros esto significa que la deliberación ayuda a dar forma a lo que hacemos. Si, sin embargo, nuestras acciones fueran influidas únicamente por la deliberación al momento de la acción, la influencia de tales deliberaciones sería mínima. Esto es así porque la deliberación requiere tiempo y otros recursos limitados y hay un obvio límite a lo que podemos deliberar exitosamente al momento de la acción. Por ello necesitamos asumir procesos de razonamiento temporal para influir en las acciones más allá del momento presente.

La última suposición nos sirve para hacer una distinción importante entre lo que es un modelo BD y lo que es un modelo BDI, lo que nos proporcionará un

parámetro para comparar y valorar los sistema formales, ya tradicionales, de Cohen y Levesque [41] y de Rao y Georgeff [134].

Asumiendo, entonces, que las intenciones están relacionadas íntimamente con el fenómeno de la planificación, y dado el supuesto 2, Bratman propone un trilema difícil de resolver.[8] Podemos apreciarlo con el siguiente ejemplo. Consideremos que el agente α tiene la intención, ϕ, de ir de compras el próximo fin de semana. Entonces α tiene una intención futura.[9] Pero ésta no es una suposición inocente. Una vez formada esta intención aparecen tres objeciones al fenómeno de la formación de tal intención:

- La *objeción metafísica* nos dice que cuando ϕ se forma no controla todas las acciones futuras, pues de ser así, ϕ implicaría acción a distancia: pero una cosa es compromiso y otra cosa es acción a distancia.

En efecto, cuando α tiene la intención de ir de compras el siguiente fin de semana, esta intención no controla todas las acciones futuras de tal manera que sea innecesario hacer cualquier otra cosa esperando que la intención se cumpla. En este sentido es famoso el problema de *Little Nell* [113]: *Little Nell* está amarrada en las vías del tren. Cuando el tren se acerca, su héroe forma la intención de rescatarla y, por lo tanto, ya no necesita hacer nada para salvarla. Si pudiéramos mapear esta idea informal a un vocabulario más formal diríamos que el compromiso, a diferencia de la acción a distancia, es una función no-monotónica.

- La *objeción racional* indica que una vez que ϕ se forma, no es preciso ni se sigue que ϕ no sea irrevocable: el mundo—el ambiente—es dinámico y los agentes no siempre anticipan el futuro del mismo.

Ciertamente, la intención de α de ir de compras el siguiente fin de semana no es irrevocable pues, como en el caso del ratón de Robert Burns, hasta el mejor plan para ir de compras es falible dado que el mundo es dinámico y no siempre es posible anticipar todos los posibles estados futuros del ambiente.

- Dadas las dos objeciones anteriores parece que ϕ debería formarse sólo si es racional para α formar ϕ, pero eso es inútil: si ese fuera el caso, no

[8] El propio Bratman detalla que los dos primeros cuernos del dilema fueron avistados por John Austin en 1873 en sus *Lectures on jurisprudence* [11].

[9] Una intención futura, a diferencia de una intención presente, es una intención de alcanzar o lograr cierto estado del mundo en el tiempo próximo: Brand distingue las intenciones prospectivas de las intenciones inmediatas [21].

tendría por qué haber planes a futuro, pero los hay: esto se conoce como *objeción pragmática*.

Así, si α debe tener la intención de ir de compras el próximo fin de semana únicamente si es racional para α formar dicha intención, entonces el proceso de planificación a futuro estaría configurado para evitar las primeras dos objeciones. Pero esto no es el caso.

Cuando buscamos la génesis de este rompecabezas y encontramos que está en la noción de intención futura, la sospecha en torno a la noción misma de intención futura emerge; y esta sospecha, una vez generada, promueve una de las posibles actitudes ante las intenciones futuras: el escepticismo. Este escepticismo de las intenciones futuras sugiere un enfoque alternativo a las intenciones en términos de cuatro tesis que algunas tradiciones en filosofía de la mente han propuesto:

- a) *La tesis de la prioridad metodológica de la intención en la acción.*

Esta es la tesis de que lo que hace que sea verdad que una acción sea intencional es una relación entre creencias e intenciones. Así, por ejemplo, lo que hace que la acción de comprar una nueva computadora sea intencional es mi deseo de jugar videojuegos de última generación y mi creencia de que con tal computadora eso es posible. Esta es, por ejemplo, la tesis propuesta por Anscombe en *Intention* [4]. Cuando esta tesis se hace más específica se convierte en:

- b) *La tesis de la relación creencia-deseo en la acción.*

Esta tesis nos dice que las acciones intencionales se definen como las compatibles con las creencias y los deseos del agente. De igual manera, podemos considerar a Anscombe como una representante de esta tesis. Pero quizás uno más representativo sería Davidson [48]. La diferencia entre Anscombe y Davidson en este punto estaría únicamente en que la primera rechaza que la relación entre creencias y deseos para llevar a cabo la acción sea causal.

Cuando se aceptan las dos tesis anteriores se recurre a:

- c) *La tesis de la extensión para las intenciones futuras.*

Porque se asume que si las tesis *a)* y *b)* son suficientes para explicar la intención presente, también lo serán para explicar el caso de las intenciones futuras. Lo cual nos conduce a:

- d) *La tesis de la reducción.*

La cual sostiene que la intención futura es explicada simplemente como un conjunto apropiado de creencias y deseos, es decir, que incluso una intención futura se reduce a una combinación de creencias y deseos. Audi [10] y Churchland [38] pueden considerarse como representantes de esta tesis.

3.4. El modelo BD

En principio, para resolver el trilema anterior, el modelo BD tiene dos aspectos: uno normativo, para encontrar qué hace que una acción sea intencional; y uno descriptivo, que concierne a cómo es que una acción es intencional. Normativamente, el modelo BD parece adecuado para explicar el papel de las creencias y los deseos en la formación de intenciones. Sin embargo, como apunta Bratman, hay dos problemas que no le permiten resolver el trilema: uno relacionado con el supuesto 2 y otro relacionado con el supuesto 4.

El primero, el *problema de las intenciones futuras* para el modelo BD consiste en que, si bien normativamente el modelo BD provee argumentos para justificar la formación de intenciones, descriptivamente no permite tratar con intenciones futuras, las cuales tienen un papel fundamental en las nociones de plan y compromiso. De acuerdo al modelo BD una intención futura sería el resultado de una combinación de creencias y deseos sobre un estado futuro, pero entonces, asumiendo esta posición, nos enfrentaríamos directamente a los cuernos del trilema.

El segundo problema es que la teoría BD reduce, descriptivamente, las intenciones a una relación entre creencias y deseos de tal suerte que *una intención es una combinación de creencias y deseos*. Sin embargo, esta reducción no puede ser correcta, dado el supuesto 4: los deseos, como las intenciones, son pro-actitudes, pero los deseos son potenciales influencias de la conducta, mientras que las intenciones son controladores efectivos y, por lo tanto, los deseos no implican compromiso. En otros términos, los deseos no explican la planificación a futuro: las intenciones sí.

Bratman argumenta, pues, que las 4 tesis anteriores, que constituyen la base del modelo BD, no proveen un marco teórico lo suficientemente poderoso como para tratar con las intenciones futuras, y por lo tanto no resuelven las objeciones del trilema. Además, sugiere que aceptar una psicología BD no da lugar a la actitud intencional e independiente de intentar hacer algo, pues con el marco

psicológico BD no podemos hacer justicia al fenómeno de la planificación, lo cual es así porque las 4 tesis anteriores son más apropiadas para entender agentes que no planifican, que agentes que sí lo hacen, como nosotros, por el supuesto 2.

Y si somos llevados a las tesis anteriores por las preocupaciones escépticas acerca de las intenciones futuras y es improbable que dichas tesis hagan justicia a nuestra concepción de nosotros mismos como agentes que planifican, entonces la otra actitud restante consiste en confrontar el escepticismo acerca de las intenciones futuras.

Por tanto, con el modelo BD el trilema no queda resuelto y entonces es preciso extenderlo: cuando Bratman extiende el modelo BD y confronta el escepticismo acerca de las intenciones futuras—y de allí viene la originalidad de su idea—nace el modelo BDI.

3.5. Extensión del modelo BD

La extensión de la teoría BD consiste en una extensión del sentido descriptivo de la misma, pero preservando parte de su sentido normativo. Para entender la extensión de este sentido descriptivo hay que seguir el supuesto 2 que dice que los agentes son agentes que planifican y que el comportamiento de esta planificación está delineado por un proceso de razonamiento (por el supuesto 3).

Este tipo de razonamiento, enfocado a realizar acciones basadas en lo que el agente cree y desea, tiene dos características importantes: por un lado, la deliberación y, por otro, el razonamiento medios-fines. El proceso de deliberación consiste en decidir qué metas debe alcanzar el agente y el razonamiento medios-fines consiste en la determinación de los medios para cumplir tales metas.

En estos procesos podemos encontrar un componente que no puede expresarse como una combinación adecuada de creencias y deseos porque tiene propiedades específicas: en términos algebraicos diríamos que es un componente extra que no puede expresarse como una combinación lineal de creencias y deseos, y que por lo tanto es independiente. Este componente singular, por tanto, puede identificarse y aislarse para explorar sus características.

A partir de esta extensión del modelo BD haremos dos extracciones de propiedades, una para la no-monotonicidad y otra para la revisión. Obtendremos ocho conjuntos de atributos del razonamiento intencional: estructurales, funcionales, descriptivos y normativos. Una vez expuestas estas propiedades las intenciones quedarán desmitificadas y aparecerán, justamente, a la par de las creencias y los deseos.

3.6. No-monotonicidad

3.6.1. Propiedades estructurales

Primero, al explorar las nociones originales del modelo BDI notamos que en la propuesta de Bratman hay una taxonomía de las intenciones que distingue tres tipos: deliberativas, no-deliberativas y basadas-en-políticas. Así, cuando un agente en t_1 intenta ϕ en t_2 a partir de una deliberación, la intención se llama *deliberativa*. Si el agente tiene una intención, no por una deliberación presente, sino porque el agente la tiene desde un momento previo t_0 y el agente ha mantenido esa intención de t_0 a t_1 sin reconsiderala, se llama *no-deliberativa*. Finalmente, cuando las intenciones son como políticas acerca de ciertas circunstancias particulares, se llaman *intenciones basadas-en-políticas* (IBP de aquí en adelante).

La importancia de las IBPs yace en sus propiedades estructurales que definen su comportamiento y su forma lógica: Bratman considera que las IBPs tienen la estructura de una regla y el comportamiento de un plan. Lo interesante de esta distinción es que, si bien podemos clasificar a las intenciones por su relación con el tiempo, las IBPs dictan la forma o estructura lógica general de las intenciones independientemente de su relación con el tiempo. Así pues, podemos inferir que para cualquier intención ϕ, sea deliberativa, no-deliberativa o basada-en-políticas, ϕ tiene la estructura lógica de una IBP, a saber, una estructura *regular*.

Esta propiedad es de gran relevancia, pues los modelos lógicos existentes que son usados para modelar razonamiento intencional están construidos en términos de lo que hemos llamado *modelo bratmaniano* y que a continuación precisaremos. Un modelo bratmaniano es un modelo que *i)* sigue las líneas generales de la teoría de razonamiento práctico de Bratman, *ii)* usa la arquitectura BDI para representar estructuras de datos y *iii)* configura una noción de consecuencia lógica a partir de las relaciones entre estados intencionales. Los sistemas lógicos existentes [41, 136, 165] que pretenden estar basados en modelos bratmanianos—revisaremos los dos sistemas más importantes en breve—, cometen dos errores que a continuación enunciaremos y que podemos resumir afirmando que cumplen con los requisitos *ii)* y *iii)*, pero omiten consecuencias del criterio *i)*.

Debido a que no se asume el criterio *i)*, el primer problema que encontramos es de tipo estructural: los sistemas lógicos tradicionales tienden a interpretar las intenciones como un operador modal único—usualmente representado por el signo INT—mientras Bratman sugiere, como hemos notado, que las intenciones tienen estructura de reglas (lo que nos permitirá construir un bosquejo más

detallado del razonamiento intencional); y funcionalmente, comportamiento de planes (por los supuestos 1 y 3).

El segundo problema es que estos sistemas no reconocen de manera simultánea que el razonamiento intencional tiene una naturaleza temporal y derrotable (por los supuestos 1, 2 y 3): el razonamiento intencional es temporal porque las intenciones—como las creencias—son estructuras temporales; pero también es derrotable porque si las estructuras son parciales y temporales, las consecuencias también pueden serlo (más adelante precisaremos detalles sobre la parcialidad y la temporalidad de las intenciones).

Por tanto, a partir de esta primera exploración podemos notar que las intenciones tienen una estructura lógica definida. Por el momento no detallaremos cómo es que es posible representar esto formalmente, pero la idea es que las intenciones tienen estructura lógica regular. Esta propiedad *estructural* nos garantiza que el estudio lógico del fragmento I dentro del modelo cognitivo BDI está justificado, al menos en principio.

3.6.2. Propiedades funcionales

En una segunda revisión podemos argumentar a favor de que las intenciones tienen propiedades que llamamos *funcionales*. Así, podemos notar que durante los procesos de razonamiento las intenciones, a diferencia de los deseos, no sólo motivan a alcanzar una meta, sino que dirigen la conducta a través de la noción de compromiso (por el supuesto 4). A esta propiedad funcional se le llama *proactividad*. Pero además las intenciones no sólo persisten sino que resisten, es decir, una vez tomadas se resisten a ser revocadas. Haciendo uso de un significado análogo, a esta propiedad se le llama *inercia*. Por otro lado, cuando las intenciones son tomadas restringen futuros razonamientos en el sentido de que mientras un agente mantiene una intención particular, no considerará opciones contradictorias y, por ende, proveen un *filtro de admisibilidad* para posibles razonamientos futuros.

Por tanto, las intenciones tienen una noción de compromiso, dado el principio de proactividad; una noción de retractabilidad, dada la propiedad de inercia; y tienen una noción de consistencia, por el criterio de admisibilidad.

Estas propiedades son importantes porque nos indican dos cosas: que las intenciones son componentes singulares que se distinguen de los fragmentos B y D (porque comparten con B la consistencia y la retractabilidad, y con D la proactividad; pero difieren de B por la proactividad, y de D por la consistencia y la retractabilidad), y que la derrotabilidad es una característica legítima del

razonamiento intencional.

3.6.3. Propiedades descriptivas

Sin embargo, este bosquejo aún está incompleto porque, por el supuesto 2, hace falta ver en qué sentido se entienden las intenciones en relación con la planificación. Con esto entramos a la tercera exploración con la que mostraremos propiedades que llamaremos *descriptivas*.

En primer lugar, puesto que las intenciones en tanto que IBPs se comportan como planes, se sigue que son parciales en el sentido de que son incompletas porque no es posible tener información completa del mundo. Esta propiedad se denomina *parcialidad* y tiene fundamento en la estructura del agente mismo y su acceso al ambiente.

Pero además de ser parciales, las intenciones no necesariamente son estructuras de información estática, pues el mundo es cambiante y, en consecuencia, es preciso tener información de los estados del mundo sólo temporalmente. En este sentido las intenciones son *dinámicamente temporales*.

Y por último, como las intenciones se organizan en razones medios-fines, tienen que estar reguladas por un criterio de orden, lo cual implica una *jerárquica* de intenciones. Junto con las propiedades funcionales, éstas nos indican que la naturaleza del razonamiento intencional no sólo implica nociones de no-monotonicidad sino también de temporalidad y orden.

3.6.4. Propiedades normativas

Finalmente, en una cuarta lectura podemos notar que las intenciones no están solas en el modelo BDI y, por lo tanto, tiene que haber cierta organización entre intenciones y el resto de los componentes del modelo. Primeramente, las intenciones deben ser ejecutables y a esto Bratman lo llama *consistencia interna*. Si no lo fueran no tendría sentido que fueran proactivas. Además deben ser consistentes con las creencias del agente, lo cual se conoce como *consistencia fuerte*.[10] Y el razonamiento medios-fines de las intenciones debe ser consistente

[10]Esta propiedad nos muestra que las creencias y las intenciones mantienen ciertas relaciones especiales. Bratman considera las siguientes relaciones como principios de racionalidad cuando un agente se enfrenta con razonamientos prácticos. Estas relaciones son las conocidas *tesis de asimetría*. La primera tesis de *inconsistencia intención-creencia* nos dice que es irracional para un agente intentar ϕ y creer al mismo tiempo que no hará ϕ. La segunda, la tesis de *incompletud intención-creencia* nos dice que es racional para un agente intentar ϕ pero no

con la red global de intenciones del agente. A esto último se le llama *coherencia medios-fines*. Estas propiedades *normativas* implican que la modificación de un conjunto de intenciones implica la modificación de un conjunto de creencias y viceversa.

Por lo tanto, a partir de estas cuatro exploraciones de la postura bratmaniana original podemos concluir provisionalmente algo muy importante: de acuerdo a un modelo bratmaniano las intenciones tienen propiedades que requieren—y justifican—una noción de derrotabilidad y temporalidad (Cuadro 3.2).

Propiedades	**Descripción**	**Justifican**
Estructurales	Las intenciones tienen estructura lógica *regular*	Análisis lógico
Funcionales	Proactividad, inercia, admisibilidad	No-monotonía
Descriptivas	Parcialidad, dinamismo temporal, jerarquía	Temporalidad
Normativas	Consistencia interna, consistencia fuerte, coherencia medios-fines	*Revisión*

Cuadro 3.2: Características informales de un modelo bratmaniano I

3.7. Revisión

En la sección inmediatamente anterior hemos visto un conjunto de propiedades relacionadas con el título de este Capítulo: la no-monotonicidad (Cuadro 3.2). Por el momento dejaremos ese tópico para exponer propiedades del otro término que también tiene lugar en el título: la revisión. Haremos tres exploraciones para argumentar la justificación de la revisión de intenciones.

De manera similar a la sección anterior, para las intenciones podemos definir ciertas propiedades que conciernen a la revisión de intenciones.[11] Aquí trataremos tres dimensiones: una de propiedades descriptivas, para extraer los tipos y los aspectos de la reconsideración; otra de propiedades funcionales acerca del comportamiento de la reconsideración; y, por último, una sobre propiedades normativas para inferir cuándo es racional reconsiderar intenciones.

---

creer que logrará ϕ.

[11]Bratman no es muy claro para hacer la distinción entre revisión y reconsideración. En algunos momentos parece considerar a la revisión como un caso más general que incluye tanto a los procesos de reconsideración como a los de no-reconsideración; en otros, reconsideración y revisión son usados como sinónimos. Aquí seguimos esta segunda interpretación para ganar uniformidad sin pérdida de generalidad.

3.7.1. Propiedades descriptivas

Descriptivamente, en una primera exploración, podemos ver que Bratman sugiere una taxonomía de la reconsideración donde distingue tres tipos de reconsideración asociados con los tres tipos de intenciones. La *reconsideración no-reflexiva* es una reconsideración sin deliberación. La *reconsideración deliberativa* es una reconsideración en la que se evalúan las creencias y los deseos durante la propia reconsideración. Y la *reconsideración basada-en-políticas* es una reconsideración que apela a una regla o política general sobre cuándo reconsiderar (Cuadro 3.3).

Intención	Reconsideración
No-deliberativa	No-reflexiva
Deliberativa	Deliberativa
Basada-en-políticas	Basada-en-políticas

Cuadro 3.3: Taxonomía de intenciones y tipos de reconsideración

Análogamente a las intenciones, una reconsideración, bajo cualquiera de estos tipos, no consiste en meramente entretenerse con la posibilidad de cambiar dicha intención, sino en considerar si, dada otra información, la intención realmente ha de continuar (a esto lo llamamos *preservación*) o debe abandonarse (a esto lo llamamos *abandono*). Por ello Bratman sugiere dos aspectos funcionales de la reconsideración: la espera y el costo.

3.7.2. Propiedades funcionales

Una reconsideración implica poner en *espera* a la intención porque dentro de una reconsideración ocurren dos tipos de procesos: procesos de cambios de razones (*reason-changing*) para evaluar si la intención se debe cambiar; y procesos de preservación de razones (*reason-preserving*) para evaluar si la intención debe continuar. Además una reconsideración implica una revisión de los planes mismos, y esto implica ciertos *costos*, en este caso en términos de tiempo, de acuerdo a la jerarquía de los planes y a su alcance en razonamientos futuros.

De estas exploraciones podemos extraer dos propiedades de la revisión: una descriptiva, que nos dice que la revisión implica procesos de adopción o abandono; y una funcional, que nos dice que la revisión funciona bajo criterios tradicionales de optimización como la espera y el costo. Como puede inferirse, una reconsideración basada en políticas provee un buen balance, en términos de optimización, entre espera y costo.

3.7.3. Propiedades normativas

Ahora, en una tercera lectura, podemos apreciar que normativamente la cuestión de la reconsideración tiene que ver con cuándo es racional para un agente reconsiderar una intención. Con respecto a la reconsideración no-reflexiva podemos volver al hecho de que una intención es estable (lo cual está garantizado por la inercia), pero tampoco es irrevocable. Entonces hay que notar en qué se basa esta estabilidad.

Esta estabilidad o firmeza de la intención está relacionada con las atenciones del agente, y estas atenciones pueden ser sólo a ciertas cosas. La racionalidad de la reconsideración se basa en una manifestación de los hábitos generales que un agente tiene para su reconsideración. Pero entonces surge la pregunta: qué hábitos son razonables. La respuesta es que son razonables los hábitos relacionados con los siguientes aspectos.

Por un lado, cuando hay *problemas para los planes* es preciso reconsiderar. Un problema para un plan puede ser de alguno de los siguientes tres tipos: *i)* el mundo esperado puede ser diferente al mundo actual; *ii)* los deseos pueden cambiar y *iii)* las propias intenciones pueden cambiar. Por otro, si hay oportunidad y recursos suficientes para llevar a cabo una reconsideración, es razonable reconsiderar la intención para tener el hábito de la *reconsideración ocasional*.

De esta segunda exploración extraemos propiedades normativas que indican que la revisión de intenciones es un proceso legítimo donde se presentan procesos de preservación (adopción), abandono y cambio de intenciones, que muestran su similitud con el repertorio de expansión, contracción y revisión del tan mencionado fragmento B del modelo AGM (Cuadro 3.4).

Propiedades	**Descripción**	**Justifican**
Descriptivas	Preservación, abandono	Revisión
Funcionales	Costo, espera	Temporalidad
Normativas	Problemas para los planes	*No-monotonía*

Cuadro 3.4: Características informales de un modelo bratmaniano II

3.7.4. Compromiso

Como hemos expuesto previamente, una de las propiedades distintivas de las intenciones es el compromiso. Siguiendo la línea expositiva, el compromiso también tiene dos aspectos: uno normativo y uno descriptivo. Descriptivamente, hay

que notar cuál es el papel de las intenciones para conectarse con futuras deliberaciones. Bratman distingue dos niveles de compromiso: el *nivel volitivo* consiste en el papel característico de la intención presente. En términos ejemplares, en la dimensión volitiva si el agente intenta ϕ ahora, normalmente hará ϕ ahora. El *nivel centrado-en-razones* consiste en el papel característico de las intenciones futuras. Para las intenciones futuras el aspecto centrado-en-razones tiene un papel fundamental en el ínterin entre la formación de la intención y su posterior ejecución.

Normativamente las intenciones tienen una dimensión que concierne a las normas o reglas de racionalidad y se distinguen dos normas: *normas internas* a las razones del agente que exigen una consistencia interna, consistencia fuerte y coherencia medios-fines. Y *normas externas* a las razones prácticas del agente que consisten en la racionalidad de la reconsideración no-reflexiva y el principio de intención-acción.[12]

Con estas exploraciones de la postura bratmaniana original en torno a la reconsideración de intenciones y al compromiso como elemento característico de las intenciones, podemos concluir provisionalmente que las intenciones, en efecto, tienen propiedades estructurales, descriptivas, funcionales y normativas que requieren—y justifican—los atributos lógicos de no-monotonicidad y revisión del razonamiento intencional.

3.8. Modelos bratmanianos

Previamente hemos caracterizado a un modelo bratmaniano como un modelo que *i)* sigue las líneas generales de la teoría de razonamiento práctico de Bratman, *ii)* usa la arquitectura BDI para representar estructuras de datos y *iii)* configura una noción de consecuencia lógica a partir de las relaciones entre estados intencionales.

En este contexto, seguir las líneas generales de la teoría de razonamiento

[12]Las intenciones, por su proactividad, son componentes controladores de la conducta: en el curso normal de acciones, si un agente racional intenta ϕ al menos tratará de cumplir ϕ. Esto sugiere el siguiente principio: si es racional para un agente intentar ϕ, y el agente ejecuta con éxito esta intención y por ende hace ϕ intencionalmente, entonces es racional para el agente intentar ϕ. La relevancia de este principio es doble. Por un lado, muestra que la intención presente y el resultado de esa intención están conectados: y esto es así porque la intención y la acción no están separadas por el agente. Al contrario, el agente tiene control de sus acciones a través de las intenciones. Por otro lado, este principio no es algo interno a los razonamientos prácticos del agente, sino más bien una norma externa para definir la racionalidad del agente.

práctico de Bratman implica seguir las propiedades del modelo BDI que definen el comportamiento lógico de los agentes BDI. Hacerlo de otro modo es construir un modelo útil o interesante, pero no bratmaniano y, por tanto, incapaz de resolver los problemas externo e interno. Resumiremos ahora algunas de las implicaturas de la relación entre un modelo bratmaniano y los problemas externo (revisión) e interno (no-monotonicidad).

3.8.1. Perspectiva externa

Quizá la mejor manera de caracterizar esta perspectiva sea mediante los postulados de Cohen y Levesque [41]—que, por cierto, se derivan de un análisis conceptual de Bratman en *Intention, plans, and practical reason*:

1. Las intenciones se presentan como problemas a los agentes, quienes tienen que determinar modos de alcanzarlas.

2. Las intenciones proveen un filtro para adoptar otras intenciones que no deben entrar en conflicto.

3. Los agentes rastrean el éxito de sus intenciones y están inclinados a reintentar si sus intentos fallan.

4. Los agentes creen que sus intenciones son posibles.

5. Los agentes no creen que no alcanzarán sus intenciones.

6. Bajo ciertas circunstancias los agentes creen que cumplirán sus intenciones.

7. Los agentes no necesitan intentar todos los efectos colaterales de sus intenciones.

Con estos criterios bratmanianos Cohen y Levesque construyeron una teoría formal de la intención usando la noción de meta persistente.[13] Sin embargo, como haremos notar, hay un par de objeciones a este planteamiento. Por un lado, éste no trata con claridad la dinámica de las intenciones [149] (veremos esto con más detalle en el próximo capítulo); y por otro, no distingue con claridad las propiedades de las intenciones que desplegamos renglones arriba.

[13] De acuerdo a su formalismo una intención es una meta persistente [41], como veremos en el siguiente capítulo.

La dinámica de las intenciones trata con el problema de cómo un agente adopta (expande) o abandona (contrae) intenciones y qué cambios producen estos procesos en el resto de los componentes del modelo BDI. Pero este tipo de cambios están relacionados con las nociones de un tipo de revisión. La dinámica de las intenciones, pues, requiere una teoría de revisión de intenciones, de la misma manera en que los cambios de creencias han requerido una teoría de revisión de creencias. Por esta razón añadimos, a partir de nuestra lectura de la revisión de intenciones, el siguiente par de postulados:

8. Los agentes pueden contraer intenciones.

9. Los agentes pueden expander intenciones.

Esto nos permite afirmar que los agentes pueden revisar intenciones y constituye lo que debe entenderse por revisión de intenciones en un modelo bratmaniano. La pregunta ahora es: ¿Cómo podemos representar esto formalmente? Cuando llegue el momento de enfrentar este problema entraremos en el terreno del aprendizaje, por el momento proponemos un ejemplo que ilustra la revisión.

Ejemplo 1 *Supongamos que un agente está inmerso en un ambiente innacesible, no-determinista, episódico, discreto y dinámico. Además, supongamos que tal agente tiene ciertas creencias e intenciones (representadas por el estado α) y que, en algún momento, desea alcanzar un nuevo estado del mundo (representado por el estado β)—representamos esta situación con la flecha negra (Figura 3.1).*

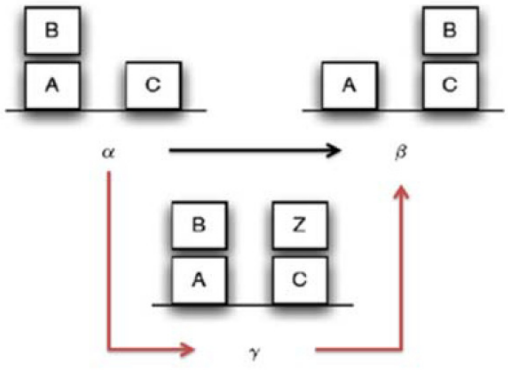


Figura 3.1: Estados de un agente

Lo que este ejemplo muestra es que el agente genera una intención de la forma $coloca(B, C)$. Ahora, dadas las propiedades del ambiente, supongamos que el agente percibe el estado γ—denotado por la flecha roja—, mientras intenta β, donde no es el caso que $libre(C)$. Por tanto, $coloca(B, C)$ fallará y el conjunto de intenciones se hará inconsistente: para evitar la *inconsistencia intención-creencia* es preciso reconsiderar.

3.8.2. Perspectiva interna

Desde una perspectiva interna hemos notado que las intenciones tienen una estructura lógica bien definida, funcionan bajo principios de proactividad, inercia y retractabilidad; se comportan como planes parcialmente instanciados, temporales y ordenados bajo ciertos criterios de consistencia y coherencia. La pregunta es, como en la perspectiva externa: ¿Cómo podemos representar esto formalmente? Cuando llegue el momento entraremos en el terreno de los sistemas formales, por el momento veamos un ejemplo.

Ejemplo 2 *Consideremos un escenario en el que un agente intenta adquirir su doctorado. Bajo un enfoque tradicional, si un agente intenta adquirir su doctorado y hay una regla*

$$Doctorado \Rightarrow Examen$$

entonces se sigue que el agente intentará hacer el examen.

No es difícil notar que este esquema de razonamiento luce familiar. En bases de conocimiento se suele denominar *omnisciencia lógica*. En el ámbito de las intenciones se denomina *efecto colateral* (volveremos a este tema más adelante). El problema de este esquema es que no nos permite distinguir entre intenciones que son mantenidas típica pero no absolutamente y constituye una aproximación BDI monotónica y, por tanto, no bratmaniana.

Todo esto nos muestra la importancia de tomar seriamente un modelo bratmaniano: el análisis y el tratamiento detallado de un modelo bratmaniano nos provee argumentos, herramientas y conceptos para comprender los problemas externo e interno; desafortunadamente, no nos provee de los formalismos adecuados para resolverlos: enfocaremos nuestra atención a este problema en los siguientes capítulos.

3.9. Resumen

Llevamos a cabo un examen del estado de la cuestion y una lectura de los aspectos descriptivos y normativos del razonamiento con intenciones según el modelo BDI de Bratman en *Intention, plans, and practical reason* para observar los problemas de la revisión y la no-monotonicidad intencional. En particular observamos que tanto la revisión de intenciones como la no-monotonicidad del razonamiento intencional son atributos lógicos legítimos que pueden estudiarse filosófica y lógicamente, especialmente dentro de un marco BDI más generoso, pero no menos riguroso.

Así, hasta este momento, tenemos que desde un punto de vista cognitivo un agente BDI es un sistema agente que se explica en términos del modelo BDI bratmaniano. Y hemos argumentado que un modelo bratmaniano de razonamiento intencional debe considerar los atributos lógicos de revisión de intenciones y no-monotonicidad del razonamiento intencional. El problema que queda abierto es el de cómo traducir los detalles de este análisis a un modelo formal.

Capítulo 4

Especificaciones formales

4.1. Introducción

En el capítulo anterior dejamos abierto el problema de cómo traducir los detalles del análisis filosófico de un modelo bratmaniano a un modelo formal. Este es un problema interesante por dos razones. Primero, porque como decíamos en el Capítulo 2, uno de los elementos que componen este trabajo tiene que ver con los sistemas formales y la especificación formal. Segundo, porque además de los problemas filosóficos, hay problemas de carácter técnico que necesitaremos resolver cuando llegue el momento.

Por lo pronto sabemos que existen diversos sistemas lógicos que han sido propuestos para especificar y verificar propiedades formales de sistemas agentes a partir de un modelo bratmaniano. En este capítulo pasamos revista al trabajo fundacional de Cohen y Levesque [41], aunque le dedicaremos más tiempo a la propuesta de Rao y Georgeff [134] debido a su poder expresivo. Revisaremos estos dos enfoques formales porque además de ser considerados como "clásicos", son fundamentales por su influencia en modelos posteriores.

La idea de este Capítulo es la siguiente: ahora que tenemos un análisis de las propiedades de un modelo BDI bratmaniano veremos que estos modelos formales tradicionales no respetan dos propiedades fundamentales del modelo bratmaniano de manera simultánea. Un modelo bratmaniano fidedigno conlleva ciertas características de temporalidad y no-monotonicidad. Pero los sistemas tradicionales no consideran los aspectos temporales y no-monotónicos de dicho modelo. De manera más específica, estos sistemas son intencional-temporales pero monotóni-

cos.[1]

Comenzamos por exponer y valorar la teoría de Cohen y Levesque (§ 4.2). Posteriormente revisamos el formalismo BDI (§ 4.3) y la axiomatización de sus componentes (§ 4.4), los tipos de realismo (§ 4.5) y el compromiso (§ 4.6). Como veremos, la propuesta de Rao y Georgeff nos resultará más útil y menos problemática—aunque no suficiente—para el desarrollo formal de nuestro modelo bratmaniano.

4.2. $C\&L$

La teoría de Cohen y Levesque ($C\&L$ de ahora en adelante) es importante por la complejidad de su análisis y porque fue el primer intento de formalizar el concepto de intención[2] siguiendo la teoría de Bratman [41]. A continuación exponemos la sintaxis y la semántica de $C\&L$.

4.2.1. Sintaxis de $C\&L$

Siguiendo la teoría de Bratman, como uno de los postulados de $C\&L$ está la consistencia mutua de las metas y la consistencia de éstas con las creencias del agente: esto, sin duda, es una formalización que se adhiere materialmente al análisis funcional que hicimos en el Capítulo 3. $C\&L$, además, comienza con la noción de meta ($GOAL$) como una primitiva del lenguaje y define una meta persistente ($P-GOAL$ por *Persistent goal*) como una meta en la que el agente persiste hasta que cree que se ha cumplido o cree que es imposible. De este modo, en $C\&L$ las intenciones son entendidas como metas persistentes.

Los componentes esenciales de $C\&L$ son un modelo M que incluye un conjunto de secuencias lineales de mundos posibles (es decir, una función que va de los enteros a los mundos posibles [88]); un dominio de discurso; una función v que une variables a objetos del dominio y Φ que interpreta los predicados de diferentes cursos de eventos en diferentes índices temporales.

[1] El caso de las lógicas derrotables, como veremos más adelante, es el complementario: son no-monotónicas pero no intencional-temporales.

[2] Si bien $C\&L$ es la primera teoría que formaliza las intenciones de manera directa, el trabajo de Allen [3] es un antecedente formal en donde las intenciones se formalizan, indirectamente, como creencias sobre estados futuros.

4.2.2. Semántica de $C\&L$

La semántica, entonces, se define dado un modelo M, una secuencia de mundos posible σ, una función de valuación v y un índice temporal n tales que $M, \sigma, v, n \models \phi$ indica que ϕ es satisfecha en tal punto en tal modelo bajo la función de valuación v. Como es usual, $\models \phi$ significa que ϕ es lógicamente válida. A continuación mostramos la semántica de $C\&L$ [81]:

- $M, \sigma, v, n \models P(t_1, \ldots, t_n) \Leftrightarrow (v(t_1), \ldots, v(t_n)) \in \Phi[P, \sigma, v]$
- Las definiciones de $\neg$, $\vee$, $\exists$ y $\equiv$ se definen de manera estándar [124].
- $M, \sigma, v, n \models (HAPPENS\ \alpha) \Leftrightarrow \exists m, m \geq n$ t.q. $M, \sigma, v, n\,[[\alpha]]\,m$ e.d., α describe una secuencia de eventos que ocurre al siguiente momento.
- $M, \sigma, v, n \models (DONE\ \alpha) \Leftrightarrow \exists m, m \leq n$ t.q. $M, \sigma, v, m\,[[\alpha]]\,|n$ e.d., α describe una secuencia de eventos que acaba de ocurrir.

$DONE\ x\ \alpha$ y $HAPPENS\ x\ \alpha$ significan que x es el agente que realiza $DONE\ \alpha$ y $HAPPENS\ \alpha$. La semántica de la relación de creencia (B) es euclidiana, transitiva y serial,[3] mientras que la relación de meta (G) es serial. Se asume que $G \subseteq B$.

- $M, \sigma, v, n \models (BELIEF\ x\ \alpha) \Leftrightarrow \forall \sigma*$ t.q. $\langle \sigma, n \rangle\, B[v(x)]\sigma*, M, \sigma*, v, n \models \alpha$. Es decir, que α es verdadera en todos los mundos accesibles con B en σ y n.
- $M, \sigma, v, n \models (GOAL\ x\ \alpha) \Leftrightarrow \forall \sigma*$ t.q. $\langle \sigma, n \rangle\, G[v(x)]\sigma*, M, \sigma*, v, n \models \alpha$. Es decir, que α es verdadera en todos los mundos accesibles con G en σ y n.

La semántica de las acciones está dada en términos de [[ ]], que denota que las acciones dadas toman lugar en tal intervalo:

- $M, \sigma, v, n\,[[e]]\,n+m \Leftrightarrow v(e) = e_1, \ldots, e_m$ y $\sigma(n+i) = e_i, 1 \leq i \leq m$. La secuencia de eventos denotados por e ocurren de n a m en σ.

[3] La relación euclidiana, como veremos más adelante (Cuadro 4.1), es codificada por el axioma S5; la relación transitiva por S4, y la serial por D.

- $M, \sigma, v, n\,[[\alpha; \beta]]\, n + m \Leftrightarrow \exists k, n \leq k \leq m$ t.q. $M, \sigma, v, n\,[[\alpha]]\, k$ y $M, \sigma, v, k\,[[\beta]]\, m$. Primero ocurre α después de n y posteriormente ocurre β terminando en m.

- $M, \sigma, v, n\,[[\alpha?]] \Leftrightarrow M, \sigma, v, n \models \alpha$ $\alpha?$ ocurre si α es verdadera, y falla si es falsa.

- $(BEFORE\ p\ q) =_{def} \forall c(HAPPENS\ c; q?) \Rightarrow \exists a(a \leq c) \wedge (HAPPENS\ \alpha; p?)$. Si q ocurre en el futuro, p ocurre antes que q, es decir, o bien p se da en el futuro y $\neg q$ se mantiene hasta entonces, o bien ni p ni q se dan.

- $(KNOW\ p) =_{def} p \wedge (BELIEF\ p)$. El conocimiento es una creencia verdadera.

- $(COMPETENT\ p) =_{def} ((BELIEF\ p) \Rightarrow (KNOW\ p))$. Un agente es competente acerca de p si y sólo si cree que p, sabe que p.

- $\diamond\alpha =_{def} \exists x(HAPPENS\ x; \alpha?)$. α es en algún momento verdadera si es verdadera después de una secuencia de eventos.

- $L\alpha =_{def} \neg \diamond \neg\alpha$

- $(LATER\, p) =_{def} \neg p \wedge \diamond p$. Esto es que $(LATER\, p)$ es verdadera si y sólo si p es verdadera en el futuro estricto.

- $(P - GOAL\ x\ p) =_{def} (GOAL\ x\ (LATER\ p)) \wedge (BELIEF\ x\ \neg p) \wedge [BEFORE((BELIEF\ x\ p) \vee (BELIEF\ x\ L\neg p))\ \neg(GOAL\ x\ (LATER\ p)]$. Un meta persistente es una proposición p t.q. *i)* es una meta del agente que p sea verdadera en el futuro estricto, *ii)* el agente no cree p ahora y *iii)* la primera condición se mantiene a menos que el agente llegue a creer que p es verdadera o que p es imposible.

- $(INTEND_1\ x\ \alpha) =_{def}$
  $(P - GOAL\, x[DONE\, x(BELIEF\, x\, (HAPPENS\, \alpha))?; \alpha])$. Un agente intenta una acción α si y sólo si tiene la meta persistente de hacer α inmediatamente después de creer que iba a suceder. Las intenciones, por tanto, son un tipo de metas persistentes y, en consecuencia, son una combinación de creencias y deseos.

A pesar de su utilidad, un análisis crítico de esta propuesta formal muestra algunos problemas. En primer lugar, de acuerdo con la teoría de Bratman, el compromiso de los agentes es meramente una condición que se mantiene en circunstancias normales. Esto significa que los agentes normalmente persisten en sus intenciones. En efecto, como decíamos renglones arriba, las intenciones tienen una inercia. Esto, por supuesto, sólo es razonable si se toma en cuenta que las intenciones tienen un papel que no se reduce a una combinación de creencias y deseos. Sin embargo, como hemos descrito, $C\&L$ mezcla la persistencia de los agentes en la semántica de las intenciones, y puesto que la intención se ve expresada como una combinación lineal de creencias y deseos, $C\&L$ es, en términos bratmanianos, la formalización de la *tesis de la reducción*.

El segundo problema es que $C\&L$ captura muy bien una noción de inferencia BD pero de manera monotónica. Debido a estos problemas recurrimos al formalismo BDI_{CTL} de Rao y Georgeff en la siguiente sección para posteriormente mostrar cómo aproximarnos a una teoría intencional que nos permita resolver estos problemas.

Finalmente, es justo mencionar que otros problemas han sido identificados previamente por Singh [149] y Sadek [142]. El primero cuestiona la complejidad operacional de $C\&L$ así como la falta de distinción entre la semántica de una intención y las políticas de revisión de una intención. El segundo apela a que el modelo $C\&L$ no da cuenta de la introspección de los agentes, por lo que no puede usarse como un modelo para hablar de la autonomía de los agentes. Rao y Georgeff, por otra parte, cuestionan la suposición de que $G \subseteq B$, pues si bien ésta captura algunos aspectos de las tesis de asimetría, también implica que cualquier creencia sobre un estado futuro se convierte en una meta.

4.3. BDI_{CTL} y BDI_{CTL*}

La lógica multimodal BDI para representar creencias, deseos e intenciones se basa en operadores modales. Si bien el trabajo fundacional sobre este tipo de formalismos se debe a Cohen y Levesque (como vimos en la sección anterior), ellos optan por una teoría de la intención reducida a una combinación de creencias y deseos. Una lógica multimodal BDI, en contraste, trata al fragmento I como un elemento que no se reduce a una combinación de B y D, lo cual es un requisito para formalizar un modelo bratmaniano. Además, el componente temporal en esta clase de sistemas está definido a través de lógicas computacionales líneales como LTL o arborescentes como CTL y $CTL*$ [54] para representar y razonar

sobre aspectos temporales con gran poder expresivo.

Como veremos cerca del final de este capítulo, las lógicas BDI pueden incluir un componente de acción para representar los eventos registrados por los agentes y sus acciones. Este componente se basa normalmente en la lógica dinámica o se define usando fórmulas de estado para expresar la ocurrencia de eventos.

En lo que sigue definiremos la sintaxis y la semántica de los sistemas lógicos BDI siguiendo un estilo similar al adoptado en la sección anterior.

4.3.1. Sintaxis de BDI_{CTL} y BDI_{CTL*}

Como los mundos posibles de estas lógicas son estructuras temporales la sintaxis de ambas es muy similar a la sintaxis de las lógicas temporales arborescentes. La diferencia con éstas es que ahora consideramos los operadores modales para las actitudes proposicionales [137].

Entonces, en primer lugar, el vocabulario está constituido por los siguientes elementos:

- Un conjunto de variables proposicionales: Var
- Operadores unarios: $\neg$, $\bigcirc$, BEL, DES, INT
- Operadores binarios: $\wedge$, U
- Cuantificadores de camino: E, A
- Signos de agrupación: $(,)$

Al igual que en las lógicas temporales arborescentes hay dos tipos de fórmulas bien formadas: las fórmulas de estado, que son evaluadas con respecto a un mundo posible en particular; y las fórmulas de camino, que son evaluadas con respecto al camino formado por una serie de transiciones entre mundos posibles. Para la lógica BDI_{CTL*} las fórmulas de estado se definen inductivamente del siguiente modo:

- Si $\phi \in Var$, ϕ es una fórmula de estado.
- Si ϕ y ψ son fórmulas de estado, $\neg\phi$ y $\phi \wedge \psi$ son fórmulas de estado.
- Si ϕ es una fórmula de estado, BEL(ϕ), DES(ϕ) e INT(ϕ) son fórmulas de estado.

- Si ϕ es una fórmula de camino, E(ϕ) es una fórmula de estado.
- Toda fórmula de estado es una fórmula de camino.
- Si ϕ y ψ son fórmulas de camino, $\neg\phi$ y $\phi \wedge \psi$ son fórmulas de camino.
- Si ϕ y ψ son fórmulas de camino, $\bigcirc\phi$ y ϕ U ψ son fórmulas de camino.

Las constantes lógicas de *disyunción* ($\vee$), *implicación* ($\Rightarrow$), *en algún momento* (*eventually*, $\Diamond$) y *siempre* (A) se definen como abreviaturas en los términos lógicos usuales.

El uso de cuantificadores de camino introduce una nueva clasificación de las fórmulas bien formadas (fbfs) en estas lógicas: una fórmula es opcional (una O-fórmula) cuando no contiene ocurrencias de A (o de la negación de $\Diamond$) fuera del alcance de los operadores BEL, DES e INT. Por otro lado, una fórmula inevitable (una I-fórmula) es una fórmula que no contiene ocurrencias de $\Diamond$ (o de la negación de A) fuera del alcance de los operadores BEL, DES e INT.[4]

4.3.2. Semántica de BDI_{CTL} y BDI_{CTL*}

En estos sistemas cada mundo posible se define como una estructura de árbol con un pasado único y un futuro arborescente. Cada estructura de árbol denota los cursos opcionales de eventos que pueden ser elegidos por el agente a partir de un mundo particular. Por tanto, los estados funcionan como índices en estas estructuras de árbol. Una relación de accesibilidad para las creencias establece qué mundos son creíbles a partir de cierto mundo y cierto estado. Las relaciones de accesibilidad para deseos e intenciones funcionan de modo similar.

Para definir la semántica de los sistemas BDI_{CTL} y BDI_{CTL*} es necesaria una estructura de Kripke K:

[4]La lógica restringida BDI_{CTL} se obtiene al prohibir el anidamiento de operadores temporales en las fórmulas de camino. Formalmente, las reglas de fórmula de camino se sustituyen por la siguiente regla:

- Si ϕ y ψ son fórmulas de estado, $\bigcirc\phi$ y ϕ U ψ son fórmulas de camino.

Definición 8 $K = \langle W, Var, \{S_w : w \in W\}, \{R_w : w \in W\}, L, \mathsf{BEL}, \mathsf{DES}, \mathsf{INT}\rangle$ *donde:*

- W *es un conjunto de mundos posibles.*
- Var *es un conjunto de variables.*
- S_w *es el conjunto de estados para cada mundo* $w \in W$*.*
- R_w *es una relación binaria serial sobre* S_w*.*
- L *es una función de asignación de valores de verdad para las proposiciones en todo* $w \in W$ *y en cada* $s \in S_w$ *t.q.* $L : W \times \cup S_w \longrightarrow Var$*.*
- BEL, DES *e* INT *son relaciones sobre los mundos y sus estados.*

Como es usual, $\phi \in Var$, la satisfacción de una fbf se denota por $\models$ y se define con respecto a una estructura de Kripke, un mundo w y un estado s. La expresión $K, w_s \models \phi$ se entiende como: la estructura K en el mundo w en el estado s satisface a ϕ. Con estas condiciones podemos mostrar la semántica de los sistemas BDI_{CTL} y BDI_{CTL*}:

- $K, w_s \models \phi \Leftrightarrow \phi \in L(w, s)$
- $K, w_s \models \neg\phi \Leftrightarrow K, w_s \not\models \phi$
- $K, w_s \models \phi \wedge \psi \Leftrightarrow K, w_s \models \phi$ y $K, w_s \models \psi$
- $K, w_s \models \mathsf{E}\phi \Leftrightarrow \exists c = (w_{s_0}, \ldots) : K, c \models \phi$
- $K, w_s \models \mathsf{A}\phi \Leftrightarrow \forall c = (w_{s_0}, \ldots) : K, c \models \phi$
- $K, w_s \models \Pi\phi \Leftrightarrow \exists v | (w, s, v) \in \Pi, K, w_s \models \phi$ donde $\Pi \in \{\mathsf{BEL}, \mathsf{DES}, \mathsf{INT}\}$
- $K, (w_{s_{0,\ldots}}) \models \phi \Leftrightarrow K, w_{s_0} \models \phi$
- $K, (w_{s_{0,\ldots}}) \models \neg\phi \Leftrightarrow K, w_{s_0} \not\models \phi$
- $K, (w_{s_{0,\ldots}}) \models \phi \wedge \psi \Leftrightarrow K, (w_{s_{0,\ldots}}) \models \phi$ y $K, (w_{s_{0,\ldots}}) \models \psi$
- $K, (w_{s_{0,\ldots}}) \models \bigcirc\phi \Leftrightarrow K, w_{s_1} \models \phi$
- $K, (w_{s_{0,\ldots}}) \models \phi$ U $\psi \Leftrightarrow$ *i)* $\exists k, k \geq 0 : K, (w_{s_k}, \ldots) \models \psi$ y $\forall j, 0 \leq j < k, K, (w_{s_j}, \ldots) \models \phi$; ó *ii)* $\forall j \geq 0, K, (w_{s_j}, \ldots) \models \phi$

La validez y la satisfacción, como veremos más adelante, se definen de manera estándar. Informalmente, una fbf es válida si y sólo si es verdadera en todo estado, en todo mundo de toda estructura.

La validez y la satisfacción con respecto a una familia de estruturas también puede definirse. Rao y Georgeff consideran dos clases de estructuras con respecto a las cuales evaluar validez y satisfacción: *i)* M^{est} que requiere que R sea total sin imponer ninguna restricción sobre los operadores intencionales; y *ii)* R^{est} que requiere que R sea total, BEL serial, transitiva y euclidiana; y DES e INT seriales. Este modelo subyace en la lógica denominada como $B^{KD45}D^{KD}I^{KD}_{CTL}$ (Cuadro 4.1).

Axioma	Relación	Sistema	Axiomatización
T: $\Box\phi \Rightarrow \phi$	Reflexiva	T	KT
D: $\Box\phi \Rightarrow \Diamond\phi$	Serial	S4	KD4
4: $\Box\phi \Rightarrow \Box\Box\phi$	Transitiva	S5 débil	KD45
5: $\Diamond\phi \Rightarrow \Box\Diamond\phi$	Euclidiana	S5	KT5

Cuadro 4.1: Caracterización de los sistemas modales normales

4.4. Axiomatización de los componentes BDI_{CTL} y BDI_{CTL*}

Dado que los componentes BDI_{CTL} y BDI_{CTL*} se comportan como sistemas modales normales [88], el axioma K se adopta para los componentes BDI:

- **Ax1.** $\mathbf{B}^K$: $\mathsf{BEL}(\phi) \wedge \mathsf{BEL}(\phi \Rightarrow \psi) \Rightarrow \mathsf{BEL}(\psi)$
- **Ax2.** $\mathbf{D}^K$: $\mathsf{DES}(\phi) \wedge \mathsf{DES}(\phi \Rightarrow \psi) \Rightarrow \mathsf{DES}(\psi)$
- **Ax3.** $\mathbf{I}^K$: $\mathsf{INT}(\phi) \wedge \mathsf{INT}(\phi \Rightarrow \psi) \Rightarrow \mathsf{INT}(\psi)$

La regla de generalización (gen) se adopta para Π y establece que toda fórmula válida es creída, deseada o intentada. A la lógica resultante se le denomina BDI^K_{CTL}:

- **Ax4.** $\mathbf{B}^{gen}$: $\vdash \phi \Rightarrow \vdash \mathsf{BEL}(\phi)$
- **Ax5.** $\mathbf{D}^{gen}$: $\vdash \phi \Rightarrow \vdash \mathsf{DES}(\phi)$

- **Ax6. I**^{gen}: $\vdash \phi \Rightarrow \vdash \mathsf{INT}(\phi)$

El sistema modal $KD45$ es adoptado para las creencias. El axioma D expresa la consistencia de las creencias, y los axiomas 4 y 5 expresan introspección positiva y negativa, respectivamente:

- **Ax7. B**^{D}: $\mathsf{BEL}(\phi) \Rightarrow \neg\mathsf{BEL}(\neg\phi)$
- **Ax8. B**^{4}: $\mathsf{BEL}(\phi) \Rightarrow \mathsf{BEL}(\mathsf{BEL}(\phi))$
- **Ax9. B**^{5}: $\neg\mathsf{BEL}\phi \Rightarrow \mathsf{BEL}(\neg\mathsf{BEL}(\phi))$

Para los deseos y las intenciones se adopta además el axioma D para expresar consistencia entre los deseos y las intenciones:

- **Ax10. D**^{D}: $\mathsf{DES}(\phi) \Rightarrow \neg\mathsf{DES}(\neg\phi)$
- **Ax11. I**^{D}: $\mathsf{INT}(\phi) \Rightarrow \neg\mathsf{INT}(\neg\phi)$

Como decíamos, la lógica resultante $B^{KD45}D^{KD}I^{KD}_{CTL}$ es consistente y completa con respecto a la familia de estructuras M^{est} [137].

4.5. Realismos

El conjunto de relaciones estructurales BDI puede combinarse para obtener una variedad de estructuras de mundos posibles diferentes. Tres de estas relaciones han sido consideradas en la literatura bajo los términos de realismo [41], realismo fuerte [134] y realismo débil [135].

4.5.1. Realismo fuerte

El realismo fuerte indica que el conjunto de mundos accesibles por las creencias es un subconjunto de los mundos accesibles por los deseos; y cada mundo accesible por creencias es un subconjunto de los mundos deseados. Como resultado, si el agente desea opcionalmente lograr una proposición, entonces cree que la proposición es una opción que, si es elegida, se logra. El realismo fuerte también puede aplicarse a los deseos e intenciones. De esta forma, si el agente intenta opcionalmente lograr una proposición, entonces también desea opcionalmente lograr esa proposición.

Los diferentes mundos accesibles por creencias, deseos e intenciones, representan diferentes posibles escenarios para el agente. Intuitivamente, el agente cree que el mundo actual es uno de sus mundos accesibles por creencias; si sucede que estuviera en el mundo accesible por creencias b_1, entonces sus deseos (con respecto a b_1) serían un mundo accesible por deseos, por ejemplo, d_1; y sus intenciones un mundo accesible por intenciones, por ejemplo, i_1. Los mundos d_1 e i_1 representan incrementalmente opciones selectivas desde b_1 acerca de los deseos y los posibles cursos de acción futuros. Si ϕ es una O-fórmula, las condiciones anteriores se expresan con los axiomas de realismo fuerte:

- **Ax12. DB-Realismo fuerte**: DES(ϕ) $\Rightarrow$ BEL(ϕ)
- **Ax13. ID-Realismo fuerte**: INT(ϕ) $\Rightarrow$ DES(ϕ)

Los axiomas anteriores expresan que si el agente tiene la intención hacia ϕ, también desea que ϕ, esto es, existe al menos un camino en el que en todos los mundos accesibles por deseo ϕ es verdadera. Esto se asegura porque las relaciones BEL y DES son seriales.

Las condiciones semánticas para el realismo fuerte se expresan como:

- **DB-Realismo fuerte**: $\forall w \forall s \forall v$ si $(w, s, v) \in$ BEL, $\exists v', (w, s, v') \in$ DES y $v \sqsubseteq v'$ (ó BEL $\subseteq$ DES)
- **ID-Realismo fuerte**: $\forall w \forall s \forall v$ si $(w, s, v) \in$ DES, $\exists v', (w, s, v') \in$ INT y $v \sqsupseteq v'$ (ó DES $\supseteq$ INT)

4.5.2. Realismo

Cohen y Levesque [41] consideran una estructura donde el conjunto de mundos accesibles por intención es un subconjunto del conjunto de mundos accesibles por creencias y las estructuras de creencia e intención son idénticas (una línea de tiempo). Esta restricción se conoce como realismo y tiene como efecto que si un agente cree una proposición también tendrá una intención con respecto a esa proposición. Los axiomas del realismo son:

- **Ax14. BD-Realismo**: BEL(ϕ) $\Rightarrow$ DES(ϕ)
- **Ax15. DI-Realismo**: DES(ϕ) $\Rightarrow$ INT(ϕ)

Las condiciones semánticas del realismo son como siguen:

- **BD-Realismo**: $\forall w \forall s \forall v$ si $(w, s, v) \in$ DES, $(w, s, v) \in$ BEL (ó DES $\subseteq$ BEL)
- **DI-Realismo**: $\forall w \forall s \forall v$ si $(w, s, v) \in$ INT, $(w, s, v) \in$ DES (ó INT $\subseteq$ DES)

4.5.3. Realismo débil

Es posible obtener un balance entre los dos enfoques anteriores si los agentes no desean aquellas proposiciones cuya negación es creída; no intentan proposiciones cuya negación es deseada; y no intentan proposiciones cuya negación es creída. A este esquema se le conoce como realismo débil y se especifica con los siguientes axiomas:

- **Ax16. DB-Realismo débil**: DES$(\phi) \Rightarrow \neg$BEL$(\neg\phi)$
- **Ax17. IB-Realismo débil**: INT$(\phi) \Rightarrow \neg$BEL$(\neg\phi)$
- **Ax18. ID-Realismo débil**: INT$(\phi) \Rightarrow \neg$DES$(\neg\phi)$

A estos axiomas les corresponde la versión multi-modal de la condición serial:

- **DB-Realismo débil**: $\forall w \forall s \exists v (w, s, v) \in$ DES si y sólo si $(w, s, v) \in$ BEL (ó BEL $\cap$ DES $\neq \varnothing$)
- **IB-Realismo débil**: $\forall w \forall s \exists v (w, s, v) \in$ INT si y sólo si $(w, s, v) \in$ BEL (ó BEL $\cap$ INT $\neq \varnothing$)
- **ID-Realismo débil**: $\forall w \forall s \exists v (w, s, v) \in$ INT si y sólo si $(w, s, v) \in$ DES (ó DES $\cap$ INT $\neq \varnothing$)

4.5.4. Otras relaciones

Es posible especificar otras relaciones. Así, por ejemplo, si un agente tiene una intención entonces cree que tiene tal intención:

- **Ax19. I-BI**: INT$(\phi) \Rightarrow$ BEL(INT(ϕ))
- **I-BI**: $\forall w \forall s \forall w' \forall w''$ si $(w, s, w') \in$ BEL y $(w, s, w'') \in$ INT entonces $(w', s, w'') \in$ BEL

Si un agente tiene un deseo, entonces cree que tiene tal deseo:

- **Ax20. D-BD**: DES(ϕ) $\Rightarrow$ BEL(DES(ϕ))
- **D-BD**: $\forall w \forall s \forall w' \forall w''$ si $(w, s, w') \in$ BEL y $(w, s, w'') \in$ DES entonces $(w', s, w'') \in$ BEL

Si un agente tiene una intención, desea tal intención:

- **Ax21. I-DI**: INT(ϕ) $\Rightarrow$ DES(INT(ϕ))
- **I-DI**: $\forall w \forall s \forall w' \forall w''$ si $(w, s, w') \in$ DES y $(w, s, w'') \in$ INT entonces $(w', s, w'') \in$ DES

Si la relación de equivalencia ($\equiv$) es usada en este axioma en lugar de la implicación ($\Rightarrow$), las modalidades anidadas se colapsan.

Finalmente, si un agente forma una intención, entonces en algún momento futuro la abandona. Esto se conoce como no-retención infinita (*no-infinite deferral*) y, siendo parte fundamental de un modelo bratmaniano, jugará un papel importante en próximos capítulos:

- **Ax22**: INT(ϕ) $\Rightarrow$ A$\Diamond$($\neg$INT(ϕ))

4.5.5. Eventos

Además, para poder describir la conducta de un agente es necesario describir la ocurrencia de acciones, aquí llamadas eventos. Las extensiones necesarias en este sentido incluyen permitir que las fórmulas bien formadas expresen el éxito y fracaso de los eventos. Si e es un evento primitivo, $succeeded(e)$ denota la ocurrencia exitosa de e en el pasado inmediato; $failed(e)$ denota el fracaso de e en el pasado inmediato; $done(e)$ denota la ocurrencia de e en el pasado inmediato (con éxito o fracaso). De manera similar, $succeeds(e)$, $fails(e)$, y $does(e)$ se usan para denotar ocurrencias futuras de e.

Sintaxis de los eventos

Primero necesitamos un conjunto de símbolos para identificar a los eventos primitivos, por ejemplo, E. Ahora, la definición de fórmula de estado debe extenderse con la inclusión de la siguiente fórmula:

- Si $e \in E$, entonces $succeeds(e)$, $fails(e)$, $does(e)$, $succeeded(e)$, $failed(e)$, y $done(e)$ son fórmulas de estado.

Semántica de los eventos

E es un conjunto de tipos de evento primitivos; $SE_w : S_w \times S_w \longrightarrow E$ y $FE_w : S_w \times S_w \longrightarrow E$ que representan ocurrencias con éxito y fracaso de los eventos. Nótese que SE_w y FE_w son disjuntos. La semántica de los eventos, en un modelo M, se define como sigue:

- $M, w_{s_1} \models succeeded(e)$ si y sólo si $SE_w(s_0, s_1) = e$
- $M, w_{s_1} \models failed(e)$ si y sólo si $FE_w(s_0, s_1) = e$

El resto de los eventos se define así: $done(e)$ se define como $succeeded(e) \vee failed(e)$, esto es, la ocurrencia del evento e independientemente de si se realizó con éxito o fracaso; $succeeds(e)$ se define como A $\bigcirc$ $(succeeded(e))$; $fails(e)$ se define como A $\bigcirc$ $(failed(e))$, esto es, el evento e se realiza con éxito o fracaso en todas las ramas a partir del estado actual; y $does(e)$ se define como A $\bigcirc$ $(done(e))$.

De este modo podemos axiomatizar los eventos que capturan el carácter volitivo del compromiso subyacente en las intenciones. Este axioma debe expresar que un agente actuará si tiene una intención dirigida hacia un tipo de evento primitivo:

- **Ax23**: INT$(does(e)) \Rightarrow does(e)$

Un agente debe ser consciente (debe creer) de todos los tipos de eventos primitivos que ocurren en su medio ambiente:

- **Ax24**: $done(e) \Rightarrow$ BEL$(done(e))$

Al elegir uno de los realismos y adoptar el resto de los axiomas descritos, configuramos lo que Rao y Georgeff [135] llaman *Basic I system*, lo que permite una formalización de un sistema intencional.

4.6. Compromiso como axioma de cambio

Con este formalismo estamos en condiciones de especificar el funcionamiento de las intenciones futuras como intenciones en proceso de mantenimiento y revisión mediante una estrategia de compromiso.

Tres estrategias son bien conocidas en la literatura multi-agentes [134]: la estrategia de compromiso ciego (*blind*), la determinada (*single-minded*) y la abierta (*open-minded*). Eligiendo una de estas tres estrategias y asumiendo el *Basic I System*, se configuran tres diferentes agentes BDI básicos:

- Un agente *blind* es un agente que se compromete ciegamente y mantiene sus intenciones hasta que cree que las ha logrado satisfacer:

$$\mathsf{INT}(\mathsf{A}\Diamond\phi) \implies \mathsf{A}(\mathsf{INT}(\mathsf{A}\Diamond\phi)\ \mathsf{U}\ \mathsf{BEL}(\phi))$$

Es importante notar que este axioma se define para I-fórmulas. Nada se dice sobre la intención de un agente por lograr opcionalmente algún medio o fin particular. Esta estrategia es demasiado fuerte, pues para un agente comprometido ciegamente resulta inevitable creer que ha logrado sus fines. Esto se debe a que este tipo de compromiso sólo permite caminos futuros en los cuales o bien el objeto de la intención es creído o bien la intención se mantiene para siempre. Sin embargo, debido a la propiedad de no-retención infinita (*no-infinite deferral*), tenemos que $(\mathsf{INT}(\phi) \Rightarrow \mathsf{A}\Diamond(\neg\mathsf{INT}(\phi)))$, por lo que tal clase de caminos no está permitida, y en consecuencia obtenemos agentes que creen que en algún momento (*eventually*) han logrado sus intenciones:

$$\mathsf{INT}(\mathsf{A}\Diamond\phi) \Rightarrow \mathsf{A}\Diamond\mathsf{BEL}(\phi)$$

- Un agente *single-minded* ocurre al relajar la estrategia anterior de modo que el agente mantenga sus intenciones en tanto considere que siguen siendo una opción viable. Formalmente:

$$\mathsf{INT}(\mathsf{A}\Diamond\phi) \implies \mathsf{A}(\mathsf{INT}(\mathsf{A}\Diamond\phi)\ \mathsf{U}\ (\mathsf{BEL}(\phi) \vee \neg\mathsf{BEL}(\mathsf{E}\Diamond\phi)))$$

En tanto el agente crea que sus intenciones se pueden lograr, un agente de este tipo no abandonará sus intenciones y seguirá comprometido. Así, un agente determinado de este tipo, de manera inevitable en algún momento creerá que ha logrado satisfacer sus fines sólo si continua creyendo que el objeto de sus intenciones sigue siendo una opción.

- Un agente *open-minded* es un agente que mantiene sus intenciones mientras éstas sigan siendo deseadas. Formalmente:

$$\mathsf{INT}(\mathsf{A}\Diamond\phi) \implies \mathsf{A}(\mathsf{INT}(\mathsf{A}\Diamond\phi)\ \mathsf{U}\ (\mathsf{BEL}(\phi) \vee \neg\mathsf{DES}(\mathsf{A}\Diamond\phi)))$$

Como podrá apreciarse ahora, en estos sistemas lógicos BDI el comportamiento de un agente BDI es especificado en términos de los cambios temporales en las actitudes intencionales del agente (creencias, deseos e intenciones), por ejemplo, mediante la definición de los realismos. Estos sistemas, en nuestra opinión, capturan de manera más clara las propiedades de racionalidad de los agentes a través de los axiomas. Entre tales propiedades encontramos las estrategias de compromiso que dictan bajo qué condiciones un agente debería abandonar o mantener una intención. Además este es un formalismo rico expresivamente y está más cercano a un modelo bratmaniano; sin embargo, también tenemos un par de objeciones: una de carácter filosófico y otra de tipo técnico-ingenieril.

El primer problema—que, por cierto, comparte con $C\&L$—es que aunque parece reconocer la importancia del tiempo en el razonamiento intencional, no toma en cuenta la naturaleza no-monotónica del mismo. Además, dadas las propiedades estructurales, está el problema de la representación de las intenciones, pues este modelo las representa con un operador único y no como reglas o planes.

El problema técnico de estos sistemas es que, si bien son usados con éxito para razonar acerca de los estados de los agentes, no son usados para programarlos (para pasar de la comprensión a la reproducción). Y ya que nuestro interés no es únicamente entender un modelo de razonamiento intencional, sino también hacerlo con miras a su reproducción, necesitamos establecer un enlace entre la especificación lógica y la implementación.

Por tanto, podemos apreciar que los modelos lógicos de agencia BDI basados en un modelo bratmaniano comprenden a las intenciones de dos maneras: o bien como una combinación de creencias y deseos (el modelo BD formalizado por $C\&L$) o bien como un elemento independiente pero de forma unitaria (el modelo BDI formalizado por Rao y Georgeff). Además estos formalismos no consideran el aspecto no-monotónico del razonamiento intencional, lo cual podemos ver al enfocarnos en el fragmento I formalizado bajo las normas del axioma K y, por ello, bajo la estructura de un sistema modal normal que induce monotonicidad.

Para resolver este último inconveniente tenemos lógicas no-monotónicas para formalizar el fenómeno del razonamiento por *default*; sin embargo, existen diferentes sistemas no-monotónicos. La mayoría de ellos, no obstante, utilizan como unidad no-estándar la noción de *default* que puede ser descrita mediante expresiones como *típicamente* o *a menos que se muestre de otro modo*.

Esta noción de *default*, común a muchos sistemas, no garantiza que los consecuentes se den siempre que se den los antecedentes y, por tanto, las lógicas no-monotónicas nos permiten derivar consecuencias de manera derrotable. Con todo, estos útiles sistemas se dedican a estudiar y formalizar el fragmento B en

vez de otros componentes del espectro cognitivo y el estudio con estructuras temporales es reciente [13, 14, 74]: tenemos sistemas no-monotónicos pero no intencional-temporales.

4.7. Resumen

Revisamos el trabajo fundacional de Cohen y Levesque, los sistemas computacionales arborescentes CTL y $CTL*$ y, finalmente, el sistema lógico BDI de Rao y Georgeff.[5] Con esta revisión hicimos una valoración, y a partir de la definición de un modelo bratmaniano, notamos dos problemas: uno técnico y otro más profundo. Ambos necesitan ser resueltos.

En resumen, desde un punto de vista cognitivo un agente BDI es un sistema agente que se explica en términos del modelo BDI. Un modelo BDI bratmaniano de razonamiento intencional debe considerar los atributos lógicos de revisión de intenciones y la no-monotonicidad del razonamiento intencional. El problema que queda abierto es el de cómo traducir estos atributos a un modelo formal. Los modelos lógicos tradicionales y más usados para llevar a cabo este proceso, a pesar de sus ventajas, tienen problemas técnicos y filosóficos que es necesario resolver.

[5] Están, además, el sistema temporal de planificación de Dean y McDermott [49] (que no revisamos aquí por estar más del lado técnico que del filosófico), y los formalismos de Seel [145] y Singh [150] (que tampoco revisamos porque, respectivamente, uno conlleva el problema del efecto colateral y el otro termina siendo equivalente a $C\&L$).

Capítulo 5

AgentSpeak(L)-Jason

5.1. Introducción

En este punto es claro que los métodos formales nos permiten especificar, diseñar y modelar agentes BDI. No obstante, las implementaciones de estos modelos, en su mayoría, asumen una simplificación con consecuencias teóricamente graves: si bien modelan las actitudes intencionales como estructuras de datos y mejoran su desempeño efectivo, carecen de los fundamentos teóricos suficientes para comprender a los programas mismos [136] y para formar parte de un modelo cognitivo.

Esto ha sido así porque los sistemas lógicos multimodales BDI empleados en la especificación formal de estos sistemas poco ofrecen en relación con los problemas prácticos de su programación o implementación.

Los primeros intentos para resolver este problema de traducción entre especificación e implementación se concentraron en proponer una arquitectura abstracta BDI como una idealización de la implementación de estos sistemas y como un vehículo para continuar investigando las propiedades teóricas de la agencia intencional. Ahora sabemos que estos intentos culminaron en el exitoso dMARS.[1] Sin embargo, debido a su nivel de abstracción, no fue posible establecer una correspondencia uno-a-uno entre la propuesta filosófica, la teoría de prueba formal y el intérprete abstracto implementado.

Con $AgentSpeak(L)$, una versión simplificada de dMARS, Rao propuso una posible solución [134, 136]. $AgentSpeak(L)$ es un lenguaje de programación

[1] La plataforma dMARS (por *distributed multi-agent reasoning system*) ha sido desarrollada para diseñar agentes de *software* con base en el modelo BDI [92].

basado en una lógica restringida de primer orden con eventos y acciones, si bien las actitudes intencionales no están representadas como expresiones modales.

En $AgentSpeak(L)$ el estado actual de un agente, que es modelo de él mismo, su ambiente y otros agentes, puede considerarse como el conjunto de creencias presentes del agente. Los estados que el agente quiere lograr con base en sus estímulos internos y externos, constituyen sus deseos. Y la adopción de planes para satisfacer estos deseos constituyen las intenciones del agente. Esta apuesta por adscribir intencionalidad a un modelo ejecutable del agente constituyó un cambio de paradigma que acercaba la teoría a la praxis de la agencia BDI.

En lo que sigue presentaremos la sintaxis y la semántica de $AgentSpeak(L)$ (§ 5.2 y § 5.3) así como su teoría de prueba (§ 5.4). Posteriormente revisaremos la sintaxis y la semántica de *Jason* (§ 5.5), el intérprete de $AgentSpeak(L)$. La idea de este capítulo es la siguiente: como para resolver los problemas técnicos necesitaremos un lenguaje, es preciso mostrar ciertos detalles del mismo de manera pormenorizada.

5.2. Sintaxis de $AgentSpeak(L)$

En esta sección abordaremos el lenguaje para escribir programas de agente en $AgentSpeak(L)$. El alfabeto de este lenguaje formal consiste en variables, constantes, símbolos funcionales, símbolos de predicado, símbolos de acciones, conectivas, cuantificadores y signos de puntuación. Además de las conectivas de primer orden se usan los operadores unarios ! y ? para identificar ciertas metas, ; para secuencias y $\leftarrow$ para la implicación. Las definiciones estándar para término, fórmula bien formada, fórmula cerrada, ocurrencia libre y acotada de variables son adoptadas de Bordini *et al* [19].

Definición 9 *(Creencias) Sea b un símbolo de predicado y $t_1, \ldots, t_n$ una secuencia de términos, entonces $b(t_1, \ldots, t_n)$ ó $b(t)$ es una creencia atómica. Si $b(t_1)$ y $b(t_2)$ son creencias atómicas, entonces $b(t_1) \wedge b(t_2)$ y $\neg b(t_1)$ son creencias. Una creencia atómica o su negación será identificada como literal de creencia. Una creencia atómica sin variables libres será llamada creencia básica.*

Tómese como ejemplo una simulación de tráfico en una autopista de cuatro carriles. Los autos pueden aparecer en cualquier carril moviéndose de norte a sur. Es posible que aparezca basura en los carriles y nuestro contratista ha comprado un robot que recoge la basura y la pone en un depósito. Necesitamos programar

al robot de tal manera que haga su trabajo sin ponerse en peligro (es decir, sin quedarse en un carril donde ha aparecido un auto).

Ejemplo 3 *A lo largo de este capítulo construiremos el programa de agente para este robot a manera de ejemplo. Las creencias de este agente representan la configuración de la autopista, agentes y objetos en ella incluidos. Las creencias lucen como* $adyacente(X, Y)$, $pos(Robot, X)$, *etc. Las creencias básicas del agente son casos sin variables libres de estas creencias atómicas, por ejemplo,* $adyacente(c1, c2)$, $pos(r1, c2)$, *etc.*

Definición 10 *(Metas) Si* g *es un símbolo de predicado y* $t_1, \ldots, t_n$ *es una secuencia de términos, entonces* $!g(t_1, \ldots, t_n)$, $!g(t)$, $?g(t_1, \ldots, t_n)$ *ó* $?g(t)$ *son metas.*

Una meta es un estado del sistema que el agente *desearía* ver logrado. Los agentes en $AgentSpeak(L)$ consideran dos tipos de metas: las metas que propiamente el agente quiere lograr (*achievement goals*, $!g(t)$) y las metas que el agente quiere verificar (*test goals*, $?g(t)$). En el ejemplo que estamos siguiendo, limpiar el carril 2 es una meta a lograr, $!limpiar(c2)$; y preguntarse si hay un auto en el carril 1 es una meta a verificar, $?pos(auto, c1)$.

Definición 11 *(Eventos disparadores) Si* $b(t)$ *es una creencia atómica y* $!g(t)$ *y* $?g(t)$ *son metas, entonces* $+b(t)$, $-b(t)$, $+!g(t)$, $+?g(t)$, $-!g(t)$ *y* $-?g(t)$ *son eventos disparadores.*

Los cambios en el ambiente del agente y en su estado interno generan eventos disparadores (*te* por *trigger events*). Estos eventos incluyen agregar $(+)$ y borrar $(-)$ metas o creencias al estado del agente. Por ejemplo, detectar basura en un carril cualquiera toma la forma del evento disparador $+pos(basura, X)$ y adquirir la meta de limpiar un carril $+limpiar(X)$.

Definición 12 *(Acciones) Si* a *es un símbolo de acción y* $t_1, \ldots, t_n$ *es una secuencia de términos de primer orden, entonces* $a(t_1, \ldots, t_n)$ *ó* $a(t)$ *es una acción.*

El agente debe ejecutar acciones para lograr el cumplimiento de sus metas. Las acciones pueden verse como procedimientos a ejecutar. Normalmente estas acciones son implementadas en el mismo lenguaje de programación en el que $AgentSpeak(L)$ ha sido implementado, pero esto no es necesariamente el caso, pues puede hacerse uso de interfaces a lenguajes foráneos en el caso de $Java$ y $Lisp$. En todo caso, una acción como $ir(X, Y)$ debería tener como resultado que el agente se halle en el carril Y y no en X.

Definición 13 *(Planes) Si te es un evento disparador, $b_1, \ldots, b_n$ es una secuencia de literales de creencia y $g_1, \ldots, g_n$ es una secuencia de metas o acciones, entonces*

$$te : b_1 \wedge \ldots \wedge b_n \leftarrow g_1, \ldots, g_n$$

es un plan.

La expresión a la izquierda de la flecha se conoce como la cabeza del plan. La expresión a la derecha de la flecha se conoce como cuerpo del plan. La expresión a la derecha de los dos puntos (:) en la cabeza del plan se conoce como contexto. Por conveniencia un cuerpo vacío se denota como *true* o $\top$. Como el resto de los agentes BDI, los agentes $AgentSpeak(L)$ poseen una librería de planes.

Ejemplo 4 *A continuación tenemos un plan para responder a la aparición de basura en algún carril. Si el agente está en el mismo carril que la basura, ejecutará la acción de levantarla, y se planteará la meta de ir al depósito para finalmente ejecutar la acción de tirar ahí la basura. El comportamiento del agente se completa con los planes para ir a algún sitio:*

```
@p0
+pos(basura,X)
   : (pos(r1,X) & pos(deposito,Y))
   <- levantar(basura);
   !pos(r1,Y);
   tirar(basura).

@p1
+!pos(r1,X)
   : pos(r1,X)
   <- true.

@p2
+!pos(r1,X)
   : (pos(r1,Y) & ((not(X=Y)) & adyacente(Y,Z)))
<- ir(Y,Z);
!pos(r1,X).
```

Esta forma de especificar un agente es similar a lo que hacemos en programación lógica al especificar hechos y reglas. Sin embargo, existen diferencias importantes entre un programa lógico y un programa de agente:

- En un programa lógico puro no hay diferencias entre una meta en el cuerpo de una regla y en su cabeza. En un programa de agente, la meta en la cabeza es un evento disparador, no una meta en sí. Esto permite más expresividad para invocar planes posibilitando procesos dirigidos por datos (al agregar y eliminar creencias) y dirigidos por metas (al agregar y eliminar metas).

- Las reglas en la programación lógica pura no son sensibles al contexto, como sí lo son los planes.[2]

- La ejecución exitosa de una regla en la programación lógica regresa una substitución; la ejecución de un plan genera secuencias de acciones que modifican el medio ambiente donde está situado el agente.

- Cuando computamos una meta en programación lógica el proceso (*querying*) no puede ser detenido; los planes de un agente, al contrario, pueden ser interrumpidos.

5.3. Semántica de $AgentSpeak(L)$

Primero definimos los conceptos de agente, intención y evento:

Definición 14 *Un agente es una 8-tupla* $\langle E, B, P, I, A, \mathcal{S}_{\mathcal{E}}, \mathcal{S}_{\mathcal{O}}, \mathcal{S}_{\mathcal{I}} \rangle$ *donde*

- E *es un conjunto de eventos.*
- B *es un conjunto de creencias.*
- P *es un conjunto de planes.*
- A *es un conjunto de acciones.*
- $\mathcal{S}_{\mathcal{E}}$ *es una función de selección eventos.*
- $\mathcal{S}_{\mathcal{O}}$ *es una función de selección de planes aplicables.*
- $\mathcal{S}_{\mathcal{I}}$ *es una función de selección de intenciones.*

[2]La cuestión del contexto es apremiante para la lógica, y por tanto, para la filosofía y la IA [108], pues formalizar los contextos permite desarrollar axiomas y reglas para comprender y programar ciertas capacidades de representación que un agente BDI sofisticado debe tener.

Definición 15 *(Intenciones) El conjunto I se compone de las intenciones del agente. Una intención es una pila de planes cerrados parcialmente (planes que pueden incluir algunas variables libres y otras con valores asignados). Una intención se denota por $[p_1 \ddagger \ldots \ddagger p_n]$, donde p_1 representa el fondo de la pila y p_n el tope de la misma. Por conveniencia, la intención $[+!true : true \leftarrow true]$ será denotada por $T = true$.*

Definición 16 *(Eventos) El conjunto E se compone de eventos. Cada evento es un par $\langle e, i \rangle$ donde e es un evento disparador e i es una intención. Si la intención $i = true$, al evento se le identifica como un evento externo; en cualquier otro caso es un evento interno.*

Ahora podemos definir formalmente los conceptos de plan relevante y aplicable. Para ello necesitamos recordar el concepto de substitución y el de unificador más general (MGU):

Definición 17 *(Substitución) Una substitución es un conjunto finito de pares de términos $\{X_1/t_1, \ldots, X_n/t_n\}$ donde cada t_i es un término y cada X_i es una variable t.q. $X_i \neq t_i$ y $X_i \neq X_j$.*

Definición 18 *(Unificador) Sean α y β términos. Una substitución θ tal que α y β sean idénticos ($\alpha\theta = \beta\theta$) es llamada unificador de α y β.*

Definición 19 *(Generalidad entre substituciones) Una substitución θ se dice más general que una substitución σ, si y sólo si existe una substitución γ tal que $\sigma = \theta\gamma$.*

Definición 20 *(MGU) Un unificador θ se dice el unificador más general (MGU) de dos términos, si y sólo si θ es más general que cualquier otro unificador entre esos términos.*

Definición 21 *(Plan relevante) Sea $\mathcal{S}_{\mathcal{E}}(E) = \epsilon = \langle d, i \rangle$ y sea el plan $p = e : b_1 \wedge \cdots \wedge b_n \leftarrow h_1; \ldots; h_m$. El plan p es relevante con respeto al evento ϵ si y sólo si existe un unificador más general (MGU) σ tal que $d\sigma = e\sigma$. A σ se le llama el unificador relevante para ϵ.*

Definición 22 *(Plan aplicable) Un plan p denotado por $e : b_1 \wedge \cdots \wedge b_n \leftarrow h_1; \ldots; h_m$ es un plan aplicable con respecto a un evento ϵ si y sólo si existe un unificador relevante σ para ϵ y existe una substitución θ tal que $\forall(b_1 \wedge \cdots \wedge b_n)\sigma\theta$ es una consecuencia lógica de B (creencias del agente). La composición $\sigma\theta$ se conoce como el unificador aplicable para ϵ; y θ se conoce como la substitución de respuesta correcta.*

Ejemplo 5 *Siguiendo con el ejemplo, si asumimos que las creencias básicas del agente son las siguientes:*

```
adyacente(c1,c2).
adyacente(c2,c3).
adyacente(c3,c4).
pos(r1,c1).
pos(basura,c2).
pos(deposito,c4).
```

Tenemos que el unificador aplicable es $\{X/c2, Y/c1, Z/c2\}$ *y por lo tanto el único plan aplicable es* $p2$*. Dependiendo del tipo de evento (interno o externo) la intención será diferente. En el caso de los eventos externos, los medios se obtienen seleccionando un plan aplicable para el evento y entonces se aplica el unificador aplicable al cuerpo del plan. Este medio es utilizado para crear una nueva intención que se agrega a* I*. En el ejemplo, el plan* $p2$ *formará una nueva intención de la forma:*

```
+!pos(r1,c2)
   : (pos(r1,c1) & ((not(c2=c1)) & adyacente(c1,c2)))
   <- ir(c1,c2);
   !pos(r1,c2).
```

Definición 23 *(Intención evento externo) Sea* $\mathcal{S}_{\mathcal{O}}(O_\epsilon) = p = e : b_1 \wedge \cdots \wedge b_n \leftarrow h_1; \ldots; h_m$ *donde* O_ϵ *es el conjunto de todos los planes aplicables u opciones para el evento* $\langle d, i\rangle$*. El plan* p *es intentado con respecto al evento* ϵ*, donde la intención* $i = T$*, si y sólo si existe un unificador aplicable* σ *tal que* $[+!true : true \leftarrow true \ddagger (e : b_1 \wedge \cdots \wedge b_n \leftarrow h_1; \ldots; h_m)\sigma] \in I$*.*

Definición 24 *(Intención evento interno) Sea* $\mathcal{S}_{\mathcal{O}}(O_\epsilon) = p = +!g(s) : b_1 \wedge \cdots \wedge b_j \leftarrow h_1; \ldots; h_k$ *donde* O_ϵ *es el conjunto de todos los planes aplicables u opciones para el evento* $\epsilon = \langle d, [p_1 \ddagger \ldots \ddagger p_n : c_1 \wedge \cdots \wedge c_m \leftarrow !g(t); k_2; \ldots; k_n]\rangle$*. El plan* p *es intentado con respecto al evento* ϵ*, si y sólo si existe un unificador aplicable* σ *tal que* $[p_1 \ddagger \ldots \ddagger p_n : c_1 \wedge \cdots \wedge c_m \leftarrow !g(t); k_2; \ldots; k_n \ddagger (+!g(s) : b_1 \wedge \cdots \wedge b_j)\sigma \leftarrow (h_1; \ldots; h_k)\sigma; (k_2; \ldots k_n)\sigma] \in I$*.*

Definición 25 *(Ejecución achieve) Sea* $\mathcal{S}_{\mathcal{I}}(I) = i = [p_1 \ddagger \ldots \ddagger p_n : c_1 \wedge \cdots \wedge c_m \leftarrow !g(t); h_2; \ldots; h_n]$*. Se dice que la intención* i *ha sido ejecutada, si y sólo si* $\langle +!g(t), i\rangle \in E$*.*

Definición 26 *(Ejecución test) Sea* $\mathcal{S}_{\mathcal{I}}(I) = i = [p_1 \ddagger \ldots \ddagger p_n : c_1 \wedge \cdots \wedge c_m \leftarrow ?g(t); h_2; \ldots; h_n]$. *Se dice que la intención* i *ha sido ejecutada, si y sólo si existe una substitución* θ *tal que* $\forall g(t)\theta$ *es una consecuencia lógica de* B *e* i *es remplazada por* $[p_1 \ddagger \ldots \ddagger (p_n : c_1; \ldots; c_m)\sigma \leftarrow (h_2; \ldots h_n)\sigma]$.

Definición 27 *(Ejecución acción) Sea* $\mathcal{S}_{\mathcal{I}}(I) = i = [p_1 \ddagger \ldots \ddagger p_n : c_1 \wedge \cdots \wedge c_m \leftarrow a(t); h_2; \ldots; h_n]$. *Se dice que la intención* i *ha sido ejecutada, si y sólo si* $a(t) \in A$ *e* i *es remplazada por* $[p_1 \ddagger \ldots \ddagger p_n : c_1 \wedge \cdots \wedge c_m \leftarrow h_2; \ldots; h_n]$

Definición 28 *(Ejecución submeta) Sea* $\mathcal{S}_{\mathcal{I}}(I) = i = [p_1 \ddagger \ldots \ddagger p_{n-1} \ddagger !g(t) : c_1 \wedge \cdots \wedge c_m \leftarrow true]$, *donde* $p_{n-1} = e : b_1 \wedge \cdots \wedge b_x \leftarrow !g(s); h_2; \ldots; h_y$. *Se dice que la intención* i *ha sido ejecutada, si y sólo si existe una substitución* θ *tal que* $g(t)\theta = g(s)\theta$ *e* i *es remplazada por* $[p_1 \ddagger \ldots \ddagger p_{n-1} \ddagger (e : b_1 \wedge \cdots \wedge b_x)\theta \leftarrow h_2; \ldots; h_y)\theta]$.

Con estos conceptos es posible definir un intérprete para $AgentSpeak(L)$ como se muestra en el Algoritmo 5. Las funciones top, $push$, pop, $head$, $first$ y $rest$ tienen una semántica de pila.

5.4. Teoría de prueba de $AgentSpeak(L)$

Para formular la teoría de prueba de $AgentSpeak(L)$ recurrimos a un sistema de transición al estilo Plotkin [127].

Definición 29 *(Sistema de transición BDI) Un sistema de transición BDI es un par* $\langle \Gamma, \vdash \rangle$ *que consiste en:*

- *Un conjunto* Γ *de configuraciones.*
- *Una relación binaria de transición* $\vdash \subseteq \Gamma \times \Gamma$.

Definición 30 *(Configuración BDI) Una quíntupla* $\langle E_i, B_i, I_i, A_i, i \rangle$, *donde* $E_i \subseteq E$, $B_i \subseteq B$, $I_i \subseteq I$, $A_i \subseteq A$, *e* i *es la etiqueta de la transición, es una configuración BDI.*

El conjunto de planes no forma parte de las configuraciones, pues se asume que permanece constante (aunque, como veremos, este no es el caso si el agente puede modificar sus planes originales, por ejemplo, mediante aprendizaje). Tampoco se lleva un registro explícito de las metas, pues se asume que estas aparecen como intenciones cuando son adoptadas por los agentes.

Algoritmo 5: El algoritmo del intérprete $AgentSpeak(L)$

mientras $E \neq \emptyset$ **hacer**
- $\epsilon \leftarrow \langle d, i\rangle \leftarrow \mathcal{S}_{\mathcal{E}}(E)$
- $E \leftarrow E \backslash \epsilon$
- $O_\epsilon \leftarrow \{p\theta | \theta$ es un unificador aplicable para ϵ y $p\}$
- **si** $externo(\epsilon)$ **entonces**
  - $I \leftarrow I \cup [\mathcal{S}_{\mathcal{O}}(O_\epsilon)]$
  - **sino**
    - $push(\mathcal{S}_{\mathcal{O}}(O_\epsilon)\sigma, i)$ donde σ es un unificador aplicable para ϵ
  - **fin**
- **fin**
- **si** $first(body(top(\mathcal{S}_{\mathcal{I}}(I)))) = true$ **entonces**
  - $x \leftarrow pop(\mathcal{S}_{\mathcal{I}}(I)$
  - $push(head(top(\mathcal{S}_{\mathcal{I}}(I)))\theta \leftarrow rest(body(top(\mathcal{S}_{\mathcal{I}}(I))))\theta, \mathcal{S}_{\mathcal{I}}(I))$ donde θ es un MGU t.q. $x\theta = head(top(\mathcal{S}_{\mathcal{I}}(I)))\theta$
  - **sino**
    - **si** $first(body(top(\mathcal{S}_{\mathcal{I}}(I)))) = !g(t)$ **entonces**
      - $E = E \cup \langle +!g(t), \mathcal{S}_{\mathcal{I}}(I)\rangle$
    - **fin**
  - **fin**
  - **sino**
    - **si** $first(body(top(\mathcal{S}_{\mathcal{I}}(I)))) = ?g(t)$ **entonces**
      - $pop(\mathcal{S}_{\mathcal{I}}(I))$
      - $push(head(top(\mathcal{S}_{\mathcal{I}}(I))))\theta \leftarrow rest(body(top(\mathcal{S}_{\mathcal{I}}(I))))\theta, \mathcal{S}_{\mathcal{I}}(I))$ donde θ es la substitución de respuesta correcta.
    - **fin**
  - **fin**
  - **sino**
    - **si** $first(body(top(\mathcal{S}_{\mathcal{I}}(I)))) = a(t)$ **entonces**
      - $pop(\mathcal{S}_{\mathcal{I}}(I))$
      - $push(head(top(\mathcal{S}_{\mathcal{I}}(I)))) \leftarrow rest(body(top(\mathcal{S}_{\mathcal{I}}(I)))), \mathcal{S}_{\mathcal{I}}(I))$; $A = A \cup \{a(t)\}$
    - **fin**
  - **fin**
- **fin**

fin

Ahora es posible escribir reglas de transición que lleven al agente de una configuración BDI a otra. La primera regla define la transición al intentar un plan al nivel más alto (un fin, en términos de razonamiento medios-fines). La regla especifica cómo el agente modifica sus intenciones en respuesta a un evento externo:

$$\textbf{(IntendEnd)}\frac{\langle\{\ldots,\langle+!g(t),T\rangle,\ldots\},B_iI_i,A_i,i\rangle}{\langle\{\ldots\},B_i,I_i\cup\{[p\sigma\theta]\},A_i,i+1\rangle}$$

donde: $p=+!g(s):b_1\wedge\cdots\wedge b_m\leftarrow h_1;\ldots;h_n\in P$, $\mathcal{S}_{\mathcal{E}}(E)=\langle+!g(t),T\rangle$, $g(t)\sigma=g(s)\sigma$ y $\forall(b_1\wedge\cdots\wedge b_m)\theta$ es consecuencia lógica de B_i.

La regla para intentar un medio es similar a la regla para intentar un fin, sólo que el plan aplicable es colocado sobre la pila cuyo tope es la intención dada como segundo argumento del evento elegido:

$$\textbf{(IntendMeans)}\frac{\langle\{\ldots,\langle+!g(t),j\rangle,\ldots\},B_i\{\ldots,[p_1\ddagger\ \ldots\ \ddagger p_n],\ldots\},A_i,i\rangle}{\langle\{\ldots\},B_i,\{\ldots,[p_1\ddagger\ \ldots\ \ddagger p_n\ddagger p\sigma\theta],\ldots\},A_i,i+1\rangle}$$

donde $p_z=f:c_1\wedge\cdots\wedge c_y\leftarrow!g(t);h_2;\ldots;h_n$, $p=+!g(s):b_1\wedge\cdots\wedge b_m\leftarrow k_1;\ldots;k:x$, $\mathcal{S}_{\mathcal{E}}(E)=\langle+!g(t),j\rangle$ es $[p_1\ddagger\ldots\ddagger p_n]$, $g(t)\sigma=g(s)\sigma$ y $\forall(c_1\wedge\cdots\wedge c_y)\theta$ es una consecuencia lógica de B_i.

Rao define una regla más para la adopción de metas y propone que el lector puede elaborar reglas parecidas para el resto de las transiciones en el sistema. De esta forma, es posible definir derivaciones y refutaciones, usando las reglas de prueba.

Definición 31 *(Derivación BDI) Una derivación BDI es una secuencia finita o infinita de configuraciones* $\gamma_0,\ldots,\gamma_i,\ldots$.

La noción de refutación en $AgentSpeak(L)$ se da con respecto a una intención particular. La refutación de una intención inicia cuando ésta es adoptada y termina cuando su pila queda vacía. Por lo tanto, usando las reglas anteriores es posible verificar seguridad y viabilidad del sistema. Además hay una correspondencia uno-a-uno entre las reglas de prueba y la semántica operacional del sistema. Dentro de las extensiones posibles se encuentran operadores más interesantes para el cuerpo de los planes (aquellos de la lógica dinámica) y post-condiciones diferenciadas para los casos de éxito y fracaso, como se especifica en dMARS [92].

5.5. $Jason$

$Jason$ [117] es un intérprete que implementa una semántica operacional extendida de $AgentSpeak(L)$. Fue desarrollado en el lenguaje de propósito general $Java$ y su IDE soporta el desarrollo y la ejecución de sistemas multi-agentes distribuidos.

Entre sus características encontramos que:

- Permite la comunicación entre agentes basada en actos de habla, incluyendo anotaciones en las creencias con información de las fuentes [118].
- Provee anotaciones sobre las etiquetas de los planes, las cuales pueden ser empleadas para diseñar funciones de selección basadas en teoría de decisión [17].
- Ofrece la posibilidad de correr un sistema multi-agente distibuido sobre una red utilizando SACI o algún *middleware* [87].
- Facilita la creación de funciones de selección totalmente configurables mediante $Java$.
- Posibilita la extensión del repertorio de acciones internas directamente en código $Java$.
- Permite una clara noción de ambiente que permite simular la situacionalidad de los agentes en cualquier ambiente implementado en $Java$.
- Incorpora un editor gráfico, $jEdit$, que facilita el desarrollo de sistemas en $Jason$.

Existen diversos sistemas que implementan agentes basados en el modelo BDI; sin embargo, uno de los principales puntos a favor de $Jason$ es el fundamento teórico que emplea: un modelo BDI bratmaniano. Con la implementación de $Jason$ se busca probar ciertas características de los agentes BDI que sirvan para trabajar con la verificación formal de los agentes programados utilizando $AgentSpeak(L)$. En la práctica, otra característica que tiene el intérprete $Jason$ a su favor, y que lo hace muy versátil, es el hecho de que está implementado en el lenguaje de propósito general $Java$, lo que le atribuye propiedades tales como ser ejecutado en diferentes plataformas y una fácil expansión.[3]

[3]No hace falta mencionar que esta plataforma y todos sus componentes están distribuidos bajo licencia libre en GNU LGPL.

En estricta comparación, por ejemplo, con $Jade$ [20], este último requiere ser programado directamente en $Java$, siendo más bien un paquete de clases y funciones de utilería para el desarrollo de sistemas basados en agentes BDI empleando el lenguaje abstracto $AgentSpeak(L)$, mientras que $Jason$ es una plataforma completa que permite interpretar directamente dicho lenguaje abstracto. A diferencia de los diversos sistemas que implementan agentes BDI, $Jason$ es un lenguaje de más alto nivel.

5.5.1. Sintaxis de $Jason$

El Cuadro 5.1 muestra una gramática detallada de $Jason$. $\langle ATOM \rangle$ es un identificador que comienza con una letra minúscula o el carácter punto $(.)$; $\langle VAR \rangle$ es un identificador que comienza con letra mayúscula; $\langle NUMERO \rangle$ es cualquier entero o número de punto flotante; y $\langle CADENA \rangle$ es una cadena de carácteres delimitada por comillas.

Entre las principales diferencias sintácticas de $Jason$ y $AgentSpeak(L)$ encontramos que:

- En lugar de átomos, $Jason$ acepta literales.
- La sintaxis de $Jason$, a diferencia de $AgentSpeak(L)$, permite el uso de anotaciones para las literales. Así, por ejemplo, es posible indicar la fuente de las percepciones, etc. El costo es una unificación más complicada.
- Es posible etiquetar los planes empleando $@ \langle STRING \rangle$.
- Las anotaciones pueden ser usadas en la definición de funciones de selección más sofisticadas.

5.5.2. Semántica de $Jason/AgentSpeak(L)$

La semántica operacional, también definida para $Jason$ [19], está dada por un sistema de transición entre configuraciones que están definidas por una quíntupla $\langle ag, C, M, T, s \rangle$ donde:

- ag es un programa agente formado por creencias bs y planes ps.
- Una circunstancia C es una tripleta $\langle I, E, A \rangle$ donde I es el conjunto de intenciones $\{i, i', \ldots, n\}$ t.q. $i \in I$ es una pila de planes parcialmente

$$
\begin{array}{rcl}
agente & \rightarrow & (creencias_0|metas_0)^*\ planes \\
creencias_0 & \rightarrow & creencias\ reglas \\
creencias & \rightarrow & (literal\ .)^* \\
reglas & \rightarrow & (literal\ :-\ exprLog\ .)^* \\
metas_0 & \rightarrow & (!\ literal\ .)^* \\
planes & \rightarrow & (plan)^* \\
plan & \rightarrow & [@atomo]\ eventoDisp\ [:\ contexto]\ [<-cuerpo]\ . \\
eventoDisp & \rightarrow & (+|-)[!|?]literal \\
literal & \rightarrow & [\ ]atomo \\
contexto & \rightarrow & exprLog|true \\
exprLog & \rightarrow & exprLogSimple \\
 & & |\ not\ exprLog \\
 & & |\ exprLog\ \&\ exprLog \\
 & & |\ exprLog\ |\ exprLog \\
 & & |\ (exprLog) \\
exprLogSimple & \rightarrow & (literal|relExpr|\langle VAR\rangle) \\
cuerpo & \rightarrow & fbfCuerpo(;\ fbfCuerpo)^* \\
 & & |\ true \\
bfbCuerpo & \rightarrow & (!|?|+|-|-+)\ literal \\
 & & |\ atomo \\
 & & |\ \langle VAR\rangle \\
 & & |\ exprLog \\
atomo & \rightarrow & (\langle ATOMO\rangle|\langle VAR\rangle)[(listaTerms)][[listaTerms]] \\
listaTerms & \rightarrow & termino(,\ termino)^* \\
termino & \rightarrow & literal \\
 & & |\ lista \\
 & & |\ exprAritm \\
 & & |\ \langle VAR\rangle \\
 & & |\ \langle CADENA\rangle \\
lista & \rightarrow & [[termino(,\ termino)^*[|\ (lista|\langle VAR\rangle)]] \\
exprRel & \rightarrow & termRel(<|<=|>|>=|==|\backslash\backslash==|=)termRel \\
exprTerm & \rightarrow & literal|exprAritm \\
exprAritm & \rightarrow & termAritm(+|-)termAritm \\
termAritm & \rightarrow & factorAritm(*|/|div|mod)factorAritm \\
factorAritm & \rightarrow & Aritm[++factorAritm] \\
Aritm & \rightarrow & \langle NUMERO\rangle \\
 & & |\ \langle VAR\rangle \\
 & & |\ -Aritm \\
 & & |\ (exprAritm)
\end{array}
$$

Cuadro 5.1: Sintaxis de $Jason$

instanciados $p \in ps$; E es un conjunto de eventos $\{\langle te, i\rangle, \langle te', i'\rangle, \ldots, n\}$, t.q. te es un evento disparador y cada i es una intención (evento interno) o una intención vacío $\top$ (evento externo); y A es un conjunto de acciones para ser ejecutadas por el agente en su ambiente.

- M es una tripleta $\langle In, Out, SI\rangle$ que funciona, para el caso multi-agente, como un buzón donde In es el buzón de entrada de un agente, Out es una lista de mensajes a entregarse y SI es un registro de intenciones suspendidas (intenciones que esperan por un mensaje de respuesta).

- T es una quíntupla $\langle R, Ap, \iota, \epsilon, \rho\rangle$ que registra información temporal: R es el conjunto de planes relevantes dado cierto evento disparador; Ap es el conjunto de planes aplicables (el subconjunto de R t.q. $bs \models ctx$); ι, ϵ y ρ registran, respectivamente, la intención, el evento y plan actual durante una ejecución.

- La etiqueta $s \in \{SelEv, RelPl, AppPl, SelAppl, SelInt, AddIM, ExecInt, ClrInt, ProcMsg\}$ indica el paso que está llevando a cabo el agente en un ciclo de razonamiento.

El intérprete para $AgentSpeak(L)$ se muestra como un sistema de transición en la Figura 5.1. El estado inicial es $s = ProcMsg$. Las reglas de la semántica operacional [19] definen el sistema de transición.

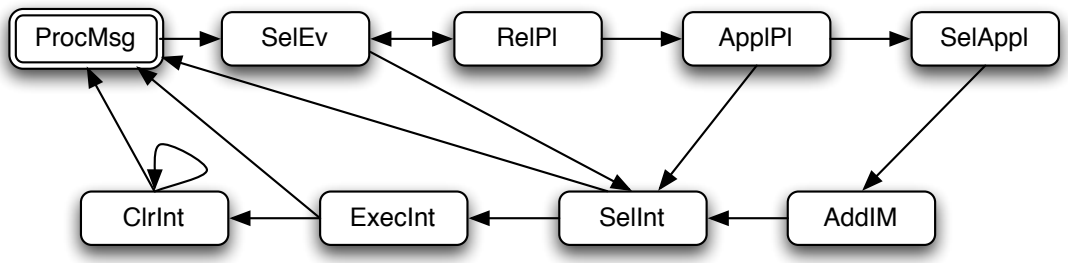


Figura 5.1: El intérprete de $Jason/AgentSpeak(L)$ como un sistema de transición

En lo que sigue adoptamos de $AgentSpeak(L)$ las definiciones de plan relevante, plan aplicable, unificador relevante, unificador aplicable, etc. Para efectos de describir las reglas semánticas adoptamos la siguiente notación:

- C denota una configuración de $Jason$; para hacer referencia al componente E (eventos) de C escribimos C_E. De manera similar accesamos a los demás componentes de C.
- Para indicar que no hay ninguna intención siendo considerada en la ejecución del agente empleamos $C_I = \emptyset$. De forma similar para C_P y C_ϵ. C_- indica que el componente en cuestión ha sido limpiado.
- Se usa $i[p]$ para denotar a la intención i que tiene al plan p como tope.

Si asumimos que p es un plan de la forma $t : \phi \leftarrow h$, entonces: $TrEv(p) = t$ y $Ctx(p) = \phi$. A continuación revisaremos las reglas de transición propuestas por Moreira [118] en su extensión a $AgentSpeak(L)$, pero antes tenemos algunas definiciones auxiliares:

Definición 32 (Planes relevantes) *El conjunto de planes relevantes con respecto a un evento disparador te está dado por:*

$$RelPlanes(ps, te) = \{p\theta | p \in ps \wedge \theta = MGU(te, TrEv(p))\}$$

Definición 33 (Planes aplicables) *Dado un conjunto R de planes relevantes, el conjunto de planes aplicables está dado por:*

$$AppPlanes(creencias, R) = \{p\theta | p \in R \wedge \theta \ t.q. \ creencias \models Ctx(p)\theta\}$$

Definición 34 (Test) *Dadas las creencias de un agente y una fbf at el conjunto de substituciones para probar la fbf contra las creencias está dado por:*

$$Test(creencias, at) = \{\theta | creencias \models at\theta\}$$

Las funciones de selección de $AgentSpeak(L)$ son denotadas aquí por S_E, S_{Ap} y S_I. En estos términos, la selección de un evento se computa como:

$$\textbf{SelEnv} \quad \frac{S_E(C_E) = (te, i)}{C, creencias \rightarrow C', creencias} \quad C_\epsilon = \_, C_{Ap} = C_R = \emptyset$$
$$donde: \quad C'_E = C_E \backslash (te, i), C'_\epsilon = (te, i)$$

La regla Rel_1 asigna a R el conjunto de planes relevantes. Si no existe ningún plan relevante, el evento es descartado de ε por la regla Rel_2.

$$\mathbf{Rel_1} \quad \frac{RelPlans(plans, te) \neq \emptyset}{C, creencias \rightarrow C', creencias} \quad C_\epsilon = (te, i), C_{Ap} = C_R = \emptyset$$
$$donde: \quad C'_R = RelPlans(plans, te)$$

$$\mathbf{Rel_2} \quad \frac{RelPlans(plans, te) = \emptyset}{C, creencias \rightarrow C', creencias} \quad C_\epsilon = (te, i), C_{Ap} = C_R = \emptyset$$
$$donde: \quad C'_R = \emptyset$$

El caso de los planes aplicables es parecido:

$$\mathbf{Appl_1} \quad \frac{AppPLans(C_R, creencias) \neq \emptyset}{C, creencias \rightarrow C', creencias} \quad C_\epsilon \neq \emptyset, C_{Ap} = \emptyset, C_R \neq \emptyset$$
$$donde: \quad C'_R = \emptyset$$
$$C'_{Ap} = AppPlanes(C_R, creencias)$$

$$\mathbf{Appl_2} \quad \frac{AppPLans(C_R, creencias) = \emptyset}{C, creencias \rightarrow C', creencias} \quad C_\epsilon \neq \emptyset, C_{Ap} = \emptyset, C_R \neq \emptyset$$
$$donde: \quad C'_R = \emptyset, C'_\epsilon = \emptyset$$

La siguiente regla asume la existencia de una función de selección S_{Ap}, la cual selecciona un plan a partir del conjunto Ap de planes aplicables.

$$\mathbf{SelAppl} \quad \frac{S_{Ap}(C_{Ap}) = p}{C, beliefs \rightarrow C', beliefs} \quad C_\epsilon \neq \emptyset, C_{Ap} \neq \emptyset$$
$$donde: \quad C'_\rho = p, C'_{Ap} = \emptyset$$

Recordemos que en $Jason$ se distinguen dos tipos de eventos, internos y externos. La regla $ExtEv$ procesa los eventos externos:

$$\textbf{ExtEv} \quad \frac{}{C, creencias \rightarrow C', creeencias} \quad C_\epsilon = (te, T), C_p = p$$
$$donde: \quad C'_I = C_I \cup \{[p]\}$$
$$C'_\epsilon = \emptyset, C'_p = \emptyset$$

Si el evento es interno, la regla $IntEv$ lo procesará:

$$\textbf{IntEv} \quad \frac{}{C, creencias \rightarrow C', creencias} \quad C_\epsilon = (te, i), C_p = p$$
$$donde: \quad C'_I = C_I \cup \{i[p]\}$$
$$C'_\epsilon = \emptyset, C'_p = \emptyset$$

La regla para seleccionar una intención a ser ejecutada es como sigue:

$$\textbf{SelInt} \quad \frac{S_I(C_I) = i}{C, creencias \rightarrow C', creencias} \quad C_\iota = \emptyset$$
$$donde: \quad C'_\iota = i$$

El grupo de reglas que describiremos a continuación expresa el efecto de la ejecución de los planes. El plan siendo ejecutado es siempre aquel que se encuentra en el tope de la intención que ha sido previamente seleccionada. Todas las reglas en este grupo terminan descartando i, por lo que otra intención puede ser seleccionada posteriormente. Las reglas se ejecutan dependiendo del componente del cuerpo del plan que se ha seleccionado:

$$\textbf{Action} \quad \frac{}{C, creencias \rightarrow C', creencias} \quad C_\iota = i[head \leftarrow a; h]$$
$$donde: \quad C'_\iota = \_, C'_A = C_A \cup \{a\}$$
$$C'_I = (C_I \backslash \{C_\iota\}) \cup \{i[head \leftarrow h]\}$$

La siguiente regla registra una nueva meta de tipo $achieve$, la cual también podrá ser seleccionada dada la regla $SelEv$:

$$\textbf{Achieve} \quad \frac{}{C, creencias \rightarrow C', creencias} \quad C_\iota = i[head \leftarrow !at; h]$$
$$donde: \quad C'_\iota = \_, C'_E = C_E \cup \{(+!at, C_\iota)\}$$
$$C'_I = C_I \backslash \{C_\iota\}$$

Nótese que la intención que generó el evento interno es removida del conjunto de intenciones C_I. Esto implementa la suspensión de una intención. Sólo cuando el curso de acción definida ha terminado se puede continuar con la ejecución de la intención que había sido suspendida, a partir de la siguiente fórmula del cuerpo de un plan dado.

Las metas de tipo *test* se procesan mediante las siguientes dos reglas:

$$\textbf{Test}_\textbf{1} \quad \frac{Test(creencias, \phi) = \emptyset}{C, creencias \rightarrow C', creencias} \quad C_\iota = i[head \leftarrow ?at; h]$$
$$donde: \quad C'_\iota = \emptyset$$
$$C'_I = C_I\{C_\iota\} \cup \{i[head \leftarrow h]\}$$

$$\textbf{Test}_\textbf{2} \quad \frac{Test(creencias, \phi) \neq \emptyset}{C, creencias \rightarrow C', creencias} \quad C_\iota = i[head \leftarrow ?at; h]$$
$$donde: \quad C'_\iota = \emptyset, C'_I = C_I\{C_\iota\} \cup \{i[(head \leftarrow h)\theta]\}$$
$$\theta \in Test(creencias, \phi)$$

Al igual que en dMARS [92], los agentes en $Jason$ pueden agregar o eliminar creencias durante la ejecución de sus planes. Las siguientes reglas se encargan de ello:

$$\textbf{AddBel}\quad \frac{}{C, creencias \rightarrow C', creencias'} \quad C_\iota = i[head \leftarrow +at; h]$$
$$donde: \quad C'_\iota = \emptyset, creencias \models at$$
$$C_E \cup \{(+at, C_\iota)\}$$
$$C'_I = C_I\{C_\iota\} \cup \{i[head \leftarrow h']\}$$

$$\textbf{DelBel}\quad \frac{}{C, creencias \rightarrow C', creencias'} \quad C_\iota = i[head \leftarrow -at; h]$$
$$donde: \quad C'_\iota = \emptyset, creencias \not\models at$$
$$C_E \cup \{(-at, C_\iota)\}$$
$$C'_I = C_I\{C_\iota\} \cup \{i[head \leftarrow h']\}$$

Para concluir con la semántica operacional de $Jason$ se definen dos reglas más, las llamadas *clearing house rules*. $ClearInt_1$ simplemente remueve una intención del conjunto de intenciones de un agente cuando no hay más que hacer al respecto, es decir, cuando ya no quedan más fórmulas (acciones o metas) a ejecutar dentro del cuerpo del plan.

$$\textbf{ClearInt}_\textbf{1}\quad \frac{}{C, creencias \rightarrow C', creencias'} \quad C_\iota = i[head \leftarrow]$$
$$donde: \quad C'_\iota = \emptyset, C'_I = C_I\{C_\iota\}$$

La segunda regla de limpieza procesa las intenciones que han sido suspendidas:

$$\textbf{ClearInt}_\textbf{2}\quad \frac{}{C, creencias \rightarrow C', creencias'} \quad C_\iota = i'[head' \leftarrow !at; h1 \ddagger head \leftarrow]$$
$$donde: \quad C'_\iota = \emptyset$$
$$C'_I = C_I\{C_\iota\} \cup \{i'[head' \leftarrow h']\}$$

5.6. Resumen

Hemos expuesto la sintaxis y la semántica del lenguaje $AgentSpeak(L)$ junto con su intérprete $Jason$. Como hemos visto, este lenguaje ha sido propuesto y usado para reducir la laguna entre la teoría (la especificación lógica BDI) y la práctica (la implementación). En este caso $AgentSpeak(L)/Jason$ tiene una semántica operacional bien definida, pero la verificación de propiedades racionales de los agentes programados no es evidente, como puede apreciarse, pues el costo de esta excelente semántica operacional es la exclusión del uso de las modalidades que hacen a los sistemas lógicos BDI lenguajes altamente expresivos.

Con este capítulo terminamos la primera parte del trabajo que podemos resumir de la siguiente forma: desde un punto de vista cognitivo un agente BDI es un sistema agente que se explica en términos del modelo BDI. Un modelo BDI bratmaniano de razonamiento intencional debe considerar los atributos lógicos de revisión de intenciones y la no-monotonicidad del razonamiento intencional. El problema que quedaba abierto era el de cómo traducir los atributos del razonamiento intencional a un modelo formal. Los modelos tradicionales, a pesar de sus ventajas, tienen problemas técnicos y profundos. Para resolver estos problemas técnicos se han propuesto lenguajes formales y de programación.

Aunque todavía no se han resuelto estos problemas consideramos que, por lo pronto, ya hemos aclarado los componentes preliminares de nuestra investigación.

Parte II

Avances y resultados

Capítulo 6

El papel de $BDI^{CTL}_{AgentSpeak(L)}$

6.1. Introducción

Los sistemas lógicos y el lenguaje de programación que hemos revisado previamente han sido propuestos, respectivamente, para modelar e implementar el comportamiento lógico de los agentes BDI a partir de un modelo bratmaniano.

Los sistemas lógicos más utilizados para hacer tales cosas han sido las lógicas BDI o sus implementaciones [137, 151, 165]. Hemos visto cuál es el problema técnico y cuál es el problema profundo de esta familia de sistemas.

Además hemos revisado $AgentSpeak(L)$ junto con su intérprete $Jason$ y, como puede apreciarse, el costo de su excelente semántica es la exclusión del uso de las modalidades que hacen a los sistemas lógicos BDI sistemas altamente expresivos.

Por estas razones proponemos, para razonar acerca de los agentes BDI dentro de un modelo bratmaniano y resolver el problema técnico, el formalismo que denominamos como $BDI^{CTL}_{AgentSpeak(L)}$. Utilizaremos esta herramienta lógica para desarrollar la especificación formal y la verificación de agentes BDI programados en $AgentSpeak(L)$.[1]

Inicialmente este sistema es similar a un sistema BDI^{CTL} definido en términos de $B^{KD45}D^{KD}I^{KD}$ con los operadores temporales usuales: *next* ($\bigcirc$), *eventually* ($\Diamond$), *always* ($\Box$), *until* (U), *optional* (E), *inevitable* (A), definidos mediante $CTL*$ [40, 54]. Pero nuestra principal contribución está en la definición de la

[1]Esta idea, por cierto, no es completamente nueva: enfoques similares se han propuesto para otros lenguajes de programación, como por ejemplo, $3APL$ [46]; lo que sí es nuevo, no obstante, es la integración con $AgentSpeak(L)$ y su aplicación en un sistema no-monotónico.

semántica de los operadores temporales CTL en términos de una estrutuctura de Kripke [88] inducida por el sistema de transición de la semántica operacional de $AgentSpeak(L)$. La semántica de los operadores intencionales es adoptada de Bordini *et al* [18]. Como resultado, la semántica de $BDI^{CTL}_{AgentSpeak(L)}$ estará basada en la semántica operacional del lenguaje de programación que usamos.[2]

El capítulo está organizado del siguiente modo: en § 6.2 recordamos las estrategias de compromiso tal como y son especificadas en el sistema lógico BDI. En § 6.3 recordamos, brevemente, la sintaxis y la semántica de $AgentSpeak(L)$ tal y como la conocemos. En § 6.4 veremos nuestra propuesta y el papel de $BDI^{CTL}_{AgentSpeak(L)}$ como una herramienta para conectar la teoría (especificación lógica) con la práctica (implementación) y resolver el problema técnico y el filosófico (§ 6.5).

6.2. Estrategias de compromiso

Ahora sabemos que diferentes teorías formales han sido propuestas para capturar las ideas de Bratman. El trabajo fundacional de Cohen y Levesque, por ejemplo, definió las intenciones como una combinación de creencias y deseos usando el concepto de meta persistente. Un análisis crítico de dicha teoría muestra que no logra capturar ciertos aspectos importantes de la semántica de las intenciones y el compromiso. El compromiso, a su vez, ha sido tratado como un proceso de mantenimiento y revisión de intenciones tomando en cuenta intenciones presentes y futuras.

Como vimos previamente, diferentes tipos de compromiso definen diferentes tipos de agentes. Tres estrategias han sido ampliamente estudiadas en el contexto de BDI_{CTL}:

[2] De este modo podremos probar que cualquier agente programado en $AgentSpeak(L)$ satisface ciertas propiedades expresadas en la especificación lógica. Es importante tener en cuenta que esta aportación es diferente de la del *model checking* [40] tradicional en el siguiente sentido: en el *model checking* tradicional el problema consiste en verificar si cierta propiedad se cumple en cierto estado para cierto agente, mientras que en nuestro trabajo trataremos de verificar propiedades generales que se cumplan para cualquier agente: la generalización y la universalidad son resultados deseables en un trabajo de lógica.

- El compromiso de tipo *blind*:

$$\mathsf{INT}(\mathsf{A}\Diamond\phi) \implies \mathsf{A}(\mathsf{INT}(\mathsf{A}\Diamond\phi)\ \mathsf{U}\ \mathsf{BEL}(\phi))$$

- El compromiso *single-minded*:

$$\mathsf{INT}(\mathsf{A}\Diamond\phi) \implies \mathsf{A}(\mathsf{INT}(\mathsf{A}\Diamond\phi)\ \mathsf{U}\ (\mathsf{BEL}(\phi) \vee \neg\mathsf{BEL}(\mathsf{E}\Diamond\phi)))$$

- Y el *open-minded*:

$$\mathsf{INT}(\mathsf{A}\Diamond\phi) \implies \mathsf{A}(\mathsf{INT}(\mathsf{A}\Diamond\phi)\ \mathsf{U}\ (\mathsf{BEL}(\phi) \vee \neg\mathsf{DES}(\mathsf{A}\Diamond\phi)))$$

Ejemplo 6 *Estas tres estrategias definen, respectivamente, tres tipos de agentes: ciegos, determinados y abiertos. Así, por ejemplo, un agente ciego que intenta en algún momento (eventually) ir a París, mantendrá tal intención, para cualquier curso de acción, hasta que crea que está rumbo a París. Un agente determinado, por otro lado, puede abandonar su intención de ir a París si llega a creer que ya no es posible ir a París. Finalmente, un agente abierto puede abandonar la intención de ir a París si de pronto deja de desear ir a París.*

Evidentemente, si un agente es ciego, no podemos hablar de ningún tipo de revisión o no-monotonicidad. Pero si el agente es determinado o abierto, podemos entonces hacer dos cosas: suponer que los razonamientos son derrotables (problema interno) y aproximar una revisión a través de funciones de aprendizaje (problema externo), como veremos en próximos capítulos.

6.3. $AgentSpeak(L)$

6.3.1. Sintaxis de $AgentSpeak(L)$

La sintaxis de $AgentSpeak(L)$ [136], tal y como es definida para su intérprete $Jason$ [19], se muestra en el cuadro 6.1.

ag	::=	$bs \quad ps$		at	::=	$P(t_1, \ldots, t_n)$ $(n \geq 0)$	
bs	::=	$b_1 \ldots b_n$	$(n \geq 0)$			$\mid P(t_1, \ldots, t_n)[s_1, \ldots, s_m] (n \geq 0, m \geq 0)$	
ps	::=	$p_1 \ldots p_n$	$(n \geq 1)$	s	::=	`percept` $\mid$ `self` $\mid id$	
p	::=	$te : ct \leftarrow h$		a	::=	$A(t_1, \ldots, t_n)$ $(n \geq 0)$	
te	::=	$+at \mid -at \mid +g \mid -g$		g	::=	$!at \mid ?at$	
ct	::=	$ct_1 \mid \top$		u	::=	$+b \mid -b$	
ct_1	::=	$at \mid \neg at \mid ct_1 \wedge ct_1$					
h	::=	$h_1; \top \mid \top$					
h_1	::=	$a \mid g \mid u \mid h_1; h_1$					

Cuadro 6.1: Sintaxis de $AgentSpeak(L)$

6.3.2. Semántica de $AgentSpeak(L)$

A continuación revisitamos la semántica operacional de $AgentSpeak(L)$ para su intérprete $Jason$ en términos de un sistema de transición, pero de un modo más compacto: mostrando sólo las reglas relevantes para los propósitos de este capítulo.

Un sistema de transición es un conjunto de reglas de transformación que van de un estado a otro [127]. Cada regla de transformación tiene la forma:

$$\frac{cond}{C \rightarrow C'}$$

donde C es una configuración o estado que puede ser transformado al estado C' si la condición $cond$ se cumple.

Asumimos las definiciones auxiliares de plan relevante y plan aplicable, tal como quedaron definidas en el capítulo anterior. Asumimos también el sistema de transición de $AgentSpeak(L)$ (Figura 6.1). En el Cuadro 6.2 recordamos sólo las reglas relevantes para este capítulo.

Así pues, aunque la semántica de este lenguaje define claramente el razonamiento realizado por el agente, es difícil probar propiedades BDI tales como las estrategias de compromiso para cualquier agente. Esto se debe al abandono de las modalidades temporales e intencionales en $AgentSpeak(L)$. No obstante, la semántica operacional nos permite definir una ejecución en $AgentSpeak(L)$ del siguiente modo:

Definición 35 *(Ejecución) Una ejecución o corrida en $AgentSpeak(L)$ es un conjunto*

$$Corrida = \{(c_i, c_j) | \Gamma \vdash c_i \rightarrow c_j \ y \ c_i, c_j \in C\}$$

donde Γ es el sistema de transición definido por la semántica operacional y C es un conjunto de configuraciones de agente.

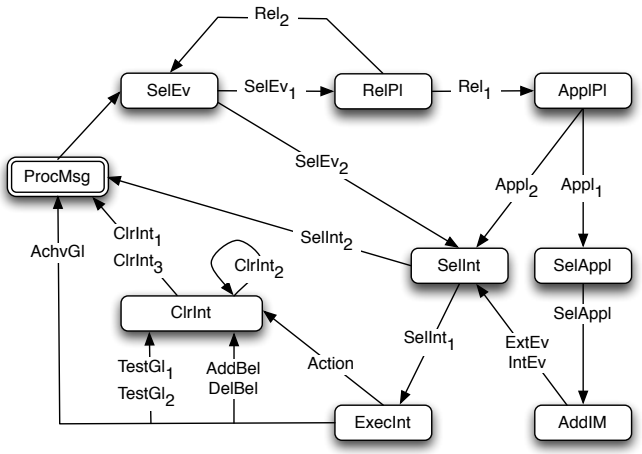


Figura 6.1: Sistema de transición $AgentSpeak(L)$

Como veremos, tal definición de ejecución nos permitirá inducir una estructura de Kripke.

6.4. $BDI^{CTL}_{AgentSpeak(L)}$

La meta ahora es definir la semántica de los operadores temporales e intencionales de $BDI^{CTL}_{AgentSpeak(L)}$ en términos de los operadores semánticos de $AgentSpeak(L)$ y el sistema BDI_{CTL}. Una vez hecho esto podremos usar el sistema para resolver tanto el problema técnico como el filosófico razonando, sin problemas, acerca de las propiedades de los agentes BDI representados mediante $AgentSpeak(L)$.

Por tanto, lo que necesitamos es un sistema lógico multimodal que use operadores temporales para lidiar con la ejecución de los agentes en el tiempo; y que use operadores intencionales para especificar los estados mentales de un agente.

6.4.1. Sintaxis de $BDI^{CTL}_{AgentSpeak(L)}$

Necesitamos, en primer lugar, una forma de expresar características intencionales, es decir, características BDI:

Definición 36 *Si ϕ es una fórmula atómica de $AgentSpeak(L)$, entonces* BEL(ϕ), DES(ϕ) *e* INT(ϕ) *son fórmulas bien formadas de $BDI^{CTL}_{AgentSpeak(L)}$.*

Para especificar la conducta temporal usamos $CTL*$ de la siguiente manera:

Definición 37 *Toda fórmula $BDI^{CTL}_{AgentSpeak(L)}$ es una formula de estado σ (por state); o una fórmula de camino π (por path):*

- $\sigma ::= \phi | \sigma \wedge \sigma | \neg\sigma$
- $\pi ::= \sigma | \neg\pi | \pi \wedge \pi | \mathsf{E}\pi | \mathsf{A}\pi | \bigcirc \pi | \Diamond\pi | \Box\pi | \pi \ \mathsf{U} \ \pi$

6.4.2. Semántica de $BDI^{CTL}_{AgentSpeak(L)}$

Inicialmente la semántica de BEL, DES e INT es adoptada de Bordini *et al* [18]. De esta manera asumimos la siguiente función que va de pilas de intenciones a conjuntos de fórmulas atómicas:

$$\begin{array}{lcll} agoals(\top) & = & \{\}, & \\ agoals(i[p]) & = & \begin{cases} \{at\} \cup agoals(i) & \text{si } p = +!at : ctx \leftarrow h, \\ agoals(i) & \text{de otro modo} \end{cases} & \end{array}$$

que nos devuelve el conjunto de fórmulas atómicas (at) adjuntas a una meta de tipo *achieve*($+!$) y donde $i[p]$ denota la pila de intenciones con p en el tope.

Definición 38 *Los operadores* BEL, DES *e* INT *se definen en términos de un agente ag y su configuración $\langle ag, C, M, T, s\rangle$:*

$$\mathsf{BEL}_{\langle ag,C,M,T,s\rangle}(\phi) \equiv \phi \in bs$$

$$\mathsf{INT}_{\langle ag,C,M,T,s\rangle}(\phi) \equiv \phi \in \bigcup_{i \in C_I} agoals(i) \text{ v } \phi \in \bigcup_{\langle te,i\rangle \in C_E} agoals(i)$$

$$\mathsf{DES}_{\langle ag,C,M,T,s\rangle}(\phi) \equiv \langle +!\phi, i\rangle \in C_E \text{ v } \mathsf{INT}(\phi)$$

donde C_I denota las intenciones activas y C_E las intenciones suspendidas.

Con estas definiciones podemos construir la semántica de este sistema usando una estructura de Kripke $K = \langle S, R, V\rangle$ donde S es el conjunto de configuraciones agente, R es una relación de acceso definida por el sistema de transición Γ y V es una función de valuación que va de las configuraciones agente a proposiciones verdaderas en esos estados.

Definición 39 *Sea* $K = \langle S, \Gamma, V\rangle$ *donde:*

- S *es un conjunto de configuraciones agente* $c = \langle ag, C, M, T, s\rangle$.
- $\Gamma \subseteq S^2$ *es una relación total t.q. para todo* $c \in S$ *existe un* $c' \in S$ *t.q.* $(c, c') \in \Gamma$.
- V *es una función de valuación t.q.:*
  - $V_{\mathsf{BEL}}(c, \phi) = \mathsf{BEL}_c(\phi)$ *donde* $c = \langle ag, C, M, T, s\rangle$.
  - $V_{\mathsf{DES}}(c, \phi) = \mathsf{DES}_c(\phi)$ *donde* $c = \langle ag, C, M, T, s\rangle$.
  - $V_{\mathsf{INT}}(c, \phi) = \mathsf{INT}_c(\phi)$ *donde* $c = \langle ag, C, M, T, s\rangle$.
- *Los caminos son series de configuraciones* $c_0, \ldots, c_n$ *t.q.* $\forall i (c_i, c_{i+1}) \in R$. *Además, si usamos* x^i *para indicar el i-ésimo estado del camino* x, *entonces:*

S1 $K, c \models \mathsf{BEL}(\phi) \Leftrightarrow \phi \in V_{\mathsf{BEL}}(c)$

S2 $K, c \models \mathsf{DES}(\phi) \Leftrightarrow \phi \in V_{\mathsf{DES}}(c)$

S3 $K, c \models \mathsf{INT}(\phi) \Leftrightarrow \phi \in V_{\mathsf{INT}}(c)$

S4 $K, c \models \mathsf{E}\phi \Leftrightarrow \exists x = c_1, \ldots \in K | K, x \models \phi$

S5 $K, c \models \mathsf{A}\phi \Leftrightarrow \forall x = c_1, \ldots \in K | K, x \models \phi$

P1 $K, c \models \phi \Leftrightarrow K, x^0 \models \phi$ *donde* ϕ *es una fórmula de estado.*

P2 $K, c \models \bigcirc\phi \Leftrightarrow K, x^1 \models \phi$.

P3 $K, c \models \Diamond\phi \Leftrightarrow K, x^n \models \phi$ *para algún* $n \geq 0$

P4 $K, c \models \Box\phi \Leftrightarrow K, x^n \models \phi$ *para todo* n

P5 $K, c \models \phi\ \mathsf{U}\ \psi \Leftrightarrow \exists k \geq 0$ *t.q.* $K, x^k \models \psi$ *y* $\forall j,\ 0 \leq j < k | K, c^j \models \phi$; *o bien* $\forall j \geq 0 : K, x^j \models \phi$

Es importante notar que la semántica de U corresponde a un *weak until* en donde ψ puede nunca ocurrir.

Para ilustrar algunas de estas definiciones entre configuraciones agente podemos ver los esquemas de las Figuras 6.2, 6.3 y 6.4.

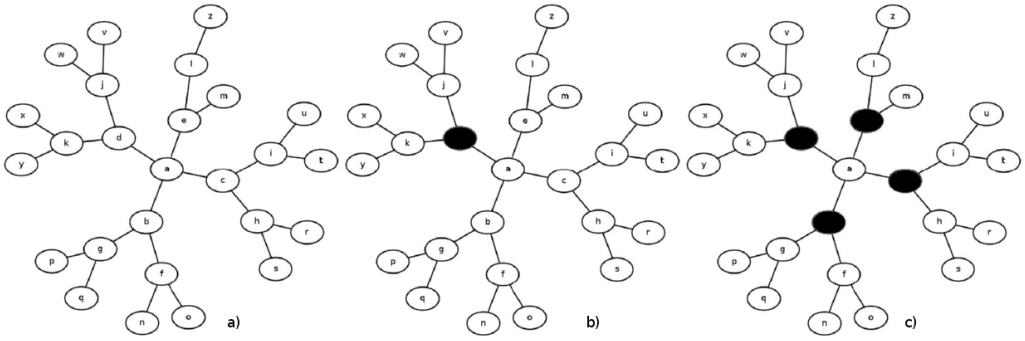


Figura 6.2: *a)* Configuración inicial, *b)* E$\Diamond\phi$, *c)* A $\bigcirc$ ϕ

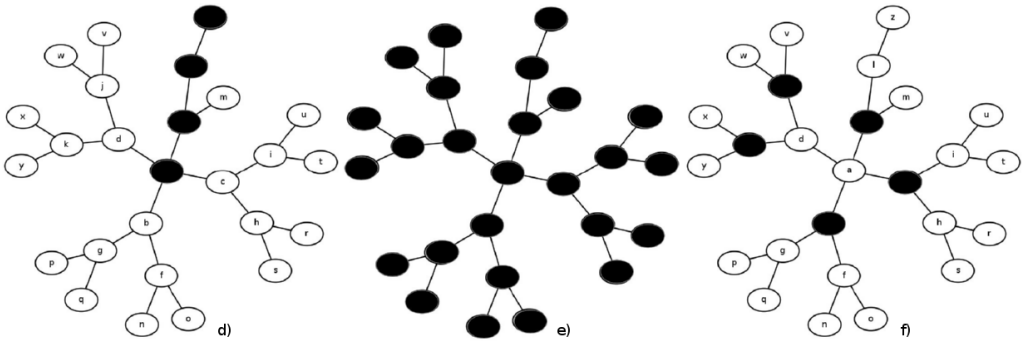


Figura 6.3: *d)* E$\Box\phi$, *e)* A$\Box\phi$, *f)* A$\Diamond\phi$

Ejemplo 7 *La fórmula bien formada*

$$\text{INT}(\text{A}\Diamond go(Paris))\ \text{U}\ \neg\text{BEL}(go(Paris, Summer))$$

expresa que un agente intentará inevitablemente, es decir, para todo curso de acción, en algún momento (eventually) ir a París en verano hasta que no crea que va a París en verano.

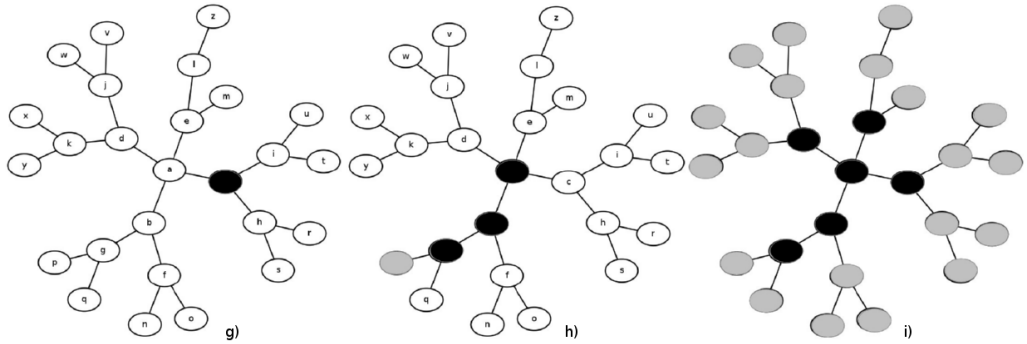


Figura 6.4: *g)* E $\bigcirc \phi$, *h)* E(ϕ U ψ), *i)* A(ϕ U ψ)

En la definición previa las fórmulas de estado son evaluadas en el modelo de Kripke K con la configuración c. Como es nuestro interés expresar que ciertas propiedades definidas en el lenguaje de especificación son satisfechas por cualquier agente BDI representado en $AgentSpeak(L)$ necesitamos definir una noción de satisfacción:

Definición 40 *(Satisfacción) Una fórmula ϕ es verdadera en $K = \langle S, R, V \rangle$ si y sólo si ϕ es verdadera en todas las configuraciones agente c en K. Esto es, $K, c \models \phi \Leftrightarrow K, c \models \phi$ para toda $c \in S$.*

Definición 41 *(Corrida de un agente en un modelo) Dada una configuración agente inicial β, un sistema de transición Γ y una valuación V, $K_\Gamma^\beta = \left\langle S_\Gamma^\beta, R_\Gamma^\beta, V \right\rangle$ denota una corrida de un agente en el modelo $K = \langle S, R, V \rangle$.*

Dado que nuestro interés es hallar que una propiedad general se cumple para cualquier agente es necesario extender la noción de modelo para todos los programas agentes:

Definición 42 *(Validez) Una fórmula $\phi \in BDI^{CTL}_{AgentSpeak(L)}$ es verdadera para toda corrida agente en Γ si y sólo si $\forall K_\Gamma^\beta \models \phi$*

6.5. Algunos resultados sobre agentes BDI

Con estos elementos podemos mostrar el papel de este formalismo mediante la presentación de algunos resultados parciales que, expresados en esta

especificación lógica, se cumplen para cualquier agente BDI representado en $AgentSpeak(L)$.

Empezamos verificando las estrategias ciega y determinada, pero antes mostramos que se satisface el axioma de no-retención infinita. Esto significa que los agentes BDI satisfacen el axioma de *no-infinite deferral*, o en otras palabras, que un agente BDI en algún momento termina de ejecutar sus intenciones:

Proposición 1 *Los agentes BDI satisfacen el axioma de no-infinite deferral:* $\mathsf{INT}(\phi) \Rightarrow \mathsf{A}\Diamond(\neg\mathsf{INT}(\phi))$.

Y lo que es más importante: los agentes BDI no sólo cumplen el axioma de no-retención infinita sino que no se comportan ciegamente y, por ello, tanto la no-monotonía como la revisión de intenciones son atributos lógicos legítimos del razonamiento intencional:

Proposición 2 *Los agentes BDI no satisfacen el axioma de compromiso ciego.*

Además, estos atributos lógicos no sólo son legítimos, sino que son compatibles, ya que exhiben una forma de compromiso determinado:

Proposición 3 *Los agentes BDI satisfacen el compromiso determinado limitado:* $\mathsf{INT}(\mathsf{A}\Diamond\phi) \Rightarrow \mathsf{A}(\mathsf{INT}(\mathsf{A}\Diamond\phi)\ \mathsf{U}\ \neg\mathsf{BEL}(\mathsf{E}\Diamond\phi))$

Es preciso mencionar que esta última proposición muestra una forma de compromiso determinado limitado porque el término $\neg\mathsf{BEL}(\mathsf{E}\Diamond\phi)$ no es representado explícitamente por el agente. De hecho, el agente no puede continuar intentando ϕ porque no hay planes para hacerlo y la intención completa falla.

Por tanto, dado que los agentes BDI no están comprometidos ciegamente y satifacen una forma limitada de compromiso determinado, el aprendizaje intencional automático [76, 77, 79] provee una alternativa para lograr una forma completa de compromiso determinado. La idea es que los agentes pueden aprender, de la misma manera en que aprenden la adopción de planes exitosos, las razones detrás de la adopción de un plan que ha fracasado.

Una semántica operacional extendida de $AgentSpeak(L)$ que implica aprendizaje, por métodos incrementales e inductivos, ha sido propuesta por Guerra *et al* [79]. Se inspira en la manera en que $Jason$ es extendido con actos de habla: añadiendo reglas operacionales que se implementan en una librería de planes. Usando las técnicas presentadas aquí es posible configurar reglas como la

siguiente:

$$(\textbf{Abandon}) \quad \frac{\mathcal{S}_{\mathcal{E}}(C_E) = \langle +abandon(\phi), \top \rangle \wedge ag_{bs} \models intending(\phi)}{\langle ag, C, M, T, SelEv \rangle \rightarrow \langle ag', C', M, T, SelEv \rangle}$$

t.q. $C'_E = C_E \backslash \{\langle +abandon(\phi), \top \rangle\}, ag_{bs} \not\models intending(\phi), C'_I = C_I \backslash \phi$

Esta regla expresa que cuando un agente está intentando ϕ y un evento con el disparador $+abandon(\phi)$ es generado, la intención es removida de C_I, el evento es removido de C_E y el agente no cree $intending(\phi)$. Con esta regla podemos ver que el compromiso determinado se hace más fuerte:

Proposición 4 *Sea Ag el conjunto de agentes BDI. El compromiso determinado se satisface por $Ag \cup$ Abandon.*

Esto permite aproximar una forma completa de compromiso determinado por medio del aprendizaje intencional, porque ahora el agente tiene una representación explícita de $\neg$BEL(E$\Diamond\phi$).

6.6. Resumen

Hemos desarrollado y usado $BDI^{CTL}_{AgentSpeak(L)}$ como una herramienta lógica para la especificación formal de agentes BDI integrados en $AgentSpeak(L)$. Hay dos resultados parciales, pero interesantes, que pueden ser extraídos a partir de este capítulo. El primero es que con este formalismo podemos evitar el problema técnico-ingenieril del que veníamos hablando; el segundo, que el cambio de intenciones no es sólo compatible con los agentes BDI modelados vía $AgentSpeak(L)$ (dado que no tienen compromiso ciego y tienen una forma de compromiso determinado), sino que los atributos de no-monotonicidad y revisión tienen un respaldo formal.

Estos dos resultados, si bien son parciales, son importantes porque nos permiten generalizar propiedades que han sido consideradas como fundamentales para un agente BDI. Al formalizar y verificar estas propiedades hemos probado que se cumplen para cualquier agente BDI interpretado en términos de $AgentSpeak(L)$, lo cual es relevante porque acerca a $AgentSpeak(L)$ a sus fundamentos filosóficos mientras ofrece un lenguaje rico para expresar y verificar sus propiedades generales.

En resumen, sabemos que desde un punto de vista cognitivo un agente BDI es un sistema agente que se explica en términos del modelo BDI. Un modelo BDI bratmaniano de razonamiento intencional debe considerar los atributos lógicos de revisión de intenciones y no-monotonicidad del razonamiento intencional: no hacerlo es construir un modelo no-bratmaniano incapaz de resolver los problemas externo e interno. El problema que quedaba abierto era, sin embargo, el de cómo traducir el análisis de las intenciones en un modelo formal, puesto que los modelos tradicionales, a pesar de sus ventajas, tienen problemas técnicos-ingenieriles y filosóficos. Y aunque para resolver estos problemas se han propuesto lenguajes formales y de programación, todavía no se resuelven por completo. Sin embargo, con el desarrollo y uso de $BDI^{CTL}_{AgentSpeak(L)}$ podemos evitar tanto el problema técnico como el filosófico al tiempo que observamos que la revisión y la no-monotonicidad intencional no sólo son atributos lógicos legítimos, sino que son compatibles con nuestra comprensión de los agentes BDI.

$(\mathbf{SelEv_1})$	$\frac{\mathcal{S}_{\mathcal{E}}(C_E)=\langle te,i\rangle}{\langle ag,C,M,T,SelEv\rangle \longrightarrow \langle ag,C',M,T',RelPl\rangle}$	t.q.	$C'_E = C_E \setminus \{\langle te,i\rangle\}$ $T'_\epsilon = \langle te,i\rangle$
$(\mathbf{Rel_1})$	$\frac{T_\epsilon=\langle te,i\rangle, RelPlans(ag_{ps},te)\neq\{\}}{\langle ag,C,M,T,RelPl\rangle \longrightarrow \langle ag,C,M,T',AppPl\rangle}$	t.q.	$T'_R = RelPlans(ag_{ps},te)$
$(\mathbf{Rel_2})$	$\frac{RelPlans(ps,te)=\{\}}{\langle ag,C,M,T,RelPl\rangle \longrightarrow \langle ag,C,M,T,SelEv\rangle}$		
$(\mathbf{Appl_1})$	$\frac{ApplPlans(ag_{bs},T_R)\neq\{\}}{\langle ag,C,M,T,ApplPl\rangle \longrightarrow \langle ag,C,M,T',SelAppl\rangle}$	t.q.	$T'_{Ap} = AppPlans(ag_{bs},T_R)$
$(\mathbf{SelAppl})$	$\frac{\mathcal{S}_{\mathcal{O}}(T_{Ap})=(p,\theta)}{\langle ag,C,M,T,SelAppl\rangle \longrightarrow \langle ag,C,M,T',AddIM\rangle}$	t.q.	$T'_\rho = (p,\theta)$
$(\mathbf{ExtEv})$	$\frac{T_\epsilon=\langle te,\top\rangle, T_\rho=(p,\theta)}{\langle ag,C,M,T,AddIM\rangle \longrightarrow \langle ag,C',M,T,SelInt\rangle}$	t.q.	$C'_I = C_I \cup \{[p\theta]\}$
$(\mathbf{SelInt_1})$	$\frac{C_I\neq\{\},\mathcal{S}_{\mathcal{I}}(C_I)=i}{\langle ag,C,M,T,SelInt\rangle \longrightarrow \langle ag,C,M,T',ExecInt\rangle}$	t.q.	$T'_\iota = i$
$(\mathbf{SelInt_2})$	$\frac{C_I=\{\ \}}{\langle ag,C,M,T,SelInt\rangle \longrightarrow \langle ag,C,M,T,ProcMsg\rangle}$		
$(\mathbf{AchvGl})$	$\frac{T_\iota=i[head\leftarrow !at;h]}{\langle ag,C,M,T,ExecInt\rangle \longrightarrow \langle ag,C',M,T,ProcMsg\rangle}$	t.q.	$C'_E = C_E \cup \{\langle +!at,T_\iota\rangle\}$ $C'_I = C_I \setminus \{T_\iota\}$
$(\mathbf{ClrInt_1})$	$\frac{T_\iota=[head\leftarrow\top]}{\langle ag,C,M,T,ClrInt\rangle \longrightarrow \langle ag,C',M,T,ProcMsg\rangle}$	t.q.	$C'_I = C_I \setminus \{T_\iota\}$
$(\mathbf{ClrInt_2})$	$\frac{T_\iota=i[head\leftarrow\top]}{\langle ag,C,M,T,ClrInt\rangle \longrightarrow \langle ag,C',M,T,ClrInt\rangle}$	t.q.	$C'_I = (C_I \setminus \{T_\iota\}) \cup$ $\{k[(head' \leftarrow h)\theta]\}$ si $i = k[head' \leftarrow g;h]$ y $g\theta = TrEv(head)$
$(\mathbf{ClrInt_3})$	$\frac{T_\iota\neq[head\leftarrow\top]\wedge T_\iota\neq i[head\leftarrow\top]}{\langle ag,C,M,T,ClrInt\rangle \longrightarrow \langle ag,C,M,T,ProcMsg\rangle}$		

Cuadro 6.2: Reglas de la semántica operacional de $AgentSpeak(L)$

Capítulo 7

El papel del aprendizaje

7.1. Introducción

En el capítulo anterior propusimos un sistema formal diseñado para resolver el problema técnico de la relación entre implementación y especificación; y el problema filosófico de la legitimidad de los atributos lógicos de revisión y no-monotonicidad. Sin embargo, a partir de este capítulo nos separaremos un momento de la problemática anterior (problema interno) para discutir otro elemento que también juega un papel importante en esta investigación: la importancia del aprendizaje en un modelo bratmaniano. Por ello, a partir de este capítulo comenzamos a enfrentarnos al problema externo. En particular, lo que haremos durante este capítulo será mostrar unos resultados experimentales cuya finalidad será evidente.

Es preciso recordar, como preámbulo a este recurso experimental, que de acuerdo a la lectura de la propuesta bratmaniana (Capítulo 3), en la revisión de intenciones hay que tomar en cuenta dos aspectos: los problemas para los planes y la reconsideración ocasional. Por el momento únicamente nos concentramos en el primer aspecto porque es más general: cuando el mundo esperado no es el mundo actual, nuestras actitudes intencionales pueden fallar, por lo que no sólo hay una exigencia de derrotabilidad sino también un problema de revisión que, como trataremos de mostrar, conlleva funciones de aprendizaje intencional.

Lo que pretendemos mostrar con estos resultados experimentales es que es posible relacionar el concepto de revisión de intenciones con el concepto de compromiso determinado mediante el aprendizaje intencional.

En § 7.2 exponemos un marco experimental para visualizar el papel del apren-

dizaje intencional. Posteriormente en § 7.3 describimos el material y los métodos. Finalmente describimos los resultados del experimento (§ 7.4).

7.2. Marco de experimentación

Supongamos que tenemos un agente situado en un mundo de bloques que percibe el estado *a* (*State-a*) mostrado en la Figura 7.1 y que desea alcanzar el estado *b* (*State-b*). De este modo, el agente forma una intención de la forma:

$$\langle +!on(b, c), \top \rangle$$

usando un plan relevante y aplicable. Supongamos, además, que dicho plan tiene la forma:

$$+!on(X, Y) \leftarrow .take(X); .put(X, Y).$$

Esto significa que el agente es ingenuo (*bold* o *naive*) con respecto a la acción de apilar objetos: cree que X puede apilarse sobre Y en cualquier contexto—o lo que es lo mismo, sin importar el contexto—, absolutamente siempre. Un agente de este tipo es, en términos de estrategias de compromiso, un agente ciego.

Ahora, supongamos que después de formar dicho plan el agente percibe el estado c (*State-c*). Entonces la acción interna $.put(X = a, Y = b)$ fallará, por lo que la intención en algún momento fallará y la meta $on(b, c)$ no será alcanzada. Un agente con aprendizaje intencional puede hallar que el contexto correcto del plan fallido es:

$$clear(Y) \wedge handfree(Ag)$$

y modificar su plan de acuerdo a tal contexto. Así, la próxima vez que procesa un evento $\langle +!on(b, c), \top \rangle$ en un estado similar a *State-c*, será el caso que $bs \not\models clear(c)$, y por tanto el plan ya no será aplicable. Un agente capaz de hacer esto usando aprendizaje intencional tiene un enfoque mejor, al menos en términos de efectividad, que el agente ingenuo, si bien no resuelve del todo el problema.

Ahora bien, así como podemos añadir las razones de éxito del plan en cuestión, como en el caso anterior, también podemos considerar el fallo del plan para añadir creencias como:

$$abandon(p) \leftarrow not(clear(Y))$$

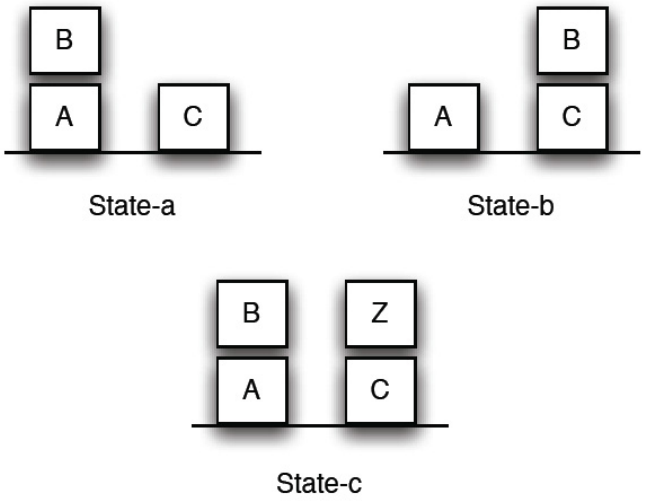


Figura 7.1: Mundo de bloques

que permiten un abandono previsorio de las intenciones a través de un mecanismo de limpiado. De este modo, por ejemplo, si tal creencia aparece en el estado mental del agente mientras el plan p está siendo intentado, el agente tiene que abandonar dicha intención (y posiblemente volver a colocar los eventos $!clear(c)$ y $!on(b, c)$). Si esto no sucede, el agente puede continuar como un agente normal. Esta estrategia corresponde a lo que hemos llamado estrategia determinada (*single-minded*).

7.3. Material y métodos

Como hemos mostrado previamente, los agentes BDI—expresados en términos $AgentSpeak(L)$—no son agentes ciegos, es decir, no siguen una estrategia de compromiso ciego, sino que se comportan de acuerdo a una estrategia de compromiso determinado; sin embargo, tal estrategia determinada es limitada. El aprendizaje intencional provee un enfoque alternativo para alcanzar un compromiso de tipo determinado completo, esto es, una forma de compromiso *single-minded* sin la limitación que los agentes BDI tienen, a saber, la carencia de elementos para representar explícitamente razones de abandono [79].

Los agentes con aprendizaje intencional usan una programación lógica induc-

tiva con el fin de aprender razones para adoptar planes como intenciones. Las razones se codifican mediante un contexto que, como decíamos renglones arriba, es apremiante porque es el núcleo de este tipo de aprendizaje.

Los ejemplos de entrenamiento que usan los agentes están representados, para cierto plan dado, de tal manera que sea más sencillo para ellos registrar lo que creen cuando el plan es adoptado como una intención y la etiqueta los ejemplos con la salida de la ejecución de la intención son *éxito* o *fallo*. Cuando cierta intención falla una inducción basada en árboles de decisión es ejecutada para aprender un nuevo contexto en función de las ramas del árbol de decisión que llevan al éxito.

La idea del aprendizaje intencional es que es posible aprender, de modo similar, las razones de abandono de una intención: la política de la reconsideración en función de las ramas del árbol de decisión que llevan al fracaso o fallo. Así, en lugar de modificar el contexto del plan que está en revisión, los agentes determinados añaden reglas (lo cual es posible en $Jason$) a las creencias del agente.[1]

Así, podemos tener creencias de la forma:

$$abandon(I) : -at_1; \ldots; at_n$$

donde I es el átomo a ser revisado y la secuencia $at_1; \ldots; at_n$ es una rama del árbol de decisión que lleva a un fallo.

Todo agente determinado tiene un plan en su librería para abandonar intenciones:

```
+abandon(I) : not I <- -intending(I); .drop_intention(I).
```

El contexto de un plan de este tipo es $not\ I$ para evitar el abandono de planes cuando otros agentes satisfacen I, cuando el ambiente lo satisface o cuando el propio agente lo satisface. La creencia $intending(I)$ es verdadera mientras el agente tiene I en el conjunto de intenciones. Además, cuando un agente verifica la secuencia $at_1; \ldots; at_n$, un evento de la forma $+abandon(I)$ es generado y el plan se convierte en un plan relevante y posiblemente aplicable.

Para integrar todos estos materiales y métodos un experimento ha sido diseñado para observar el comportamiento del compromiso determinado como un caso de revisión basada-en-políticas. El experimento ha sido implementado en

[1] Una semántica operacional extendida de $AgentSpeak(L)$ que se las ve con el aprendizaje intencional, por medio de métodos inductivos e incrementales, ha sido propuesto por Guerra *et al* [79] con excelentes resultados.

$Jason$ [19] y comienza situando cuatro agentes en un ambiente definido como un mundo de bloques:

```
MAS policy-based-rec {
    infrastructure: Centralised
       environment: Blocks
    agents:
       bold;
       learner agentClass Learner;
       single-minded agentClass SingleMinded;
       experimenter; }
```

El agente *experimenter* propone a los otros agentes (*bold*, *learner* y *single-minded*) la tarea de colocar un bloque b sobre un bloque c y, con cierta probabilidad, introduce ruido en el experimento al colocar un bloque z sobre c antes de su pedido (esto para representar la dinamicidad del ambiente). Después recolecta la información sobre la eficiencia de los otros agentes para cierto número de iteraciones.

Todos los demás agentes son, inicialmente, audaces (*bold*) con respecto a sus intenciones en el sentido de que los contextos de sus planes son vacíos ($\top$). Esto es, todos ellos comparten el plan:

```
+!on(X,Y) <- put(X,Y).
```

El agente audaz (*bold*) no puede aprender intencionalmente: es un agente por *default*. El agente con aprendizaje intencional (*learner*) es capaz de aprender el contexto de sus planes. El agente determinado (*single-minded*) puede aprender, además, tanto el contexto de los planes como razones de abandono, por lo que el agente con aprendizaje y el agente determinado pertenecen a una diferente clase de agentes debido a que usan acciones primitivas para lograr sus acciones de aprendizaje y generan eventos especializados para usar su conocimiento generado por aprendizaje.

Tenemos, por tanto, dos posibles resultados en el experimento: la intención adoptada para colocar b sobre c falla o tiene éxito. Cuando ocurre un fallo, el agente con aprendizaje intencional modifica el contexto de su plan del siguiente modo:

```
+!on(X,Y) : clear(Y) <- put(X,Y).
```

Mientras que el agente determinado genera la regla:

```
abandon(on(X,Y)) :-
intending(on(X,Y)) & not clear(Y).
```

7.4. Resultados

Dentro del ambiente el ruido puede aparecer antes, después o durante la adopción de la intención de cada agente. Esto depende de la organización de $Java$ con respecto a los procesos de $Jason$. Aún cuando es posible controlar el momento donde el ruido aparece, es importante notar que los agentes en condiciones normales carecen de dicho control. Algunas veces el ruido aparece después de la ejecución de la intención, de tal suerte que el experimentador falla (esa es la razón por la cual el número de iteraciones no siempre es igual a la suma de fallos y éxitos en el Cuadro 7.1).

Ags	Iters	Ruido	Fallo				Éxito			
			antes	después	acción	total	no plan	abandon	éxito	total
B	1,000	90 %	735	160	0	895	0	0	101	101
L	1,000	90 %	0	69	0	69	724	0	188	912
S	1,000	90 %	0	1	5	6	242	381	356	982
B	1,000	50 %	417	69	0	486	0	0	502	502
L	1,000	50 %	0	54	0	54	355	0	576	931
S	1,000	50 %	0	1	3	4	116	158	715	989
B	1,000	30 %	249	52	0	301	0	0	624	624
L	1,000	30 %	0	26	0	26	131	0	837	968
S	1,000	30 %	0	0	4	4	72	76	838	986
B	1,000	10 %	90	23	0	113	0	0	797	797
L	1,000	10 %	0	19	0	19	26	0	951	977
S	1,000	10 %	0	0	2	2	24	21	950	995
B	100	90 %	67	23	0	90	0	0	10	10
L	100	90 %	0	5	0	5	47	0	48	95
S	100	90 %	0	2	0	2	12	25	60	97
B	100	50 %	34	11	0	45	0	0	47	47
L	100	50 %	0	10	0	10	41	0	49	90
S	100	50 %	0	2	0	2	20	18	59	97
B	100	30 %	19	10	0	29	0	0	64	64
L	100	30 %	0	4	0	4	21	0	75	96
S	100	30 %	0	1	0	1	9	8	82	99
B	100	10 %	4	2	0	6	0	0	81	81
L	100	10 %	0	0	0	0	0	0	100	100
S	100	10 %	0	0	0	0	0	0	98	98

Cuadro 7.1: Resultados experimentales

El Cuadro 7.1 muestra el resultado de dos corridas de diferente tamaño: una con 1000 iteraciones y otra con 100. La dinámica del ambiente depende de

la probabilidad de ocurrencia de ruido (90 %, 50 % y 30 %). Los valores bajos inducen ambientes estáticos, mientras los valores altos determinan ambientes dinámicos. La intención adoptada para colocar b sobre c o bien falla o bien tiene éxito. Los resultados relevantes son los siguientes:

- El agente audaz (*bold* o *B*) siempre falla cuando hay ruido, ya que no puede aprender nada acerca de la adopción o abandono de las intenciones.
- El agente con aprendizaje intencional (*learner* o *L*), por otro lado, reduce el número de fallos debidos al ruido, puesto que aprende cierto contexto: en este caso, poner X sobre Y requiere que Y esté libre. Y cuando esto no es el caso, el plan deja de ser aplicable.
- El agente determinado (*single-minded* o *S*) reduce drásticamente el número de fallos al prevenir la adopción inconveniente de ciertos planes al abandonar intenciones cuando es necesario: cuando el agente adopta cierta intención pero el agente experimentador coloca un bloque z sobre c. En efecto, el agente determinado sólo falla cuando está listo para ejecutar su acción *put* y el ruido aparece.

Para los casos exitosos hemos considerado la conducta previsora del agente como un éxito. Así, si el agente se rehusa a adoptar cierto plan, se considera como un caso exitoso. Usualmente se espera que el agente tenga diferentes planes para cierto evento. El rechazo de un plan resultaría en la generación de un plan diferente a ser adoptado para solventar el problema.

Abandonar un plan también se considera como un caso de éxito: significa que el agente usó su reconsideración basada-en-políticas para prevenir un fallo real. En la práctica esto resulta en la eliminación del evento asociado a la intención perteneciente a los eventos del agente.

Claramente, tanto el agente audaz como el agente con aprendizaje intencional no pueden abandonar intenciones. Por tanto, como se esperaba, el agente determinado es más exitoso que el agente con aprendizaje intencional cuando hay tazas altas de ruido; mientras que estos dos agentes son más exitosos que un agente audaz.[2]

[2]Este es un resultado muy interesante desde el punto de vista técnico porque el grado de audacia o precaución de un agente es algo difícil de definir *a priori*. Experimentalmente se sabe que en ambientes dinámicos un agente cauto se comporta mejor que un agente audaz; inversamente, en ambientes estáticos los agentes audaces se comportan mejor que los cautos [97]: el problema, sin embargo, consiste en cómo definir estos grados de manera automática y no-supervisada.

7.5. Resumen

Con estos resultados se pueden ver, literalmente, las bondades del aprendizaje intencional en el marco de toda esta discusión. Los resultados experimentales son prometedores: el aprendizaje permite una aproximación a la revisión. Bastaría, pues, con mostrar cómo podemos integrar estos resultados de manera orgánica en el marco de la revisión y la no-monotonicidad.

Por lo pronto podemos asegurar que este tipo de aprendizaje puede interpretarse en términos bratmanianos de reconsideración: al obtener dos tipos de ramas en el árbol inductivo de decisión [79] obtenemos ramas con éxito, que representan los procesos de preservación de razones (*reason preserving*); y ramas con fallo, que representan los procesos de modificación de razones (*reason changing*). Además, este resultado nos permite ver que lo expresado por la Proposición 4 se corrobora en la práctica.

En resumen: sabemos que desde un punto de vista cognitivo un agente BDI es un sistema agente que se explica en términos del modelo BDI. Un modelo BDI bratmaniano de razonamiento intencional debe considerar los atributos lógicos de revisión de intenciones y de no-monotonicidad del razonamiento intencional: no hacerlo es construir un modelo no-bratmaniano que no resuelve los problemas externo e interno. El problema que quedaba abierto era el de cómo traducir el análisis del modelo BDI bratmaniano a un modelo formal que resolviera los problemas que los modelos tradicionales no terminaban de resolver.

Con el desarrollo y uso de $BDI^{CTL}_{AgentSpeak(L)}$ evitamos el problema técnico y observamos que la revisión de intenciones y la no-monotonicidad intencional eran atributos lógicos legítimos de los agentes BDI modelados vía $AgentSpeak(L)$. Y aunque nos separamos momentáneamente del problema interno, lo que hemos hecho ahora es mostrar que además de que nuestro análisis es compatible con los agentes BDI interna y externamente desde un punto de vista formal, también es correcto experimentalmente. Sin embargo, aunque este resultado es relevante, su carácter experimental aún se queda corto para nuestros fines.

Capítulo 8

Hacia la revisión de intenciones

8.1. Introducción

En el capítulo precedente hemos visto que el aprendizaje intencional permite acercarnos a una forma de revisión de intenciones. En este capítulo pretendemos encarar de manera teórica la conjetura que inferimos previamente a partir de los resultados experimentales.

Lo que sabemos al momento es que la revisión de intenciones está justificada en un modelo BDI bratmaniano, en principio, en términos declarativos pero no procedimentales. Para acercar los términos declarativos a los procedimentales veremos que el modelo general de revisión es lo suficientemente general como para ser compatible con el fragmento I del modelo BDI bratmaniano.

El capítulo está distribuido en la siguiente manera: en la primera Sección describimos lo que queremos decir cuando usamos la expresión *revisión de intenciones* y mostramos algunos problemas metodológicos. En § 8.3 discutimos algunas cuestiones relacionadas con la representación y en § 8.5 adaptamos y sugerimos algunos postulados generales para la revisión de intenciones.

8.2. Hacia la revisión de intenciones

Desde el Capítulo 3 hemos insistido en que las nociones de revisión, expansión y contracción no son exclusivas de las creencias. Y cuando hablábamos de la perspectiva externa decíamos que a los siete postulados de $C\&L$ podíamos agregar dos más, de tal suerte que la lista luciría así:

1. Las intenciones ponen problemas para los agentes, quienes necesitan determinar maneras de alcazarlas.
2. Las intenciones proveen un filtro para adoptar otras intenciones que no deben entrar en conflicto.
3. Los agentes rastrean el éxito de sus intenciones y están inclinados a tratar de nuevo si sus intentos fallan.
4. Los agentes creen que sus intenciones son posibles.
5. Los agentes no creen que no lograrán sus intenciones.
6. Bajo ciertas circunstancias los agentes creen que alcanzarán cumplir sus intenciones.
7. Los agentes no necesitan intentar todos los efectos colaterales de sus intenciones.
8. *Los agentes pueden expandir sus intenciones*.
9. *Los agentes pueden contraer sus intenciones*.

Y decíamos que, de manera general, los postulados 8 y 9 garantizaban la posibilidad de la revisión de intenciones. Sin embargo, hasta ese momento hablábamos en un sentido declarativo, de alto nivel, y no nos preocupábamos por los detalles procedimentales.

Para acercarnos a nuestra meta veamos ahora, a manera de ejemplo, cómo luce el proceso de revisión de intenciones. Asumamos un agente inmerso en un ambiente que es innacesible, no-determinista, episódico, discreto y dinámico. Admitamos, además, que tal agente tiene ciertas creencias e intenciones (un estado α) y que, en cierto momento, forma la intención de modificar el estado del mundo (estado β)—representamos esta situación con la flecha sólida en negro (Figura 8.1).

De esta manera el agente genera una intención de la forma $coloca(B, C)$. Ahora, dadas las propiedades del ambiente, supongamos que el agente percibe la información denotada por γ—señalada por la flecha roja—donde no es el caso que $libre(C)$ (tal como ocurría en el experimento en 7.4). En dicho caso la intención fallará y el conjunto de intenciones se hará inconsistente y, en última instancia, el agente no logrará sus metas (o intentará todas las metas, dada la inconsistencia).

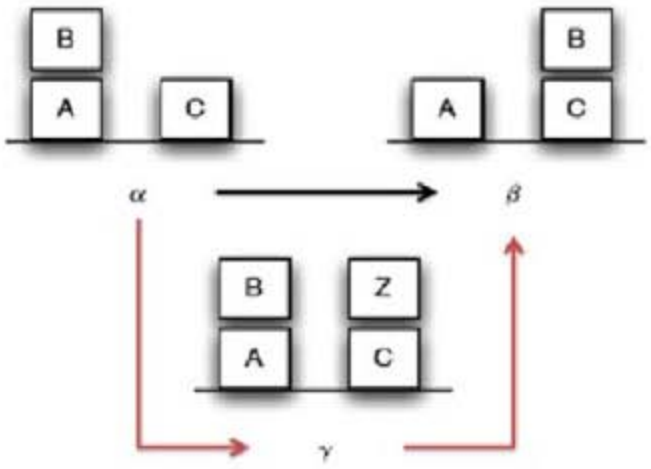


Figura 8.1: Estados del ambiente y del agente

Siguiendo con el ejemplo, si pudiéramos apreciar esta situación en una base de datos, podríamos ver algo como lo siguiente:

p_1: $!coloca(x, y)$.

p_2: $+!coloca(x, y) : -libre(x)$.

p_3: $+!coloca(x, y) : -libre(y)$.

p_4: $+!coloca(x, y) : -!mueve(x)$.

donde $!\phi$ representa una intención y $+\phi$ la adición de dicha intención. Si este conjunto de cláusulas viene equipado con algún motor inferencial, la siguiente fórmula se requiere para alcanzar la intención:

p_5: $libre(x)$.

Ahora, supongamos que es el caso que x no está libre. Esto significa que tenemos que añadir la negación de p_5 a la base de datos. Pero entonces el conjunto de intenciones se vuelve inconsistente

Si queremos mantener la consistencia de las intenciones, lo cual no sólo es un requirimiento de nuestra teoría sino también una condición razonable, es preciso revisar la base de datos. Esto implica que algunas intenciones, seleccionadas de alguna manera particular, tienen que ser retractadas.

El problema de la revisión de intenciones es, entonces, doble: en primer lugar, porque está íntimamente relacionada a otros estados intencionales (como las creencias); y, en segundo lugar, porque la representación lógica por sí sola no es suficiente para determinar qué intenciones deberían ser retractadas.

8.2.1. Detalles metodológicos

Los problemas son, entonces, ¿cómo podemos hacer esos cambios de actitudes intencionales? ¿Y qué significan en el contexto BDI? Para responder a estas preguntas hay que hacer referencia a algunas cuestiones metodológicas relacionadas con la representación, la noción de inferencia y las funciones de selección.

- Cuando nos preguntamos cómo es que hay que representar a las intenciones nos enfrentamos con el *problema de la representación*.

Por ejemplo, la mayoría de las bases de datos trabajan con hechos y reglas de algún tipo. El lenguaje usado para representar intenciones—y creencias—puede estar relacionado con un formalismo lógico. El problema es, entonces, determinar qué lenguaje debemos usar para representar nuestros datos.

- Cuando consideramos cuáles son las mutuas relaciones entre los elementos del modelo podemos ver que hay una relación especial que llamamos el *problema de la consecuencia*.

Este problema puede ser pensado en términos de los elementos que pueden ser inferidos y los elementos que son de base. En algunos casos, por ejemplo, los inferidos tienen un estatus diferente de los elementos básicos.

- Y finalmente, cuando nos preguntamos cómo deberíamos escoger los elementos a retractar nos adentramos en el *problema de la función de selección*.

La lógica por sí sola es insuficiente para decidir qué intenciones o creencias deben mantenerse y cuáles deben abandonarse, por lo que es necesaria alguna heurística para definir esta función.

8.3. Modelos para representar estados intencionales

Respecto al problema de la representación por el momento usaremos un modelo proposicional. Asumiremos, además, que el lenguaje está cerrado por los operadores $\neg$, $\wedge$, $\vee$, $\Rightarrow$ evaluados de manera booleana. Usamos también $\phi, \psi, ...$ como variables proposicionales. La idea es que el lenguaje así definido no sólo acepta lo que está explícitamente expresado en la base de datos sino también sus consecuencias. Por esto, otro factor que hay que considerar es el del sistema lógico asociado. Sin embargo, al hacer este análisis, en principio teorético, comenzaremos haciendo una declaración general de la funciones de revisión.

8.3.1. Conjuntos de estados y bases intencionales

La manera más sencilla para representar estados intencionales es usando fórmulas bien formadas (fbf) de algún lenguaje. De acuerdo a esto podemos definir un conjunto de estados intencionales[1] (conjunto intencional, de ahora en adelante) como un conjunto Σ de fbfs que satisfacen el axioma de reflexividad generalizada: si $\Sigma \vdash \phi$ entonces $\phi \in \Sigma$. Esta condición asegura que Σ está cerrada bajo una relación de consecuencia lógica. Sin embargo, por las propiedades de la lógica estándar, siempre que Σ sea inconsistente, entonces para toda ϕ, $\Sigma \vdash \phi$. Denotamos esto como $\Sigma_\perp$ y decimos que el conjunto intencional es inconsistente.

Sin embargo, puede ser el caso que consideremos que algunas intenciones no son básicas sino inferidas. Y no es posible expresar estas distinciones a través de conjuntos intencionales, puesto que la representación conjuntista no provee de manera natural con marcadores o banderas explícitas para indicar qué intenciones son básicas y cuáles no lo son. Por este motivo haremos uso de las bases intencionales cuando sea pertinente.

[1] Hay una cercanía entre los conjuntos intencionales y lo modelos de mundos posibles puesto que, para cualquier conjunto W_Σ de mundos posibles podemos definir un conjunto intencional correspondiente Σ como el conjunto de las proposiciones que son verdaderas en todos los mundos en W_Σ. Desde un punto de vista computacional, sin embargo, los estados intencionales son menos complejos.

8.4. El problema de la consecuencia

Para caracterizar la revisión de intenciones notamos que hay dos estrategias a seguir: una, que consiste en presentar de manera explícita la construcción de los procesos de revisión; la otra, que formula las ideas generales para llevar a cabo tales construcciones. Así, la primera solución consiste en desarrollar algoritmos para computar las funciones; la segunda, en describir los postulados para definir tales funciones. En esta sección seguiremos la primera estrategia; en la siguiente sección, la segunda.

Para entender el problema de la consecuencia hagamos uso de las relaciones de consecuencia de las creencias. Siguiendo a Shoham [146], en el fragmento B es posible caracterizar estos problemas usando los siguientes procedimientos algorítmicos. Los Algoritmos 6 y 7 muestran las funciones que implementan una expansión y una contracción de creencias, respectivamente:

Algoritmo 6: Expansión de creencias

Datos : B, b
$B \leftarrow B \cup b$
mientras $b \in precondiciones(B)$ **hacer**
  | $B \leftarrow consistencia(B, b)$
fin
return B

Algoritmo 7: Contracción de creencias

Datos : B, b
$B \leftarrow B - b$
mientras $b \in poscondiciones(B)$ **hacer**
  | $B \leftarrow consistencia(B, b)$
fin
return B

El Algoritmo 6 indica que cuando se expanden las creencias, no sólo se añade la nueva creencia b a la base de creencias B, sino que además se tiene que resguardar la consistencia para todas las creencias cuyas precondiciones incluyan a b.

Por su parte, el Algoritmo 7 indica que cuando se contrae una creencia, b se remueve de la base B y se mantiene la consistencia atendiendo a las creencias de B cuyas poscondiciones incluyan b. De manera análoga, podemos entender los mecanismos de expansión o contracción de intenciones [146] (Algoritmos 8 y 9):

Algoritmo 8: Expansión de intenciones

Datos : Σ, B, ϕ
$\Sigma \leftarrow \Sigma \cup \phi$
mientras $\phi \in precondiciones(B)$ **hacer**
  para $\phi \in B$ **hacer**
    $\Sigma \leftarrow \Sigma \cup poscondiciones(\phi)$
    $\Sigma \leftarrow consistencia(\Sigma, \phi)$
  fin
fin
return Σ

Algoritmo 9: Contracción de intenciones

Datos : Σ, ϕ
$\Sigma \leftarrow \Sigma - \phi$
para $\phi \in poscondiciones(\Sigma)$ **hacer**
  $poscondiciones(\Sigma) \leftarrow poscondiciones(\Sigma) - \phi$
fin
return Σ

El Algoritmo 8 indica que cuando expandemos intenciones no sólo añadimos la intención ϕ al conjunto de intenciones, sino que añadimos la poscondición de ϕ al conjunto de intenciones y mantenemos consistencia con las creencias mientras haya creencias cuyas precondiciones incluyan ϕ.

El Algoritmo 9 nos dice que cuando se contrae una intención ϕ, ésta se remueve del conjunto de intenciones y se remueven todas las intenciones anotadas por ϕ. En el siguiente capítulo exploraremos más detalles de estos algoritmos.

8.5. Postulados

Ahora presentaremos los postulados clásicos de la revisión de creencias y observaremos cómo la revisión de intenciones ocurre de manera similar cuando una nueva pieza de información inconsistente con la base de información es añadida a un sistema intencional.

Podemos distinguir, por lo menos, los tres cambios tradicionales dadas las propiedades que los fragmentos B e I comparten, a saber, retractabilidad e inercia:

- La *expansión* (o *adopción* en términos bratmanianos) que ocurre cuando una fórmula ϕ es añadida a Σ junto con sus consecuencias: $\Sigma \oplus \phi$.

- La *revisión* (o *reconsideración* en términos bratmanianos) que ocurre cuando una fórmula ϕ inconsistente con Σ es añadida de tal manera que para mantener consistencia en el sistema resultante algunas fórmulas de Σ tienen que ser abandonadas: $\Sigma \odot \phi$.

- La *contracción* (o *abandono* en términos bratmanianos) que ocurre cuando una fórmula ϕ en Σ es abandonada sin añadir ningún dato nuevo y mantener a Σ cerrado: $\Sigma \ominus \phi$.

La expansión es cerrada bajo consecuencia lógica ($\Sigma \oplus \phi = \{\psi | \Sigma \cup \phi \vdash \psi\}$); sin embargo, no es posible dar una caracterización similar de los otros dos procesos, pues los problemas de la revisión y de la contracción están en la manera extralógica de llevar a cabo estos procesos, como se puede ver en los Algoritmos 8 y 9. Por tanto, podemos tener diferentes maneras de investigarlos, especificarlos y verificarlos.

Por el momento asumiremos que los conjuntos intencionales modelan bases intencionales. La motivación de esta interpretación de la revisión de intenciones en términos de los postulados de AGM es que cuando modificamos el fragmento I de manera racional modificamos las intenciones de manera mínima y preservando consistencia. A continuación exponemos los postulados clásicos.

8.5.1. Revisión

Para la revisión requerimos cerradura:

Postulado 1 *($\odot 1$) Para toda ϕ y cualquier conjunto intencional Σ, $\Sigma \odot \phi$ es un conjunto intencional.*

El segundo postulado garantiza que la fórmula de entrada es aceptada en la revisión:

Postulado 2 *($\odot 2$)* $\phi \in \Sigma \odot \phi$

Un proceso de revisión debería ocurrir cuando la fórmula ϕ contradice algo que ya pertencece a Σ, esto es, $\neg\phi \in \Sigma$. Sin embargo, para generalizar este proceso tenemos que cubrir el caso cuando $\neg\phi \notin \Sigma$, por lo que la revisión coincide con la expansión:

Postulado 3 *($\odot 3$)* $\Sigma \odot \phi \subseteq \Sigma \oplus \phi$

Postulado 4 *($\odot 4$) Si $\neg\phi \notin \Sigma$, entonces $\Sigma \oplus \phi \subseteq \Sigma \odot \phi$*

El objetivo de una revisión es producir un nuevo conjunto intencional consistente. Por tanto, $\Sigma \odot \phi$ debería ser consistente, a menos que ϕ sea lógicamente imposible:

Postulado 5 *($\odot 5$) $\Sigma \odot \phi = K_\perp$ si y sólo si $\vdash \neg\phi$*

Y finalmente, requerimos equivalencia:

Postulado 6 *($\odot 6$) Si $\vdash \phi \Leftrightarrow \psi$, entonces $\Sigma \odot \phi = \Sigma \odot \psi$*

Estos son los postulados básicos de la revisión. A continuación mostramos los postulados básicos de la contracción.

8.5.2. Contracción

También se necesita cerradura:

Postulado 7 *($\ominus 1$) Para toda ϕ y cualquier conjunto intencional Σ, $\Sigma \ominus \phi$ es un conjunto intencional.*

Además, dado que $\Sigma \ominus \phi$ se forma a partir de Σ al abandonar ciertas intenciones, ninguna intención nueva debería aparecer:

Postulado 8 *($\ominus 2$) $\Sigma \ominus \phi \subseteq \Sigma$*

Cuando $\phi \notin \Sigma$, una heurística de optimización requiere que nada tiene que ser retraído:

Postulado 9 *($\ominus 3$) Si $\phi \notin \Sigma$, entonces $\Sigma \ominus \phi = \Sigma$*

La fórmula que va a ser contraída no debe ser una consecuencia de las intenciones en $\Sigma \ominus \phi$:

Postulado 10 *($\ominus 4$) Si $\nvdash \phi$, entonces $\phi \notin \Sigma \ominus \phi$*

De $(\ominus 1)$ a $(\ominus 4)$ se sigue que si $\phi \notin \Sigma$, entonces $(\Sigma \ominus \phi) \oplus \phi \subseteq \Sigma$. En otras palabras, si primero retractamos ϕ y después añadimos ϕ de nuevo, no hay intenciones nuevas que no hubieran estado antes.

La heurística de optimización demanda una forma de recuperación:[2]

[2]Este postulado nos permite deshacer contracciones, y aunque es controversial, lo asumiremos como parte de nuestro modelo general.

Postulado 11 *($\ominus 5$) Si $\phi \in \Sigma$, entonces $\Sigma \subseteq (\Sigma \ominus \phi) \oplus \phi$*

El sexto postulado es análogo a $(\Sigma \odot 6)$:

Postulado 12 *($\ominus 6$) Si $\vdash \phi \Leftrightarrow \psi$, entonces $\Sigma \ominus \phi = \Sigma \ominus \psi$*

Estos son los postulados básicos de la contracción. Y en consecuencia, al considerar los postulados de manera general, podemos hacer las siguientes observaciones compatibles y relevantes para un modelo bratmaniano:

Observación 1 *Se cumple:*

1. *Reconsideración bratmaniana: si ϕ es revisada, entonces ϕ es abandonada o mantenida:*
$$\Sigma \odot \phi \Rightarrow (\Sigma \ominus \phi) \vee ((\Sigma \ominus \phi) \oplus \phi)$$

2. *Consistencia interna: la inconsistencia de una revisión resulta de la inconsistencia de intenciones:*
$$\Sigma_{\perp} \Rightarrow \Sigma \odot \phi = \Sigma_{\perp}$$

3. *Reconsideración ocasional: revisar un Σ consistente con las intenciones actuales no remueve intenciones:*
$$(\Sigma \odot \phi) \oplus \phi = \Sigma$$

Esto nos da evidencias para concluir que el modelo AGM de revisión de creencias no sólo es un modelo para creencias y es, en cambio, un modelo más general para representar y entender otras estructuras de datos: en particular, AGM constituye un modelo para la revisión de intenciones en un modelo bratmaniano.

8.6. Resumen

En este capítulo empezamos a enfrentar de manera teórica la conjetura que inferimos de los resultados experimentales del capítulo previo. La revisión de intenciones, sabemos, está justificada en un modelo BDI bratmaniano en términos declarativos pero no procedimentales.

Lo que buscábamos era interpretar el fragmento I para observar si era compatible con el modelo general de revisión para mostrar que la metodología de

revisión era lo suficientemente general para ser compatible con el fragmento I del modelo BDI bratmaniano.

Lo que hemos bosquejado con el uso del marco AGM de revisión para entender la revisión de intenciones de Bratman puede resumirse de la siguiente manera: el modelo AGM, por ser un modelo general, es compatible con las intenciones; el problema, evidentemente, es que este resultado es muy débil, pues no nos dice nada realmente sobre la revisión de intenciones. Seguimos con el problema: ¿Cómo *ver* la revisión de intenciones en agentes BDI si ya tenemos evidencia de que la revisión de intenciones es un desiderátum de racionalidad y además el modelo BDI bratamaniano es compatible con este marco teórico?

Resumimos: sabemos que desde un punto de vista cognitivo un agente BDI es un sistema agente que se explica en términos del modelo BDI. Un modelo BDI bratmaniano de razonamiento intencional debe considerar los atributos lógicos de revisión de intenciones y de no-monotonicidad del razonamiento intencional: no hacerlo es construir un modelo no-bratmaniano que no resuelve los problemas externo e interno. El problema que quedaba abierto era el de cómo traducir el análisis del modelo BDI bratmaniano a un modelo formal que resolviera los problemas que los modelos tradicionales no terminaban de resolver.

Con el desarrollo y uso de $BDI^{CTL}_{AgentSpeak(L)}$ evitamos el problema técnico y observamos que la revisión de intenciones y la no-monotonicidad intencional eran atributos lógicos legítimos de los agentes BDI modelados vía $AgentSpeak(L)$. Y aunque nos separamos momentáneamente del problema interno, lo que hemos hecho ahora es mostrar que además de que nuestro análisis es compatible con los agentes BDI interna y externamente desde un punto de vista formal, también es correcto experimentalmente. Sin embargo, aunque este resultado es relevante, su carácter experimental aún se queda corto para nuestros fines. Por ello nos embarcamos en la tarea de interpretar la revisión de intenciones mediante el aprendizaje intencional, de manera precisa, explorando cómo algunos procedimientos de aprendizaje pueden usarse como mecanismos para la revisión de intenciones acercando los términos declarativos de la revisión a los términos procedimentales.

Capítulo 9

Revisión de intenciones

9.1. Introducción

Mientras la adaptación del modelo AGM para analizar los cambios intencionales provee una especificación abstracta, compatible y útil, como vimos en el capítulo anterior, no está comprometida o relacionada con ningún mecanismo fijo o con alguna implementación de agentes BDI como la que también mostramos en el mismo capítulo.

Por otro lado, $AgentSpeak(L)$ tiene una semántica operacional bien definida que nos ofrece un marco para analizar los cambios explícitos en los estados del agente y los eventos en el ambiente pero no cuenta con un proceso de revisión de intenciones al estilo AGM.

El marco de revisión de intenciones no analiza explícitamente los eventos que producen cambios intencionales ni los mecanismos que definen tales cambios, a la vez que se enfoca únicamente en las tres operaciones de expansión, contracción y revisión, cuya completud como un repertorio de acciones aún está abierta [16]. Afortunadamente, como vimos en el Capítulo 7, existe un procedimiento de aprendizaje para agentes BDI en $AgentSpeak(L)$ [77, 79] que puede ser usado para fijar un mecanismo particular de revisión de intenciones.

Lo que proponemos ahora es que podemos entender a la revisión de intenciones como un mecanismo de aprendizaje de intenciones: con esto nos enfrentamos directamente al problema externo.

Recordemos que cuando planteábamos este problema nos preguntábamos si habían características en las intenciones que permitieran un mecanismo de cambio racional de intenciones. Partiendo de la lectura del modelo BDI bratmaniano

y del análisis del capítulo anterior podemos inferir que las intenciones poseen propiedades que justifican una noción mínima de revisión. Ahora veremos cómo se comporta esta noción de revisión.

El capítulo está organizado del siguiente modo. En la Seccion 9.2 generalizamos el proceso de aprendizaje intencional y posteriormente (§ 9.3) desplegamos los resultados de esta generalización.

9.2. Aprendizaje BDI

Como hemos intentado mostrar, los agentes BDI interpretados mediante $AgentSpeak(L)$ no siguen una forma de compromiso determinado de manera explícita [80]. Sin embargo, el uso de mecanismos de aprendizaje intencional se convierte en este momento en una manera alternativa de alcanzar tal estrategia usando reglas como la siguiente:

$$\frac{\mathcal{S}_{\mathcal{E}}(C_E) = \langle +abandon(\phi), \top\rangle \wedge ag_{bs} \models intending(\phi)}{\langle ag, C, M, T, SelEv\rangle \rightarrow \langle ag', C', M, T, SelEv\rangle}$$

t.q. $C'_E = C_E \backslash \{\langle +abandon(\phi), \top\rangle\}, ag_{bs} \not\models intending(\phi), C'_I = C_I \backslash \phi$.

Esta regla, llamada $Abandon$, como decíamos previamente, dicta que cuando un agente intenta ϕ y un evento de la forma $+abandon(\phi)$ se genera, el evento es removido de C_E (la lista de eventos), el agente deja de creer $intending(\phi)$, la intención es removida de C_I (el conjunto de intenciones) y un nuevo evento pasa a ser seleccionado.

Podemos reducir la regla anterior a una función de intenciones sin perder las propiedades globales acerca de ellas:

Definición 43 (Abandon) *La regla $abandon$ es una función t.q.*

$$abandon(\phi, C_I) \;=\; \begin{cases} \{C_I - \phi\} & \textit{si } \phi \in C_I, \\ \{C_I\} & \textit{si } \phi \notin C_I | \phi = \top \end{cases}$$

Conversamente, podemos abstraer y definir una función de adopción que se comporte de la siguiente manera:

Definición 44 (Learn) *La regla $learn$ es una función t.q.*

$$learn(\phi, C_I) \;=\; \begin{cases} \{C_I\} & \textit{si } \phi \in C_I | \phi = \top, \\ \{C_I \cup \phi\} & \textit{si } \phi \notin C_I \end{cases}$$

Notemos que en esta representación el componente C_I de la circunstancia agente denota las intenciones del agente dentro de un ciclo de razonamiento. Esta suposición tiene la ventaja de permitirnos trabajar dentro de cada ciclo de razonamiento. Además, es importante notar que la aplicación de *abandon* deja $C'_I \subseteq C_I$, lo cual es trivialmente cierto; y además $\phi \notin Cn(C'_I)$ dado que si $\phi \in Cn(C'_I)$ entonces ϕ fallará y la regla $ClrInt$ sacará ϕ de C'_I [19].

9.3. Revisión mediante aprendizaje

Usando estas funciones podemos visualizar los siguientes resultados que relacionan los mecanismos de aprendizaje intencional con la especificación abstracta de la revisión de intenciones.

En primer lugar, notamos que el abandono de intenciones se comporta como la contracción (*abandono*, en términos bratmanianos):

Proposición 5

$$abandon(\phi, C_I) \Rightarrow C_I \ominus \phi$$

Mientras que el aprendizaje de intenciones se comporta como la expansión (*adopción*, en términos bratmanianos):

Proposición 6

$$learn(\phi, C_I) \Rightarrow C_I \oplus \phi$$

Además, resulta interesante notar que *learn* y *abandon* son duales:

Lema 1

$$learn(\phi, abandon(\phi^c, C_I)) \Leftrightarrow abandon(\phi^c, learn(\phi, C_I))$$

Observación 2 *Usando el lema anterior, las siguientes observaciones son directas:*
Si $\phi \in C_I$ *ó* $\phi \notin C_I$ *entonces* $learn(\phi, abandon(\phi^c, C_I)) \subseteq learn(\phi, C_I)$
Si $\phi^c \notin C_I$ *y* $\phi \in C_I$ *entonces* $learn(\phi, abandon(\phi^c, C_I)) = C_I$
Si $\phi^c \notin C_I$ *y* $\phi \notin C_I$ *entonces* $learn(\phi, abandon(\phi^c, C_I)) = C_I \cup \phi$

Estos resultados nos permiten representar la composición de *learn* y *abandon* como un proceso de revisión (*reconsideración* en términos bratmanianos):

Proposición 7

$$learn(\phi, C_I) \circ abandon(\phi^c, C_I) \Rightarrow C_I \odot \phi$$

La idea de modificar C_I representa la suposición de que los agentes cambian de intenciones durante ciclos de razonamiento, pero este resultado se puede extender a los planes ps de la librería de un agente BDI de tal manera que:

Definición 45 ($Abandon_p$) *La regla $abandon_p$ es una función t.q.*

$$abandon_p(\phi, ps) \quad = \quad \begin{cases} \{ps - \phi\} & \text{si } \phi \in ps, \\ \{ps\} & \text{si } \phi \notin ps | \phi = \top \end{cases}$$

Conversamente:

Definición 46 ($Learn_p$) *La regla $learn_p$ es una función t.q.*

$$learn_p(\phi, ps) \quad = \quad \begin{cases} \{ps\} & \text{si } \phi \in ps | \phi = \top, \\ \{ps \cup \phi\} & \text{si } \phi \notin ps \end{cases}$$

Por tanto, por hipótesis, esta generalización debería aplicarse a ps:

$$abandon_p(\phi, ps) \Rightarrow ps \ominus \phi$$
$$learn_p(\phi, ps) \Rightarrow ps \oplus \phi$$
$$learn_p(\phi, ps) \circ abandon_p(\phi^c, ps) \Rightarrow ps \odot \phi$$

9.3.1. Traducción

Con estos resultados en mente podemos traducir la especificación abstracta en términos de agentes BDI a través del formalismo $AgentSpeak(L)$.

Cerradura para la revisión:

Traducción 1 *($\odot 1$) Dado un agente $ag = \langle bs, ps \rangle$, y una intención $\phi \in ag_{ps}$, $ag_{ps} \odot \phi \subseteq ag_{ps}$.*

Aceptación para la revisión:

Traducción 2 *($\odot 2$) $\phi \in ag_{ps} \odot \phi$, dado que ag_{ps} es un conjunto intencional.*

La revisión de ϕ implica la expansión del conjunto de intenciones:

Traducción 3 *($\odot 3$)* $ag_{ps} \odot \phi \subseteq (ag_{ps} \oplus \phi)$.

Si una intención ya pertenece al conjunto de intenciones, entonces su expansión está incluida en la revisión:

Traducción 4 *($\odot 4$) Si* $\neg\phi \notin ag_{ps}$, *entonces* $ag_{ps} \oplus \phi \subseteq ag_{ps} \odot \phi$.

El siguiente postulado es muy importante porque define una propiedad bratmaniana. *Prima facie*, define lo que entendemos como consistencia intencional. En términos de este formalismo, si una intención ϕ no se puede seguir de ag_{ps}, entonces $\phi \in ag_{ps}$ haría a ag_{ps} inconsistente:

Traducción 5 *($\odot 5$)* $ag_{ps} \odot \phi = ag_{ps\perp}$ *si* $\vdash \neg\phi$.

Las intenciones tienen que tratarse de la misma manera al hacer revisiones:

Traducción 6 *($\odot 6$) Si* $\phi \Leftrightarrow \psi$ *entonces* $\phi \odot ag_{ps} = ag_{ps} \odot \psi$.

Cerradura para la contracción:

Traducción 7 *($\ominus 1$) Si* ag_{ps} *es un conjunto intencional, la contracción de* ag_{ps} *es cerrada.*

La siguiente inclusión indica que el conjunto intencional contraído es un subconjunto del original:

Traducción 8 *($\ominus 2$)* $ag_{ps} \ominus \phi \subseteq ag_{ps}$.

Si una intención que va ser contraída no está en el conjunto, la contracción es vacua:

Traducción 9 *($\ominus 3$) Si* $\phi \notin ag_{ps}$, *entonces* $ag_{ps} - \phi = ag_{ps}$.

Traducción 10 *($\ominus 4$) Si* ϕ *no se puede derivar de* ag_{ps}, *entonces* ϕ *no puede ser contraída a partir de* ag_{ps}, *puesto que* $\phi \notin ag_{ps}$.

Traducción 11 *($\ominus 5$) Si* $\phi \in ag_{ps}$, *entonces* $ag_{ps} \subseteq (ag_{bs} \ominus \phi) \oplus \phi$.

Las intenciones deben tratarse igualmente al hacer contracciones.

Traducción 12 *($\ominus 6$) Si* $\phi \Leftrightarrow \phi$, *entonces* $ag_{ps} \ominus \phi = ag_{ps} \ominus \psi$.

9.4. Resumen

Con estos resultados podemos entender, dentro de una implementación particular, el sentido de la especificación abstracta. No es difícil ver que estos procedimientos de aprendizaje intencional proveen un mecanismo para la revisión de intenciones. La importancia de este resultado radica en la preservación de las propiedades de la especificación abstracta en la implementación.

Partiendo de la lectura del modelo BDI bratmaniano y del análisis del capítulo anterior podemos concluir ahora que las intenciones poseen propiedades que justifican una noción mínima de revisión y que ésta se puede modelar interpretándola como un proceso de aprendizaje intencional.

Lo que tenemos, por tanto, es que desde un punto de vista cognitivo un agente BDI es un sistema agente que se explica en términos del modelo BDI. Un modelo BDI bratmaniano de razonamiento intencional debe considerar los atributos lógicos de revisión de intenciones y de no-monotonicidad del razonamiento intencional: no hacerlo es construir un modelo no-bratmaniano que no resuelve los problemas externo e interno. El problema que quedaba abierto era el de cómo traducir el análisis del modelo BDI bratmaniano a un modelo formal que resolviera los problemas que los modelos tradicionales no terminaban de resolver.

Con el desarrollo y uso de $BDI^{CTL}_{AgentSpeak(L)}$ evitamos el problema técnico y observamos que la revisión de intenciones y la no-monotonicidad intencional eran atributos lógicos legítimos de los agentes BDI modelados vía $AgentSpeak(L)$. Y aunque nos separamos momentáneamente del problema interno, lo que hemos hecho ahora es mostrar que además de que nuestro análisis es compatible con los agentes BDI interna y externamente desde un punto de vista formal, también es correcto experimentalmente. Sin embargo, aunque este resultado es relevante, su carácter experimental aún se queda corto para nuestros fines. Por ello nos embarcamos en la tarea de interpretar la revisión de intenciones mediante el aprendizaje intencional, de manera precisa, explorando cómo algunos procedimientos de aprendizaje pueden usarse como mecanismos para la revisión de intenciones acercando los términos declarativos de la revisión a los términos procedimentales: lo que hemos hecho es responder el *problema externo* a través de la construcción de un modelo bratmaniano de revisión.

Capítulo 10

Hacia la no-monotonicidad

10.1. Introducción

Hasta el capítulo anterior desarrollamos un aparato analítico y crítico para responder al problema externo. A partir de aquí volvemos nuestra atención al problema interno relacionado con la noción de no-monotonicidad. El objetivo de este capítulo consiste en acercarnos al desarrollo de un marco no-monotónico para representar intenciones desde un enfoque material.

Sabemos que la revisión de intenciones es posible dado que el análisis filosófico muestra que las intenciones poseen características como proactividad, inercia y admisibilidad. Pero estas propiedades garantizan que las intenciones necesitan, respectivamente, mecanismos de compromiso, derrotabilidad y consistencia que, a su vez, permiten la investigación sobre intenciones en términos de no-monotonicidad.

El problema es que los sistemas formales tradicionales [41, 136, 165] para estudiar el razonamiento intencional están basados en un modelo bratmaniano pero no consideran los aspectos temporales y no-monotónicos de dicho modelo. De manera más específica estos sistemas son intencional-temporales pero monotónicos; complementariamente, las lógicas derrotables consideran el aspecto no-monotónico pero no el intencional-temporal.

Basados en la lectura del Capítulo 3, el objetivo de éste consiste en acercarnos a una formalización de dicho estudio para desarrollar un marco no-monotónico que nos permita representar intenciones pues, como trataremos de mostrar, la inferencia intencional es tan lógica como cualquier otro tipo de inferencia.

El capítulo está organizado del siguiente modo. En las Secciones 10.2 y 10.3

revisamos parte del análisis filosófico de las intenciones y el problema de la monotonicidad. En la Sección 10.4 desplegamos una representación de un modelo bratmaniano en términos de un marco no-monotónico.

10.2. Omnisciencia lógica e intenciones

Comencemos esta sección con un ejemplo:

Ejemplo 8 *Consideremos la expresión:*

$$coloca(X, Y) \leftarrow cerca(X)$$

Esta fórmula nos indica que colocar X sobre Y típicamente implica que el agente se tiene que acercar a X. Si imaginamos que hay un agente moviéndose en el ambiente y adquiere la meta de colocar X sobre Y, entonces, típicamente el agente se acercará a X; sin embargo, esta intención es derrotable en el siguiente sentido. El agente hará lo que tenga que hacer a menos que haya evidencia—esto es, un contexto—que sugiera que tal cosa no es posible. Por tanto, podemos decir que el agente intenta su ejecución típicamente, pero no de manera absoluta e indiscutible.

Podemos notar que para un plan como el anterior resulta razonable asumir un requisito de derrotabilidad que difiere de las interpretaciones tradicionales de la agencia BDI que modelan la conducta intencional bajo la batuta de un sistema modal normal y, por ende, de un sistema monotónico.

Sin duda los sistemas monotónicos son útiles en sus contextos, pero fuera de ellos dan lugar a problemas que ya están bien estudiados o documentados, como el problema de la *omnisciencia lógica*, que consiste en dar licencia a los agentes para inferir todas las consecuencias de su base de creencias o de conocimiento. Así, solemos decir que un agente es lógicamente omnisciente si cree (o sabe) todas las consecuencias de su base de de creencias (o conocimiento).

De manera análoga, así como se habla de omnisciencia lógica para el fragmento B, se habla del problema del *efecto colateral* cuando se trata del fragmento I [27, 41] y, si como decíamos antes, para modelar agentes BDI es usual seguir un sistema lógico de la forma $B^{KD45}D^{KD}I^{KD}$, al enfocarnos en el fragmento I podemos ver que este funciona bajo las normas del axioma K y por ello esta estructura induce un sistema modal normal para las intenciones, y por tanto, el entendimiento de las intenciones en dicha estructura da lugar a los siguientes

problemas (originalmente diseñados para explicar la omnisciencia lógica [101], aquí los seguimos para caracterizar el efecto colateral):

- Si un agente intenta un conjunto de fórmulas Γ y Γ implica una fórmula γ, entonces el agente también intenta γ. A esto se le llama propiedad de *cerradura*.
- Además, bajo la interpretación de mundos posibles, una fórmula válida es aquella que es satisfecha en todo mundo posible. Esto implica que el agente ha de intentar todas las fórmulas válidas de su base de datos, pero este es un problema de potencial *irrelevancia*.
- Un agente no puede intentar γ y $\neg\gamma$ sin intentar cualquier otra cosa. Este es el problema de la *explosión*.
- Computacionalmente, dados los problemas anteriores, un agente tiene que computar todas las consecuencias de sus acciones, lo cual, en la práctica, no parece posible dados los recursos limitados. A esto se le llama *intratabilidad*.

Estos problemas de la noción de efecto colateral pueden resumirse, formalmente, de la siguiente manera [104]:

- $\mathsf{INT}(\phi) \wedge \phi \rightarrow \psi \vdash \mathsf{INT}(\psi)$ (efecto colateral)
- $\phi \rightarrow \psi \vdash \mathsf{INT}(\phi) \rightarrow \mathsf{INT}(\psi)$ (efecto colateral)
- $\vdash \psi \Leftrightarrow \psi \rightarrow\vdash \mathsf{INT}(\phi) \Leftrightarrow \mathsf{INT}(\psi)$ (efecto colateral)
- $\vdash \phi \rightarrow\vdash \mathsf{INT}(\phi)$ (problema de transferencia)
- $(\mathsf{INT}(\phi) \wedge \mathsf{INT}(\psi) \rightarrow \mathsf{INT}(\psi \wedge \phi))$ (combinación irrestricta)
- $\mathsf{INT}(\phi) \rightarrow \mathsf{INT}(\phi \vee \psi)$ (debilitamiento irrestricto)
- $\neg(\mathsf{INT}(\phi) \wedge \mathsf{INT}(\neg\phi))$ (axioma D)

Ninguna de estas propiedades es válida para modelar el comportamiento de las intenciones, excepto por la última expresión, la cual es una norma para representar la consistencia intencional [72].

10.3. Intenciones y no-monotonicidad

Para evitar el problema de la omnisciencia lógica las lógicas no-monotónicas formalizan el fenómeno del razonamiento por *default*; sin embargo, lo hacen de diferentes formas y, por tanto, existen diferentes sistemas no-monotónicos. La mayoría de ellos, no obstante, utilizan como unidad no-estándar la noción de *default* que puede ser descrita mediante expresiones como *típicamente* o *a menos que se muestre de otro modo*.

Esta noción de *default*, común a muchos sistemas, no garantiza que los consecuentes se den siempre que se den los antecedentes y, por tanto, las lógicas no-monotónicas nos permiten derivar consecuencias de manera derrotable para evitar los problemas de la omnisciencia lógica. Con todo, estos útiles sistemas generalmente se las ven con creencias en vez de otros componentes del espectro cognitivo: son no-monotónicos pero no intencional-temporales.

Los avances que se han hecho sobre la propuesta filosófica de Bratman, como hemos visto, son sumamente complejos e interesantes, pero aún existen detalles de su análisis que no han sido explorados completamente. Así, por ejemplo, notamos que los modelos lógicos de agencia BDI basados en un modelo bratmaniano comprenden a las intenciones de dos maneras: o bien como una combinación de creencias y deseos (el modelo BD que es formalizado por $C\&L$) o bien como un elemento independiente pero de forma unitaria (el modelo BDI que es formalizado como INT por Rao y Georgeff).

Ahora, como argumentamos antes, las intenciones tienen una estructura y un comportamiento particular. Esta observación es importante porque las formalizaciones más usadas parecen olvidar que las intenciones tienen una estructura y conducta de reglas y que funcionan como planes, por lo que los formalismos que hemos estudiado no pueden ser completamente útiles para nuestros fines. Sin embargo, a partir del análisis de los lenguajes de programación para agentes BDI, podemos resumir estas propiedades estructurales y funcionales al fijar las intenciones como patrones de la forma $g : ctx \leftarrow body$ [19] donde g denota la meta de la intención, ctx el contexto de dicha intención y $body$ un conjunto de acciones o sub-intenciones. La relevancia de preservar esta estructura es que nos permitirá representar a las intenciones como intenciones derrotables [72] dentro de un modelo bratmaniano.

Lo que haremos ahora será tratar de conservar las características del modelo bratmaniano en un modelo formal. Para proponer un modelo formal cumpliremos los dos requisitos generales que se tienen que tomar en cuenta al desarrollar una teoría lógica: la adecuación material y la formal [5]. La primera tiene que ver con

la captura de un fenómeno objetivo; la segunda con las propiedades lógicas que la teoría satisface.

La cuestión de la naturaleza del razonamiento intencional está relacionada con el aspecto material mientras la cuestión del estatus está directamente conectada con la noción de aspecto formal. En esta sección dedicaremos nuestra atención al primer aspecto.

La formalización del modelo bratmaniano que proponemos trata de representar la doble naturaleza del razonamiento intencional siguiendo las líneas generales del análisis de las propiedades de las intenciones que nos permite distinguir entre propiedades estructurales, funcionales, descriptivas y normativas del modelo BDI.

10.3.1. Una teoría no-monotónica

Por ello, nuestro objetivo ahora será construir un marco lógico en donde podamos expresar los detalles del análisis filosófico del Capítulo 3: en otras palabras, nuestro objetivo ahora es proponer el concepto formal de un modelo bratmaniano.

Definición 47 *(Teoría no-monotónica intencional) Una teoría no-monotónica intencional es una 9-tupla* $\langle bs, ps, F_{bs}, F_{ps}, \vdash, \mid\sim, \dashv, \sim\mid, \succ\rangle$ *donde:*

- bs *denota el conjunto creencias*
- ps *denota el conjunto de intenciones*
- $F_{bs} \subseteq bs$ *denota las creencias básicas*
- $F_{ps} \subseteq ps$ *denota las intenciones básicas*
- $\vdash$ *y* $\dashv$ *son relaciones de consecuencia fuerte*
- $\mid\sim$ *y* $\sim\mid$ *son relaciones de consecuencia débil*
- $\succ \subseteq ps^2$ *t.q.* $\succ$ *es acíclica*

Donde el ítem bs denota las creencias, que entendemos como literales. F_{bs} representa las creencias que se consideran básicas y, similarmente, F_{ps} representa las intenciones básicas. Cada intención $\phi \in ps$ es una estructura $g : ctx \leftarrow body$ donde g representa la meta de la intención—de tal suerte que se preserva la noción de *proactividad*—, ctx es un contexto y el resto denota el cuerpo. Además se asume que las intenciones así representadas están *parcialmente* instanciadas por definición.

La *consistencia interna* se preserva al permitir que el contexto de una intención sea parte de las creencias, $ctx(\phi) \in bs$, y al dejar que g sea la cabeza de la intención. Así la *consistencia fuerte* es implicada por la consistencia interna (dado que la consistencia interna es $ctx(\phi) \in bs$). La *coherencia medios-fines* es representada por el criterio de *admisibilidad*—la restricción de que un agente no intenta opciones contradictorias—y la *jerarquía* de las intenciones es representada por la relación de orden que necesitamos que sea acíclica para evitar conflictos entre intenciones. No es difícil notar, por tanto, que esta teoría es un modelo bratmaniano de inferencia.

Como parte de nuestra notación denotaremos una intención ϕ con cabeza g por $\phi|[g]|$. Además, denotaremos la intención complementaria de g como $\phi|[g^c]|$, es decir, la intención ϕ con $\neg g$ como cabeza. La semántica de esta teoría requerirá una estructura de Kripke $K = \langle S, R, V\rangle$ donde S es el conjunto de configuraciones agente, R es una relación de acceso definida por el sistema de transición de $AgentSpeak(L)$ y V es una función de valuación que va de las configuraciones agente a proposiciones verdaderas en esos estados.

Con este marco podemos proponer una noción de inferencia donde diremos que ϕ es fuertemente (débilmente) derivable a partir de una secuencia Δ si y sólo si hay una prueba de $\Delta \vdash \phi$ ($\Delta \mid\sim \phi$). Y que ϕ no es fuertemente (débilmente) derivable si y sólo si hay una prueba de $\Delta \dashv \phi$ ($\Delta \sim\!\!\mid \phi$), donde $\Delta = \langle bs, ps\rangle$. Para definir la sintaxis y la semántica comenzamos usando $BDI^{CTL}_{AgentSpeak(L)}$ como fue expuesto en el Capítulo 6, usando inicialmente un enfoque similar al de un sistema BDI^{CTL} definido en términos de $B^{KD45}D^{KD}I^{KD}$ con los operadores temporales usuales [40, 54].

10.4. Aspecto material de la inferencia intencional

Podemos definir, en consecuencia, una noción de inferencia en cuatro formas: si la secuencia es $\Delta \vdash \phi$, decimos que ϕ es fuertemente derivable. La idea es que una intención es fuertemente derivable si para cualquier posible camino y estado de la configuración del agente, esta intención pertenece a C_I o C_E. Si, por otro lado, decimos que es débilmente derivable, lo que queremos decir es que o bien es fuertemente derivable o bien es derivable hasta un punto en el que dicha intención ya no es compatible con la configuración del agente.

Finalmente, la construcción de los restantes fragmentos la hacemos por vía

negativa, de tal suerte que si la secuencia es $\Delta \dashv \phi$ decimos que ϕ no es fuertemente derivable; y si es $\Delta \nsim \phi$, entonces ϕ no es débilmente derivable. Si colocamos estas ideas en una definición esquemática obtenemos:

Definición 48 *(Prueba) Sea ϕ una intención. Una prueba de ϕ a partir de Δ es una secuencia finita de fórmulas $BDI^{CTL}_{AgentSpeak(L)}$ que satisface:*

1. $\Delta \vdash \phi$ *syss*

   1.1. $\Box$A(INT(ϕ)) *ó*

   1.2. $\Box$A($\exists \phi[[g]] \in F_{ps}$: BEL($ctx(\phi)$) $\wedge \forall \psi[[g']] \in body(\phi)$ $\vdash \psi[[g']]$)

2. $\Delta \mid\sim \phi$ *syss*

   2.1. $\Delta \vdash \phi$ *ó*

   2.2. $\Delta \dashv \neg\phi$ *y*

      2.2.1. $\Diamond$E(INT(ϕ) U $\neg$BEL($ctx(\phi)$)) *ó*

      2.2.2. $\Diamond$E($\exists \phi[[g]] \in ps$: BEL($ctx(\phi)$) $\wedge \forall \psi[[g']] \in body(\phi)$ $\mid\sim \psi[[g']]$) *y*

         2.2.2.1. $\forall \gamma[[g^c]] \in ps$, $\gamma[[g^c]]$ *falla*[1] *en* Δ *ó*

         2.2.2.2. $\psi[[g']] \succ \gamma[[g^c]]$

3. $\Delta \dashv \phi$ *syss*

   3.1. $\Diamond$E(INT($\neg\phi$)) *y*

   3.2. $\Diamond$E($\forall \phi[[g]] \in F_{ps}$: $\neg$BEL($ctx(\phi)$) $\vee \exists \psi[[g']] \in body(\phi)$ $\dashv \psi[[g']]$)

4. $\Delta \nsim \phi$ *syss*

   4.1. $\Delta \dashv \phi$ *y*

   4.2. $\Delta \vdash \neg\phi$ *ó*

      4.2.1. $\Box$A$\neg$(INT(ϕ) U $\neg$BEL($ctx(\phi)$)) *y*

      4.2.2. $\Box$A($\forall \phi[[g]] \in ps$: $\neg$BEL($ctx(\phi)$) $\vee \exists \psi[[g']] \in body(\phi)$ $\nsim \psi[[g']]$) *ó*

         4.2.2.1. $\exists \gamma[[g^c]] \in ps$ *t.q.* $\gamma[[g^c]]$ *tiene éxito en* Δ *y*

         4.2.2.2. $\psi[[g']] \not\succ \gamma[[g^c]]$

[1] Que una intención *falle* significa que no es aplicable en el sentido de la Definición 22.

10.5. Resumen

Lo que sabemos ahora es que desde un punto de vista cognitivo un agente BDI es un sistema agente que se explica en términos del modelo BDI. Un modelo BDI bratmaniano de razonamiento intencional debe considerar los atributos lógicos de revisión de intenciones y de no-monotonicidad del razonamiento intencional: no hacerlo es construir un modelo no-bratmaniano que no resuelve los problemas externo e interno. El problema que quedaba abierto era el de cómo traducir el análisis del modelo BDI bratmaniano a un modelo formal que resolviera los problemas que los modelos tradicionales no terminaban de resolver.

Con el desarrollo y uso de $BDI^{CTL}_{AgentSpeak(L)}$ evitamos el problema técnico y observamos que la revisión de intenciones y la no-monotonicidad intencional eran atributos lógicos legítimos de los agentes BDI modelados vía $AgentSpeak(L)$. Y aunque nos separamos momentáneamente del problema interno, lo que hemos hecho ahora es mostrar que además de que nuestro análisis es compatible con los agentes BDI interna y externamente desde un punto de vista formal, también es correcto experimentalmente. Sin embargo, aunque este resultado es relevante, su carácter experimental aún se queda corto para nuestros fines. Por ello nos embarcamos en la tarea de interpretar la revisión de intenciones mediante el aprendizaje intencional, de manera precisa, explorando cómo algunos procedimientos de aprendizaje pueden usarse como mecanismos para la revisión de intenciones acercando los términos declarativos de la revisión a los términos procedimentales: lo que hemos hecho es responder el *problema externo* a través de la construcción de un modelo bratmaniano de revisión.

El objetivo de este capítulo ha sido el de acercarnos al desarrollo de un marco no-monotónico para responder al problema interno. Lo que hemos hecho ha sido representar un modelo bratmaniano desde el punto de vista material. Al hacer esto hemos explorado detalles del modelo BDI bratmaniano que no habían sido desarrollados previamente y hemos observado cómo es que la no-monotonicidad del razonamiento intencional es compatible con un modelo BDI bratmaniano.

Capítulo 11

No-monotonicidad de intenciones

11.1. Introducción

En el capítulo anterior sugerimos una representación del razonamiento intencional haciendo énfasis en el aspecto material mediante una representación formal de un modelo bratmaniano. Lo que ahora proponemos es de carácter formal: nos preguntamos por el estatus metalógico de este tipo de razonamiento. Siguiendo la hipótesis de que el razonamiento intencional es un modo de razonamiento *sui generis* por su naturaleza derrotable y temporal, sugerimos que también tiene el derecho a ser llamado *razonamiento lógico* por su comportamiento.

La idea es que si la monotonía no es una propiedad del razonamiento intencional y pretendemos dar una descripción más o menos adecuada de su noción de inferencia entonces debemos investigar cuáles son las propiedades metalógicas que aparecen en lugar de la monotonía puesto que, una vez que ésta es abandonada emerge una pregunta de manera muy orgánica: ¿Por qué deberíamos considerar la inferencia intencional como una instancia de lógica *bona fide*?

Este capítulo está organizado de manera muy simple: en la Sección 11.2 mostramos directamente la cuestión de la adecuación formal mediante el despligue de algunas propiedades metalógicas interesantes.

11.2. Aspecto formal de la inferencia intencional

Ahora que nos enframos al aspecto formal nuestra hipótesis es que el modelo propuesto en el capítulo anterior constituye una lógica bien comportada [64] que satisface condiciones básicas de Consistencia, Corrección, Supraclasicalidad,

Corte y Monotonía Cauta: el modelo bratmaniano que proponemos tiene un buen comportamiento.

11.2.1. Consistencia

Comenzamos sugiriendo un cuadrado de oposición para mostrar relaciones lógicas de consistencia y coherencia:

Proposición 8 *($Subalternas_1$) Si $\vdash \phi$ entonces $\mid\!\sim \phi$.*

Corolario 1 *($Subalternas_2$) Si $\sim\!\mid \phi$ entonces $\dashv \phi$.*

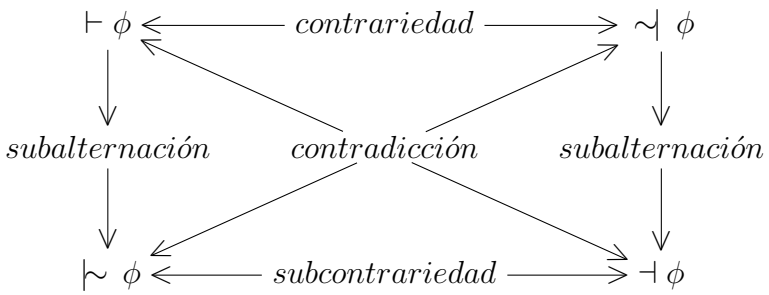
Proposición 9 *($Contradictorias_1$) No existe ϕ t.q. $\vdash \phi$ y $\dashv \phi$.*

Corolario 2 *($Contradictorias_2$) No existe ϕ t.q. $\mid\!\sim \phi$ y $\sim\!\mid \phi$.*

Proposición 10 *($Contrarias$) No existe ϕ t.q. $\vdash \phi$ y $\sim\!\mid \phi$.*

Proposición 11 *($Subcontrarias$) Para toda ϕ, $\mid\!\sim \phi$ ó $\dashv \phi$.*

Un arreglo conveniente de estas proposiciones nos permite formar el siguiente cuadrado de oposición:



Así pues, la Proposición 8 y el Corolario 1 representan Supraclasicalidad; la Proposición 9 y el Corolario 2 manifiestan la Consistencia mientras las restantes proposiciones especifican la coherencia del cuadrado, y así, la coherencia global del modelo bratmaniano.

Ejemplo 9 *Para ilustrar todo esto consideremos un escenario en el que un agente intenta adquirir su doctorado bajo la configuración Δ tal que $F_{bs} = \{\top\}$, $bs = \{beca\}$, $F_{ps} = \{investigar : \top \leftarrow \top\}$, $ps = \{investigar : \top \leftarrow \top; doctorado : \top \leftarrow tesis, examen; tesis : beca \leftarrow investigar; examen : \top \leftarrow investigar\}$. Y supongamos que enviamos la pregunta (query) 'doctorado?'. La búsqueda de intenciones con cabeza '$doctorado$' en F_{ps} falla, por lo que la alternativa $\vdash \phi[doctorado]$ no es el caso. Por tanto, podemos inferir, por contradicción (Proposición 9), que no es fuertemente probable que '$doctorado$', es decir, que en algún momento, en algún estado la intención '$doctorado$' no es el caso. Así, el resultado de la pregunta debería ser que el agente intentará obtener su doctorado derrotablemente bajo la configuración Δ. Por el contrario, la interrogante 'investigar?' tendrá éxito como $\vdash \phi[investigar]$, y por ello diríamos que '$investigar$' es fuertemente (y débilmente) probable (usando la Proposición 8 convenientemente).*

11.2.2. Corrección

Además este modelo bratmaniano es correcto con respesto a su semántica.

Definición 49 *(Satisfacción) ϕ es verdadera en un modelo K si y sólo si ϕ es verdadera en todas las configuraciones agente σ en K. Esto es, $K \models \phi \Leftrightarrow K, \sigma \models \phi$ para toda $\sigma \in S$.*

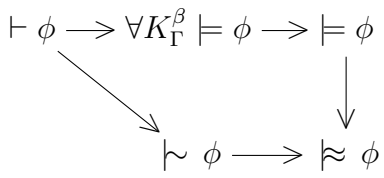
Definición 50 *(Corrida de un agente en un modelo) Dada una configuración inicial β, un sistema de transición Γ y una valuación V, $K^{\beta}_{\Gamma} = \left\langle S^{\beta}_{\Gamma}, R^{\beta}_{\Gamma}, V \right\rangle$ denota una corrida de un agente en un modelo.*

Definición 51 *(Validez) Una fórmula $\phi \in BDI^{CTL}_{AgentSpeak(L)}$ es verdadera para cualquier corrida agente en Γ si y sólo si $\forall K^{\beta}_{\Gamma} \models \phi$.*

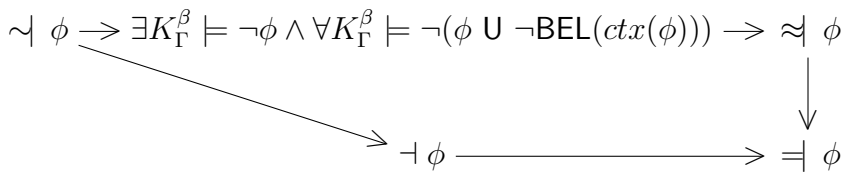
Al denotar

$$(\exists K^{\beta}_{\Gamma} \models \phi \ \mathsf{U}\ \neg \mathsf{BEL}(ctx(\phi)))\ \mathsf{v}\ \models \phi$$

como $\mid\!\approx\ \phi$, y asumiendo $\models \phi \geq \mid\!\approx\ \phi$ y $\approx\!\mid\ \phi \geq =\!\mid\ \phi$, es posible encontrar una serie de *traducciones* tales que:



Y de la misma manera para el resto de los fragmentos:



De tal suerte que:

Proposición 12 *Las siguientes relaciones son el caso:*

$$a)\ Si \vdash \phi\ entonces \models \phi \qquad b)\ Si\ \mid\!\sim\ \phi\ entonces\ \mid\!\approx\ \phi$$

Corolario 3 *Análogamente:*

$$a)\ Si\ \dashv \phi\ entonces =\!\!|\ \phi \qquad b)\ Si \sim\!\!|\ \phi\ entonces \approx\!\!|\ \phi$$

11.2.3. Más propiedades metalógicas

Pero hay otras propiedades metalógicas que pueden ser exploradas para definir la racionalidad del razonamiento intencional, esto es, para determinar su buen comportamiento.

En primer lugar, resulta bastante razonable la propiedad de Supraclasicalidad: si ϕ se sigue de Δ de manera monotónica, entonces también debe seguirse de manera no-monótona. Así, necesitamos el requisito razonable de que las intenciones fuertemente mantenidas tienen que mantenerse también débilmente, pero no a la inversa:

Proposición 13 *($Supraclasicalidad$) Si $\Delta \vdash \phi$, entonces $\Delta\ \mid\!\sim\ \phi$.*

Otra propiedad interesante nos dice que si un conjunto intencional es clásicamente consistente, entonces también lo es el conjunto de consecuencias derrotables de éste. Esta propiedad se denomina:

Proposición 14 *($Preservación\ de\ Consistencia$) Si $\Delta\ \mid\!\sim\ \bot$, $\Delta \vdash \bot$.*

Una propiedad más interesante es una forma de Corte Cauto. Esta dicta que si ϕ es una consecuencia de Δ, entonces ψ es una consecuencia de Δ y ϕ sólo si ya es una consecuencia de Δ por sí sola. En otras palabras, que añadir algunas intenciones a Δ que ya son consecuencia de Δ no conlleva a ningún *incremento* de información. En términos bratmanianos, esto significa que la longitud de una planificación no afecta el grado en que la información de Δ apoya ϕ:

Proposición 15 *(Corte Cauto) Si* $\Delta \mid\sim \phi$ *y* $\Delta, \phi \mid\sim \psi$ *entonces* $\Delta \mid\sim \psi$.

Si seguimos buscando encontramos una forma de Monotonicidad Cauta que funciona como la conversa del Corte. Esta propiedad nos dice, en términos bratmanianos, que añadir una consecuencia ϕ de vuelta en Δ no conlleva a ningún *decremento* de información en la planificación, esto es, que añadir información implícita a los planes o a las intenciones es una tarea monotónica.

Así pues, Monotonicidad Cauta nos dice que el razonamiento intencional es un razonamiento cumulativo en el sentido de que podemos extraer consecuencias que pueden ser usadas como premisas adicionales sin afectar el conjunto de conclusiones:

Proposición 16 *(Monotonicidad Cauta) Si* $\Delta \mid\sim \psi$ *y* $\Delta \mid\sim \gamma$ *entonces* $\Delta, \psi \mid\sim \gamma$.

Si atamos Corte Cauto y Monotonicidad Cauta podemos decir que si ϕ es una consecuencia de Δ entonces para cualquier ψ, ψ es consecuencia de Δ si y sólo si es una consecuencia de Δ junto con ϕ.

11.3. Resumen

Lo que sabemos ahora es que desde un punto de vista cognitivo un agente BDI es un sistema agente que se explica en términos del modelo BDI. Un modelo BDI bratmaniano de razonamiento intencional debe considerar los atributos lógicos de revisión de intenciones y de no-monotonicidad del razonamiento intencional: no hacerlo es construir un modelo no-bratmaniano que no resuelve los problemas externo e interno. El problema que quedaba abierto era el de cómo traducir el análisis del modelo BDI bratmaniano a un modelo formal que resolviera los problemas que los modelos tradicionales no terminaban de resolver.

Con el desarrollo y uso de $BDI^{CTL}_{AgentSpeak(L)}$ evitamos el problema técnico y observamos que la revisión de intenciones y la no-monotonicidad intencional eran atributos lógicos legítimos de los agentes BDI modelados vía $AgentSpeak(L)$. Y aunque nos separamos momentáneamente del problema interno, lo que hemos hecho ahora es mostrar que además de que nuestro análisis es compatible con los agentes BDI interna y externamente desde un punto de vista formal, también es correcto experimentalmente. Sin embargo, aunque este resultado es relevante, su carácter experimental aún se queda corto para nuestros fines. Por

ello nos embarcamos en la tarea de interpretar la revisión de intenciones mediante el aprendizaje intencional, de manera precisa, explorando cómo algunos procedimientos de aprendizaje pueden usarse como mecanismos para la revisión de intenciones acercando los términos declarativos de la revisión a los términos procedimentales: lo que hemos hecho es responder el *problema externo* a través de la construcción de un modelo bratmaniano de revisión.

En el capítulo anterior nos acercamos al desarrollo de un marco no-monotónico para responder al problema interno. Lo que hemos hecho ha sido representar un modelo bratmaniano desde el punto de vista material. Al hacer esto hemos explorado detalles del modelo BDI bratmaniano que no habían sido desarrollados previamente y hemos observado cómo es que la no-monotonicidad del razonamiento intencional es compatible con un modelo BDI bratmaniano.

Con lo que hemos hecho en este último capítulo parece razonable concluir que el modelo bratmaniano aquí desarrollado permite capturar aspectos relevantes de la naturaleza y el estatus de la inferencia intencional. Más aún, este modelo así desarrollado permite ver que tal inferencia es bien comportada dado que satisface condiciones de Consistencia, Corrección, Supraclasicalidad, Corte y Monotonía Cauta. Por tanto, es razonable concluir que este tipo de inferencia intencional se comporta lógicamente en el buen sentido del término; estrictamente, se comporta como una lógica no-monotónica. Y con esto respondemos al *problema interno*.

Capítulo 12

Conclusiones

Antes de presentar un balance final de nuestros resultados nos gustaría hacer un resumen general.

12.1. Resumen

Después de una exposición de carácter protocolar sobre el origen y los fines de este trabajo (Capítulo 1) nos enfrentamos al problema de caracterizar un agente BDI como un sistema agente que se explica en términos del modelo BDI bratmaniano (Capítulo 2). Posteriormente observamos que un modelo BDI bratmaniano de razonamiento intencional debe considerar los atributos lógicos de revisión y no-monotonicidad intencional (Capítulo 3). El problema que quedaba abierto era, sin embargo, el de cómo traducir el análisis de las propiedades del modelo BDI bratmaniano para construir un modelo formal de revisión e inferencia que nos permitiera responder tanto al problema externo como al problema interno y que, de algún modo, resolviera los problemas que los modelos tradicionales no terminaban de resolver (Capítulos 4 y 5). Con el desarrollo y uso de $BDI^{CTL}_{AgentSpeak(L)}$ evitamos el problema técnico y además observamos que los atributos lógicos de revisión y no-monotonicidad eran compatibles y relevantes para los agentes BDI (Capítulo 6).

En seguida argumentamos que además de que nuestro análisis era compatible con los agentes BDI, también era aplicable experimentalmente y era fidedigno con el modelo BDI bratmaniano (Capítulo 7). Pero si bien el resultado experimental era prometedor nos quedaba corto, por lo que nos embarcamos en la tarea de verificar que la revisión de intenciones podía entenderse como un mecanismo de

aprendizaje (Capítulos 8 y 9). Y con esto construimos un modelo bratmaniano de revisión de intenciones con el que ofrecimos una respuesta al problema externo.

Después, como nos preguntábamos por los mecanismos subyacentes a un tipo de razonamiento que conlleva creencias e intenciones, también distinguíamos el problema interno. Partiendo del análisis del Capítulo 3 y aplicando la herramienta lógica desarrollada en el Capítulo 6 propusimos una respuesta al problema interno argumentando que, en efecto, es posible construir un modelo bratmaniano de inferencia con ciertos mecanismos de razonamiento cuyo comportamiento es estrictamente lógico (Capítulos 10 y 11), con lo que ofrecimos una respuesta al problema interno.

12.2. Balance

Al principio nos preguntábamos si había características en las intenciones que permitieran *i)* tanto un mecanismo de revisión como *ii)* un mecanismo de inferencia. Siguiendo el enfoque BDI original desarrollamos el concepto de *modelo bratmaniano* con el que argumentamos que sí las hay. Lo interesante, claro, era saber cómo se comportaban tales mecanismos.

A la primera pregunta (problema externo) respondimos afirmando que la revisión funciona como aprendizaje intencional; a la segunda pregunta (problema interno) respondimos que la inferencia funciona como una inferencia lógica en sentido estricto. Y con esto tenemos elementos de juicio suficientes para concluir que, en efecto, el razonamiento intencional sí es razonamiento lógico *bona fide* porque responde a las exigencias externas e internas.

Así, considerando un conjunto de procedimientos de aprendizaje pudimos visualizar cómo podemos entender la revisión de intenciones mediante el aprendizaje de intenciones. La relevancia de este punto consistió en visualizar cómo es que cambian esas estructuras de datos llamadas intenciones. El eslogan del problema externo diría: *las intenciones son tan revisables como las creencias*.

Esto último es interesante porque no sólo nos permite argumentar que las intenciones son objetos legítimos para una teoría de revisión de intenciones, sino que además esta teoría puede implementarse y es, por ello, mecanizable.

Por otro lado, bajo el problema interno, nuestra principal contribución ha sido el estudio de la naturaleza y el estatus de la inferencia intencional para mostrar que tal inferencia es lógica. La relevancia de este segundo punto consistió en visualizar cómo es la inferencia intencional y lo que pudimos observar puede resumirse en la fórmula: *la inferencia intencional es tan lógica como la inferencia*

con creencias.

Por tanto, en el marco del problema de la persistencia podemos ver que el razonamiento intencional es razonamiento lógico *bona fide*. Y por ello, no sólo puede reproducirse mecánicamente sino que merece el mismo tipo de atención que el razonamiento que utiliza creencias.

En consecuencia, si bien estamos de acuerdo en que para que un agente sea racional es necesario tomar en cuenta un poder inferencial contextual, recursos limitados y falibilidad bajo condiciones de optimización y racionalidad acotada, creemos que estas no son condiciones conjuntamente suficientes pues, ciertamente, hemos sugerido que nuestra noción de racionalidad está más allá del fragmento B.

En efecto, dado que es posible considerar otros componentes del modelo BDI sin perder rigor, podemos desarrollar sistemas lógicos dentro de un marco BDI más generoso, pero no menos riguroso. Podemos tener sistemas y modelos lógicos *stricto sensu* dentro de un modelo cognitivo de agencia BDI a través del concepto de *modelo bratmaniano*.

Esto último es de gran importancia, pues cuando Bratman afirmaba que su teoría definía un marco de tipo cognitivo para entender a *otros*[1] y a nosotros mismos, apuntaba a que podemos exigir el comportamiento lógico como un desiderátum a nuestro alcance[2] a través de la *desmitificación* de las intenciones.

Nosotros pensamos que, además de esto último, el estudio de los modelos bratmanianos nos permite ver que las intenciones no sólo son componentes propios y singulares del modelo BDI, sino que además tienen un mecanismo *lógico*—no fenomenológico o psicológico—; un mecanismo que, en principio, se puede comprender y reproducir automáticamente en nosotros y en *otros* agentes, lo cual es alentador desde nuestro enfoque cognitivo: para la filosofía como para la IA, el *proyecto lógico* de la comprensión y la reproducción del comportamiento intencional no sólo sigue vivo, sino que está vigente.

12.3. Trabajo futuro

El trabajo futuro está compuesto por todas las preguntas que no pudimos responder en este trabajo pero que logramos visualizar. Podríamos enunciar centenas

[1] De nuevo, el subrayado es nuestro.

[2] Un desiderátum que nos otorga el deber y nos garantiza el derecho a demandar la identificación, evaluación y construcción de razones de carácter intencional

de este tipo de preguntas, pero sólo mencionaremos las que podemos distinguir con mucha claridad y las que estamos trabajando actualmente:

- El *problema de la mecanización* que también es el problema de la algoritmia. Aunque dimos algunos detalles algorítmicos en § 8.4 y acercamos la revisión a una implementación, faltan detalles algorítmicos y de complejidad sobre la noción de inferencia intencional que necesitan ser explorados incluso experimentalmente.

- Tenemos además el *problema de la representación* que también es el problema de la formalización. Si bien nuestro formalismo hace uso de estructuras lógicas multimodales BDI no se hace cargo de una lógica de eventos que puede ofrecer más expresividad, por supuesto, a cambio de mayor complejidad; asimismo, es necesario investigar los resultados de esta investigación a la luz de una álgebra de procesos.

- Además podemos mencionar el *problema de la simetría* que también es el problema de la equivalencia entre la revisión y la no-monotonicidad intencional. Este es el problema de la correspondencia entre la revisión de intenciones y la no-monotonicidad intencional. Dado que el modelo BDI permite la inferencia intencional y la revisión de intenciones, es fácil sospechar que hay una relación entre éstas. La principal idea consiste en identificar la revisión de un conjunto Δ por una intención ϕ con una inferencia derrotable a partir de Δ.

- Y, finalmente, un problema interesante de carácter lógico y filosófico es el *problema de la extensión*, que consiste en definir a un modelo bratmaniano como la clase de extensiones no-monotónicas de los modelos de Kripke, de tal manera que podamos considerar a los modelos bratmanianos como estructuras lógicas legítimas y singulares tanto para la representación de información como para la generación de nuevos modelos inferenciales.

Bibliografía

[1] Alchourrón, C.E., Gärdenfors, P., Makinson, D.: "On the logic of theory change: partial meet contraction and revision functions". *Journal of Symbolic Logic*, 50, 1985.

[2] Alechina, N., Bordini. R.H., Hübner, J.F., Jago, M., Logan,B.: "Belief revision for AgentSpeak agents". *Proceedings of the Sixth International Joint Conference on Autonomous Agents and Multi-Agent Systems*, AAMAS 2006, 812, Hakodate, Hokkaido, Japan, 2006.

[3] Allen, J.F.: "Towards a general theory of action and time". *Artificial Intelligence*, 23(2):123-154, 1984.

[4] Anscombe, G.E.M.: *Intention*. Cornell University Press, Ithaca, 1963.

[5] Antonelli, A.: *Grounded consequence for defeasible logic*. Cambridge University Press, Cambridge, 2005.

[6] Antoniou, G.: "A discussion of some intuitions of defeasible reasoning". Vouros, G.A., Panayiotopoulos, T. (eds.) *SETN 2004*, LNAI 3025, pp. 311–320, Springer-Verlag, 2004.

[7] Aristóteles: *Acerca del alma*. Madrid, Gredos, 1978.

[8] Aristóteles: "Primeros Analíticos". *Tratados de Lógica. Órganon*. Madrid, Gredos, 1998.

[9] Aristóteles: *Ética Nicomáquea*. Madrid, Gredos, 1985.

[10] Audi, R.: *Intending*. Journal of philosophy, 70, 1973.

[11] Austin, J.: *Lectures on jurisprudence*. London, John Murray, 1873.

[12] Austin, J.L.: *Cómo hacer cosas con palabras*. Madrid, Paidós, 1990.

[13] Baral, C., Zhao, J.: "Non-monotonic temporal logics for goal specification". *IJCAI*, 2007.

[14] Baral, C., Zhao, J.: "Non-monotonic temporal logics that facilitate elaboration tolerant revision of goals". *Proceedings of the Twenty-Third AAAI Conference on Artificial Intelligence*, 2008.

[15] Bates, J.: "The role of emotion in believable agents". *Communications of the ACM*, 37(7), 122-125, 1994.

[16] Benthem van, J.: "Dynamic logic for belief revision". *Journal of Applied Non-Classical Logics*, 17, 2007.

[17] Bordini, R.H., Bazzan, A.L.C., Jannone, R.O., Basso, D.M., Vicari, R.M., Lesser, V. R.: "AgentSpeak(XL): efficient intention selection in BDI agents via decision-theoretic task scheduling". *Proceedings of the First International Joint Conference on Autonomous Agents and Multi-Agent Systems*, AAMAS 2002, 15-19 July, 1294-1302, 2002.

[18] Bordini, R.H., Moreira, Á.F.: "Proving BDI properties of agent-oriented programming languages". *Annals of Mathematics and Artificial Intelligence*, 42, 197-226, 2004.

[19] Bordini, R.H., Hübner, J.F., Wooldridge, M.: *Programming multi-agent systems in AgentSpeak using Jason*. Wiley, England, 2007.

[20] Bordini, R.H., Dastani, M., Dix, J., El-Fallah-Seghrouchni, A.: *Multi-agent programming. languages, platforms and applications*. Springer, USA, 2005.

[21] Brand, M.: *Intending and Acting*. MIT Press, 1984.

[22] Bratko, I.: *Prolog Programming for artificial intelligence*. Addison-Wesley, USA, 1986.

[23] Bratman, M.: "Intention and means-end reasoning". *The Philosophical Review*, 90, 1981.

[24] Bratman, M.: "Taking plans seriously". *Social Theory and Practice*, 9, 1983.

[25] Bratman, M.: "Two faces of intention". *The Philosophical Review*, 93, 1984.

[26] Bratman, M.: "Davidson's theory of intentions". Vermazen, B., Hintikka, M.B. (eds.) *Essays on Davidson: actions and events*, Oxford University Press, 1985.

[27] Bratman, M.: *Intention, plans, and practical reason*. Harvard University Press, Cambridge, MA., 1987.

[28] Bratman, M.: *Faces of intention*. Cambridge University Press, Cambridge, MA., 1999.

[29] Bratman, M.: *Structures of agency. Essays*. Oxford University Press, Oxford, 2007.

[30] Bratman, M., Pollak, M.E., Israel D.J.: "Plans and resource-bounded practical reasoning". *Computer Intelligence*, 4, 349-355, 1988.

[31] Brentano, F.: *Psychology from an empirical standpoint*. Routledge, London, 1973.

[32] Brooks, R.A.: *Cambrian intelligence: the early history of the new AI*. MIT Press, Cambridge, MA., 1999.

[33] Brooks, R.A.: "Elephants don't play chess". *Robotics and Autonomous Systems*, 6, 3-15, 1990.

[34] Buchanan, B.G.: "A (very) brief history of artificial intelligence". *AAAI*, 2005.

[35] Chen, X., Liu, G.: "A logic of intention". *ICJAI-99*, 1999.

[36] Chisholm, R.: *Perceiving*. Ithaca, Cornell University Press, 1957.

[37] Chisholm, R.: *Theory of knowledge*. Englewood Cliffs, NJ, Prentice-Hall, 1966.

[38] Churchland, P.: "The logical character of action-explanations". *The philosophical review*, 79, 1970.

[39] Churchland, P.: "Eliminative materialism and the propositional attitudes". *Journal of Philosophy*, 78, 67-90, 1981.

[40] Clarke, E. M. Jr., Grumberg, O., Peled, D.A.: *Model checking*. MIT Press, Boston, MA., 1999.

[41] Cohen, P., Levesque, H.: "Intention is choice with commitment". *Artificial Intelligence*, 42(3), 213-261, 1990.

[42] Collins, A.: "Why Cognitive Science". *Cognitive Science*, 1977.

[43] Copeland, J.: *Artificial intelligence. A philosophical introduction*. Blackwell, 1993.

[44] Couturat, L.: *La logique de Leibniz d'aprés de documents inédits*. G. Olms, Hildesheim, 1962.

[45] Dastani, M., Governatori, G., Rotolo, A., van der Torre, L.: "Preferences of agents in defeasible logic". Zhang, S., Jarvis, R. (eds.) *AI 2005*, LNAI 3809, pp. 695–704, Springer-Verlag Berlin-Heidelberg, 2005.

[46] Dastani, M., Birna van Riemsdijk, M., Meyer, J.J.Ch.: "A grounded specification for angent programs". *Proceedings of the Seventh International Joint Conference on Autonomous Agents and Multi-Agent Systems*, AAMAS 2007, Mayo 14 - 18, 2007.

[47] Dastani, M., Hindriks, K.V., John-Jules, Ch.M.: *Specification and verification of multi-agent systems*. Springer, 2010.

[48] Davidson, D.: *Essays on actions and events*. Oxford University Press, New York, 1980.

[49] Dean, T., McDermott, D.: "Temporal data base management". *Artificial Intelligence*, 32(1), 1–55, 1987.

[50] Dennett, D.: *Brainstorms*. Harvesters Press, 1981.

[51] Dennett, D.: *The intentional stance*. MIT Press, Cambridge, MA., 1987.

[52] Dignum, F., Meyer, J.J.Ch., Wieringa, R.J., Kuiper, R.: "A modal approach to intentions, commitments and obligations: Intention plus commitment yields obligation". *Deontic logic, agency and normative systems*, 80-97, 1996.

[53] Dreyfus, H.L.: *What Computers Still Can't Do. A Critique of Artificial Reason* MIT Press, Cambridge, MA., 1992.

[54] Emerson, A.: "Temporal and modal logic". *Handbook of Theoretical Computer Science*, Elsevier Science Publishers B.V., Amsterdam, 1990.

[55] Etzioni, O.: "Intelligence without robots". *AI Magazine*, 14(4), 1993.

[56] Fagin, R., Ullman, J.D., Vardi, M.Y.: "On the semantics of updates in databases". *Proceedings of Second ACM*, SIGACT-SIGMOD, 352-365, 1983.

[57] Feinberg, B., Kasrils, R. (eds): *Dear Bertrand Russell... A selection of his correspondence with the general public 1950–1968*. London, George Allen and Unwin, 1969.

[58] Ferber, J.: *Les systemes multi-Agents: vers une intelligence collective*. InterEditions, Paris, 1995.

[59] Fisher, M.: *An introduction to practical formal methods using temporal logic*. John Wiley & Sons, 2011.

[60] Floridi, L.: *Philosophy and computing. An introduction*. Routledge, New York, 1999.

[61] Floridi, L. (ed.): *The Blackwell guide to the philosophy of computing and information*. Blackwell, Oxford, 2004.

[62] Fodor, J.: *La explicación psicológica*. Cátedra, Madrid, 1980.

[63] Franklin, S., Graesser, A.: "Is it an agent, or just a program?: A taxonomy for autonomous agents". *Proceedings of the Third International Workshop on Agent Theories, Architectures, and Languages*, Springer-Verlag, 1996.

[64] Gabbay, D. M.: "Theoretical foundations for nonmonotonic reasoning in expert systems". Apt, K. (ed.) *Logics and Models of Concurrent Systems*, Berlin and New York: Springer Verlag, pp. 439–459, 1985.

[65] Galliers, J.R.: *A theoretical framework for computer models of cooperative dialogue, acknowledging multi-agent conflict*. Open University, UK, 1998.

[66] Gardenfors, P., Makinson, D.: "Revisions of knowledge systems using epistemic entrenchment". *Proceedings of the Second Conference on Theoretical Aspects of Reasoning about Knowledge*, 352-365, 1988.

[67] Gardner, H.: *La nueva ciencia de la mente*. Paidós, Barcelona, 1988.

[68] Genesereth, M.R., Nilsson, N.J.: *Logical foundations for artificial intelligence*. Morgan Kauffman, Palo Alto, CA., 1987.

[69] Genesereth, M.R., Ketchpel, S.P.: "Software agents". *Communications of the ACM*, 37(7), 48-53, 1994.

[70] Georgeff, M., Pell, B., Pollack, M., Tambe, M., and Wooldridge, M.: "The Belief-Desire-Intention model of agency". Muller. J.P., Singh, M., Rao, A. (eds.) Intelligent Agents V, LNAI-1365, Springer-Verlag, 1999.

[71] Goodwin, R.: *Formalizing properties of agents*. Technical Report CMU-CS-93-159, School of Computer Science, Carnegie-Mellon University, Pittsburgh, 1993.

[72] Governatori, G., Padmanabhan, V., Sattar, A.: "A defeasible logic of policy-based intentions". *Advances in Artificial Intelligence*, LNAI-2557, 2002.

[73] Governatori, G., Rotolo, A.: "Defeasible logic: agency, intention and obligation". Lomuscio, A., Nute, D. (eds.) *DEON 2004*, LNAI 3065, pp. 114–128, Springer-Verlag, 2004.

[74] Governatori, G., Terenziani, P.: "Temporal extensions to defeasible Logic". *Proceedings of IJCAI'07 Workshop on Spatial and Temporal Reasoning*, India, 2007.

[75] Grim, P., Mar, G., St. Denis, P.: *The philosophical computer. Exploratory essays in philosophically computer modeling*. Bradford books, 1998.

[76] Guerra-Hernández, A., El-Fallah-Seghrouchni, A., Soldano, H.: "Learning in BDI multi-agent systems". *CLIMA IV, Revised and Selected Papers*, CLIMA IV, LNCS 3259:218–233, 2004.

[77] Guerra-Hernández, A., Ortíz-Hernández, G.: "Toward BDI sapient agents: learning intentionally". *Toward Artificial Sapience: Principles and Methods for Wise Systems*, London, Springer, 2008.

[78] Guerra-Hernández, A., Ortíz-Hernández, G., Luna-Ramírez, W. A.: "Jason smiles: incremental BDI MAS learning". *MICAI 2007. Special Session*, IEEE CSP, 2008.

[79] Guerra-Hernández, A., Castro-Manzano, J.M., El-Fallah-Seghrouchni, A.: "Toward an AgentSpeak(L) theory of commitment and intentional learning". *MICAI 2008*, LNCS, vol. 5317, 2008.

[80] Guerra-Hernández A., Castro-Manzano, J.M., El-Fallah-Seghrouchni, A.: "CTLAgentSpeak(L): a specification language for agent programs". *Journal of Algorithms in Cognition, Informatics and Logic*, 2009.

[81] Haddadi, A.: *Communication and cooperation in agent systems: a pragmatic theory*. Springer Verlag, Berlin-Heidelberg, 1995.

[82] Hájek P.: *Metamathematics of Fuzzy Logic*. Dordrecht, Kluwer, 1998.

[83] Harper, W.L.: *Rational conceptual change*. East Lansing, Michigan, 1977.

[84] Hart, H.L.A.: "The ascription of responsibility and rights". *Proceedings of the Aristotelian Society*, 1948-9, 1948.

[85] Haugeland, J.: *Artificial intelligence: the very idea*. MIT Press, Cambridge, 1985.

[86] Hoek, W. van der, Jamroga, W., Wooldridge, M.: "Towards a theory of intention revision". *Synthese*, Springer-Verlag, 2007.

[87] Hübner, J.F.: *Um modelo de reorganizacão de sistemas multiagentes*. Universidad de de São Paulo, Escola Politecnica, 2003.

[88] Hughes, G.E., Cresswell, M.J.: *A new introduction to modal logic*. Routledge, London, 1998.

[89] Hume, D.: *Investigación sobre el conocimiento humano. Investigación sobre los principios de la moral*. Tecnos, 2007.

[90] Hunsberger, L., Ortiz, Ch.: "Dynamic intention structures I: a theory of intention representation". *Auton. Agent Multi-Agent Syst* 16:298–326, 2008.

[91] Husserl, E.: *Ideas relativas a una fenomenología pura y una filosofía fenomenológica* FCE, México, 1949.

[92] D'Inverno, M.: "A formal specification of dMARS". *Intelligent Agents IV: Proceedings of the Fourth International Workshop on Agent Theories, Architectures and Languages*, LNAI 1365, 1997.

[93] D'Inverno, M., Luck, M.: "Engineering AgentSpeak(L): a formal computational model". *J. Logic Computat.*, Vol. 8 No. 3, pp. 1-27, Oxford University Press, 1998.

[94] Jahne, B., Haußbecker, H. (ed.): *Computer vision and applications. A guide for students and practicioners*. Academic Press, San Diego, 2000.

[95] Jörgensen, J.: "Imperatives and Logic". *Erkenntnis* 7, p. 288-296, 1937/8.

[96] Katsuno, H., Mendelzon, A.: "On the difference between updating a knowledge base and revising". *Principles of knowledge representation and reasoning*, Morgan Kaufmann, San Mateo, CA, 1991.

[97] Kinny, D., Georgeff, M.P.: "Commitment and effectiveness of situated agents". *Proceedings of the twelfth international joint conference on artificial intelligence*, IJCAI-91, Sydney, Australia, 1991.

[98] Kiss, G.: "Variable coupling of agents to their environment: combining situated and symbolic automata". Demazeau, Y., Muller, J.P. (eds.) *Descentralized A.I.3.*, 231-248, North-Holland, 1992.

[99] Kneale, W.C., Kneale, M. *The Development of logic*. Oxford University Press, 1962.

[100] Konolige, K., Pollack, M.E.: "A representationalist theory of intentions". *Proceedings of International Joint Conference on Artificial Intelligence*, IJCAI-93, 1993.

[101] Sim, K.M.: "Epistemic logic and logical omniscience: A survey". *International Journal of Intelligent Systems*, 12, 57-81, 1997.

[102] Leibniz, G.W.: "Lengua universal, característica y lógica". *Obras filosóficas y científicas*, 5. Comares, Granada, 2013.

[103] Levi, I.: *The enterprise of knowledge*. MIT Press, Cambridge, MA, 1980.

[104] Linder van, B.: *Modal logic for rational agents*. Department of Computer Science, Utrecht University, 1996.

[105] Lorini, E., Herzig, A.: "A logic of intention and attempt". *Knowledge, Rationality & Action. Synthese* 163:45–77, 2008.

[106] Macnish, C.K.: "Nonmonotonic temporal reasoning: a logic-based approach to declarative modeling". *Cybernetics and Systems: An International Journal*, 29:42, 459, 1998.

[107] Makinson, D.: "How to go nonmonotonic". Gabbay, D., Guenthner, F. (eds.) *Handbook of Philosophical Logic*, Volume 12, 175–278, Springer, 2005.

[108] McCarthy, J.: "Notes on formalizing context". *Proceedings of the Thirtheen International Joint Conference on Artificial Intelligence*, 1993.

[109] McCarthy, J., Minsky, M., Rochester, N., Shannon, C.: *A proposal for the Dartmouth Summer Research Project on Artificial Intelligence*, 1955.

[110] McCarthy, J.: *Ascribing mental qualities to machines*. Technical report, Computer Science Department, Stanford University, Stanford, 1979.

[111] McCarthy, J., Hayes, P.: "Some philosophical problems from the standpoint of artificial intelligence". Metzer, B., Michie, D. (eds.) *Machine Intelligence 4*. Edinburgh, Edinburgh University Press, 1969.

[112] McDermott, D., Doyle, J.: "Non-monotonic logic I". *Artificial Intelligence* 13: 41-72, 1980.

[113] McDermott, D.: "A temporal logic for reasoning about processes and plans". *Cognitive Science*, 6(2), 101–155, 1982.

[114] McGown, G. W. Th.: *A Primer of Burns*. A. Gardner, 1907.

[115] Meseguer, J.: "General logics". *Logic Colloquium* 87, 275-329. North Holland, 1989.

[116] Morado, R.: "Problemas filosóficos de la lógica no-monotónica". *Enciclopedia Iberoamericana de Filosofía*, vol. 27, Filosofía de la lógica, Trotta, Madrid, 2004

[117] Moreira, Á.F., Bordini, R.H.: "An operational semantics for a BDI agent-oriented programming language". *Proceedings of the Workshop on Logics for Agent-Based Systems*, LABS-2002, Toulouse, France, 2002.

[118] Moreira, Á.F., Vieira, R., Bordini, R.H.: "Extending the operational semantics of a BDI agent-oriented programming language for introducing speech-act based communication". *Declarative Agent Languages and Technologies,*

Proceedings of the First International Workshop, DALT-03, LNAI 2990:135-154, 2004.

[119] Moro Simpson, T.: *Semántica y filosofía: Problemas y discusiones*. Eudeba, Buenos Aires, Argentina, 1964.

[120] Narayanan.: *On being a machine*. Ellis Horwood, 1991.

[121] Newell, A; Simon, H. A.: "GPS: A program that simulates human thought". Feigenbaum, E.A.; Feldman, J. (eds.) *Computers and Thought*, New York: McGraw-Hill, 1963.

[122] Newell, A., Simon, H. A.: "Computer science as empirical inquiry: symbols and search". *Communications of the ACM* 19 (3): 113–126, 1976.

[123] Nilsson, N.: *Inteligencia artificial. Una nueva síntesis*. McGraw Hill, México, 2006.

[124] Nilsson, U., Maluszyński, J.: *Logic, programming and prolog* 2ed. John Wiley and Sons, London, 1995.

[125] Nute, D.: "Defeasible logic". *INAP 2001*, LNAI 2543M 151-169, Springer-Verlag, 2003.

[126] Icard, Th., Pacuit. E., Shoham, Y.: "Joint revision of belief and intention". *Proceedings of the Twelfth International Conference on the Principles of Knowledge Representation and Reasoning*, 2010.

[127] Plotkin, G.: *A structural approach to operational semantics*. Laboratory for Foundations of Computer Science, School of Informatics, University of Edinburgh, Scotland, 2004.

[128] Pollack, M.: "Overloading practical reasoning". *Nous* 25:513-536, 1991.

[129] Pollock, J.: "The structure of epistemic justification". *American Philosophical Quarterly*, monograph series 4: 62-78, 1970.

[130] Pollock, J.: "Perceptual knowledge". *Philosophical Review*, 80, 287-319, 1971.

[131] Pollock, J.: *Knowledge and justification*. Princeton University Press, 1974.

[132] Pollock, J., Iris, O.: "Vision, knowledge, and the mystery link". *Philosophical Perspectives* 19, 2006.

[133] Prakken, H., Vreeswijk, G.: "Logics for defeasible argumentation". Gabbay, D., Guenthner, F. (eds.) *Handbook of Philosophical Logic*, Vol 4, pp. 219-318. Kluwer Academic Publishers, Dordrecht, 2002.

[134] Rao, A.S., Georgeff, M.P.: "Modelling rational agents within a BDI architecture". *Readings in Agents*, 317-328, 1998.

[135] Rao, A.S., Georgeff, M.P.: *Asymmetry thesis and side-effect problems in lineartime and branching-time intention logics*. Technical Note 13, Australian Artificial Intelligence Institute, Carlton, Victoria, IJCAI-91, LNAI 2990:135-154, 1991.

[136] Rao, A.S.: "AgentSpeak(L): BDI agents speak out in a logical computable language". *MAAMAW*, LNCS 1038:42-55, 1996.

[137] Rao, A.S., Georgeff, M.P.: "Decision procedures for BDI logics". *Journal of Logic and Computation* 8(3), 293-342, 1998.

[138] Reiter, R.: "A logic for default reasoning". *Artificial Intelligence* 13, 81–132, 1980.

[139] Rescher, N.: *Dialectics*. Albany, SUNY Albany Press, 1977.

[140] Rosenschein, J.S., Genesereth, M.R.: "Deals among rational agents". *Proceedings of the Ninth International Joint Conference on Artificial Intelligence*, IJCAI-85, 91-99, Los Angeles, CA, 1985.

[141] Russell, S.J., Norvig, P.: *Artificial intelligence, a modern approach*. Prentice Hall, New Jersey, 1995.

[142] Sadek, M.D.: "A logic in the study of intention". *Proceedings of the Conference in Knowledge Representation and Reasoning*, 1992.

[143] Schut, M., Wooldridge, M., Parsons, S.: "The theory and practice of intention reconsideration". *J. Expt. Theor. Artif. Intell.* 16, No. 4, 261–293, 2004.

[144] Searle, J.R.: *Speech acts: An essay in the philosophy of language*. Cambridge University Press, England, 1969.

[145] Seel, N.: *Formalising first-order intentional system theory*. Technical report, STC Technology LTD, 1989.

[146] Shoham, Y.: "Logical theories of intention and the database perspective". *Journal of Philosophical Logic*, 38:633–647, 2009.

[147] Simon, H.A.: "A Behavioral Model of Rational Choice". *Models of Man, Social and Rational: Mathematical Essays on Rational Human Behavior in a Social Setting*. New York, Wiley, 1957.

[148] Simon, H.A.: *Las ciencias de lo artificial*. Editorial Comares, 2006.

[149] Singh, M.: "A critical examination of the Cohen-Levesque theory of intentions". *Proceedings of the European Conference on Artificial Intelligence*, ECAI, 1992.

[150] Singh, M.: *Multiagent systems: A theoretical framework for intentions, know how, and communication*. LNCS, Springer Verlag, Berlin-Heidelberg, 1995.

[151] Singh, M., Rao, A.S., Georgeff, M.P.: "Multiagent systems: A modern approach to distributed artificial intelligence". *Formal Methods in DAI: Logic-Based Representation and Reasoning*, 331-376, 1999.

[152] Shoham, Y.: *Agent-oriented programming*. Technical Report STAN - CS - 1335 - 90, Computer Science Department, Stanford University, 1990.

[153] Shoham, Y., Cousins, S.B.: "Logics of mental attitudes in AI". Lakemeyer, G., Nebel, B. (eds.) *Foundations of Knowledge Representation and Reasoning*, LNAI, Springer-Verlag, 1994.

[154] Sudeikat, J., Braubach, L., Pokahr, A., Lamersdorf, W., Renz, W.: "Validation of BDI agents". Bordini, R.H. *et al*. (eds.) *ProMAS 2006*, LNAI 4411, pp. 185–200, Springer-Verlag, Berlin-Heidelberg, 2007

[155] Sugimoto, T.: "A preference-based theory of intention". *PRICAI-2000*, Springer-Verlag, 2000.

[156] Tarski, A.: "On some fundamental concepts of metamathematics". *Logic, semantics, metamathematics*. Hackett Publishing Company, 1928.

[157] Toulmin, S.: *The place of reason in ethics*. Cambridge, 1958.

[158] Turner, R., Eden, A.: "The philosophy of computer science". *The Stanford Encyclopedia of Philosophy (Summer 2009 Edition)*, Zalta, E.N. (ed.) http://plato.stanford.edu/archives/sum2009/entries/computer-science/.

[159] Verhagen, H.: "Autonomy and reasoning for natural and artificial agents". Nickles, M., Rovatsos, M., Weiß, G. (eds.) *Agents and Computational Autonomy - Potential, Risks, and Solutions*, Lecture Notes in Computer Science 2969 Springer, p. 83-94, 2004.

[160] White, J.E.: *Telescript technology: The foundation for the electronic marketplace*. White paper, General Magic, Inc., 2465 Latham Street, Mountain View, CA, 1994.

[161] Wobcke, W.: "Intention and rationality for PRS-like agents". McKay, R.I., Slaney, J. (eds.) *AI 2002*, LNAI 2557, pp. 167–178, Springer-Verlag, Berlin Heidelberg, 2002.

[162] Wobcke, W.: "An analysis of three puzzles in the logic of intention". Sattar, A., Kang, B.H. (eds.) *AI 2006*, LNAI 4304, pp. 403–412, Springer-Verlag, Berlin-Heidelberg, 2006.

[163] Wobcke, W.: "Plans and the revision of intentions". *Distributed Artificial Intelligence Architecture and Modelling*, LNCS 1087, 100-114, Springer, Heidelberg, 2006.

[164] Wooldridge, M., Parsons, S.: "Intention reconsideration reconsidered". Muller, J.P. *et al.* (eds.) *ATAL'98*, LNAI 1555, pp. 63-79, Springer-Verlag, Berlin-Heidelberg, 1999.

[165] Wooldridge, M.: *Reasoning about rational agents*. MIT Press, Cambridge, MA., 2000.

[166] Wooldridge, M.: *Introduction to multiagent systems*. John Wiley and Sons, London, 2001.

Parte III

Apéndices

Apéndice A

Demostraciones y bosquejos

Capítulo 6

Proposición 1 *Los agentes BDI satisfacen el axioma de no-retención infinita:* $\mathsf{INT}(\phi) \Rightarrow \mathsf{A}\Diamond(\neg\mathsf{INT}(\phi))$.

Prueba. Asumamos $K, c_0 \models \mathsf{INT}(\phi)$. Entonces, por la definición de INT, existe un plan $p \in C_I \vee C_E$ con cabeza $+!\phi$ en c_0. El axioma de no-retención infinita expresa que para toda corrida K^0_Γ en algún momento p será removido de C_I (intenciones activas) o C_E (intenciones suspendidas). Mientras p está siendo exitosamente ejecutado hay tres posibles corridas dadas las reglas de transición: $ClrInt_i \in \Gamma$: i) $ClrInt_3$ se aplica cuando el cuerpo de p no es vacío, y nada es limpiado; ii) $ClrInt_2$ se aplica cuando el plan p, con un cuerpo vacío, está en el tope de i, por lo que p es extraída de i; $ClrInt_1$ se aplica cuando i tiene sólo un plan p con un cuerpo vacío, la intención completa i es extraída. Dada la naturaleza finita de los planes, las condiciones *ii)* o *iii)* son finalmente alcanzadas. Alternativamente, si algo sale mal con p, un mecanismo de falla se activa generando un evento de la forma $\langle -!\phi, i[p]\rangle$ resultando en el abandono de p. Por tanto, sea la ejecución exitosa o fallida, toda intención adoptada es en algún momento abandonada. ■

Proposición 2 *Los agentes BDI no satisfacen el axioma de compromiso ciego.*[1]

Prueba. Dado que los agentes BDI satisfacen el axioma de no-retención infinita (Proposición 1) el axioma de compromiso ciego se reduce a:[2]

$$\mathsf{INT}(\mathsf{A}\Diamond(\phi)) \Rightarrow \mathsf{A}\Diamond(\mathsf{BEL}(\phi))$$

Mostramos un contra-ejemplo. Considérese una configuración inicial c_0 t.q. $ag = \langle bs, ps\rangle$ donde $bs = \{\}$ y $ps = \{+b(t_1) : \top \leftarrow p(t_2). \quad +!p(t_2) : \top \leftarrow +b(t_3).\}$. Supóngase que a partir de la percepción del ambiente se añade $ag_{bs} = \{b(t_1)\}$. Un evento es generado por esta actualización de creencias, de tal manera que $C_E = \{\langle +b(t_1), \top\rangle\}$. Entonces, siguiendo las reglas semánticas de Γ, $SelEv_1$, Rel_1, $ApplPl_1$ se aplican obteniendo una configuración donde $C_I = \{[+b(t_1) : \top \leftarrow !p(t_2).]\}$ y $C_E = \{\}$. Entonces, siguiendo con las reglas $SelAppl$, $ExtEv$, $SelInt_1$, $AchvGl$ se obtiene una configuración donde $C_E = \{\langle +!p(t_2), +b(t_1) : \top \leftarrow \top\rangle\}$, $C_I = \{\}$. En esta configuración c', $K, c' \models \mathsf{DES}(p(t_2))$. Si aplicamos $SelEv_1$, Rel_1, $AppPl_1$, $SelAppl$, se obtiene una configuración donde $C_I = \{[+!p(t_2) : \top \leftarrow \quad +b(t_3).]\}$ y $C_E = \{\}$. En esta configuración c'' $K, c'' \models \mathsf{INT}(p(t_2))$. Entonces, procediendo con $IntEv$, $SelInt_1$, $AddBel$ se llega a $C_E = \{\langle +b(t_3), \top\rangle\}$, $ag_{bs} = \{b(t_1)\}$ y $C_I = \{[+b(t_1) : \top \leftarrow \top \ddagger +!p(t_2) : \top \leftarrow \top]\}$ y $bs = \{b(t_1), b(t_3)\}$. La intención de que $p(t_2)$ se mantiene. Nótese que los cuerpos de los planes son vacíos, de tal suerte que las reglas $ClrInt$ descartarán la intención completa, por lo que en la siguiente configuración c''', $K, c''' \models \neg\mathsf{INT}(p(t_2))$ y $K, c''' \models \neg\mathsf{BEL}(p(t_2))$. ■

Proposición 3 *Los agentes BDI satisfacen una forma limitada de compromiso flexible:* $\mathsf{INT}(\mathsf{A}\Diamond(\phi)) \Rightarrow \mathsf{A}(\mathsf{INT}(\mathsf{A}\Diamond(\phi))\ \mathsf{U}\ \neg\mathsf{BEL}(\mathsf{E}\Diamond(\phi)))$

Prueba. Este caso es similar a la prueba de no-retención infinita. Asumamos que el agente intenta $\mathsf{INT}(\mathsf{A}\Diamond(\phi))$ en c_0, entonces hay un $p \in C_I \vee C_E$ con cabeza $+!\phi$ en c_0. Si hay una configuración $c_k \geq 0$ donde $\neg\mathsf{BEL}(\mathsf{E}\Diamond(\phi))$ (usando *until* débil), entonces $K, x^{0,\ldots,k} \models \mathsf{INT}(\mathsf{A}\Diamond(\phi))$. Siguiendo la estrategia de prueba de la Proposición 1, en los casos de fallo el agente satisfará $\bigcirc\neg\mathsf{INT}(\phi)$ dada Rel_2, lo cual implica que para un evento $\langle te, i[+!\phi : c \leftarrow h.]\rangle$ no existen planes relevantes y la intención asociada será descartada, es decir, no hay una secuencia

[1] De hecho, no satisfacen el axioma de compromiso ciego extendido dado que el agente no mantiene su intención $p(t_2)$ hasta que esta es creída. Este razonamiento es similar al de la demostración de incompletud intención-creencia (AT2) en $AgentSpeak(L)$ [18].

[2] Una reducción similar es usada por Rao *et al* [137].

de configuraciones en las que en algún momento ocurra ϕ, de tal manera que es racional abandonar $\mathsf{INT}_{\langle ag,C\rangle}(\phi)$. El caso de no-retención infinita por ejecución exitosa de intenciones cubre la segunda condición del *until* débil, de tal manera que $\neg\mathsf{BEL}(\mathsf{E}\Diamond(\phi))$ puede no ocurrir. ■

Proposición 4 *Sea Ag el conjunto de agentes BDI. El compromiso flexible completo se satisface por $Ag \cup$ Abandon.*

Prueba. Tenemos que probar que

$$\mathsf{INT}(\mathsf{A}\Diamond(\phi)) \Rightarrow \mathsf{A}(\mathsf{INT}(\mathsf{A}\Diamond(\phi))\ \mathsf{U}\ (\mathsf{BEL}(\phi) \vee \neg\mathsf{BEL}(\mathsf{E}\Diamond(\phi))))$$

Asumimos $\mathsf{INT}(\mathsf{A}\Diamond(\phi))$. Primero probamos que, para cualquier curso de configuraciones, $\mathsf{INT}(\mathsf{A}\Diamond(\phi))$ se da hasta $(\mathsf{BEL}(\phi) \vee \neg\mathsf{BEL}(\mathsf{E}\Diamond(\phi)))$. Tenemos, pues, por la definición de *until*, dos alternativas: $K^0_\Gamma, x^i \models \mathsf{INT}(\mathsf{A}\Diamond(\phi))$ ó $K^0_\Gamma, x^i \models (\mathsf{BEL}(\phi) \vee \neg\mathsf{BEL}(\mathsf{E}\Diamond(\phi)))$. Si se da la primera alternativa, entonces $\forall j, K^0_\Gamma, x^j \models \mathsf{INT}(\mathsf{A}\Diamond(\phi))$, que es la definición de *until* débil. Si la segunda opción se da, entonces, para algún x^k, tenemos que $K^0_\Gamma, x^k \models (\mathsf{BEL}(\phi) \vee \neg\mathsf{BEL}(\mathsf{E}\Diamond(\phi)))$, y para todo $0 \leq j < k, K^0_\Gamma, x^j \models \mathsf{INT}(\mathsf{A}\Diamond(\phi))$, que es de nuevo la definición del *until* débil.

En los casos exitosos, $\mathsf{INT}(\mathsf{A}\Diamond(\phi))$ se da hasta que $\mathsf{BEL}(\phi)$. Sin embargo, en los casos de fallo el agente satisfará $\bigcirc\neg\mathsf{INT}(\phi)$, en cuyo caso tendríamos un evento de la forma $\langle +abandon(\phi), \top\rangle$ t.q. $C'_E = C_E - \{+abandon(\phi), \top\}$ y $C'_I = C_I - \phi$, y $ag_{bs} \not\models intending(\phi)$; lo que a su vez significa que el agente abadona ϕ y no cree que ϕ es una opción, es decir, $\neg\mathsf{BEL}(\mathsf{E}\Diamond(\phi))$. ■

Capítulo 9

Proposición 5

$$abandon(\phi, C_I) \Rightarrow C_I \ominus \phi$$

Prueba. Si asumimos *abandon*, los seis postulados para la contracción se deben mantener. Caso 1: Por definición el resultado de *abandon* es un conjunto de intenciones C_I t.q. C_I es un conjunto intencional. Caso 2: La aplicación de $abandon(\phi, C_I)$ produce $C'_I = C_I - \phi_i$, con $C'_I \subseteq C_I$. Caso 3: Si $\phi \notin C_I$, entonces $abandon(\phi, C_I) = C_I$. Caso 4: Si $\phi \notin Cn(C_I)$, entonces $\phi \notin C_I$. Luego, $abandon(\phi, C_I) = C_I$. Caso 5: Si $\phi \in C_I$, entonces $C_I \subseteq abandon(\phi, C_I) \cup \phi$. Caso 6: Si $\phi \Leftrightarrow \psi$, $abandon(\phi, C_I) = abandon(\psi, C_I)$. ■

Proposición 6

$$learn(\phi, C_I) \Rightarrow C_I \oplus \phi$$

Prueba. Asumiendo *learn* la definición de *expansión* debería mantenerse, es decir, $\Sigma \oplus \phi = C_I \cup \phi$. Sólo hay un caso en el que $learn(\phi, C_I) = C_I \cup \phi$ ya sea que $\phi \in C_I$ ó $\phi \notin C_I$. ■

Proposición 7

$$learn(\phi, abandon(\phi^c, C_I)) \Rightarrow C_I \odot \phi$$

Prueba. Asumiendo la composición de ambas funciones, la *revisión* debería mantenerse. Caso 1: $learn(\phi, abandon(\phi^c, C_I))$ produce C_I, que es un conjunto intencional. Caso 2: Cualquier ϕ es aceptada en $learn(\phi, abandon(\phi^c, C_I))$. Caso 3: $learn(\phi, abandon(\phi^c, C_I) \subseteq learn(\phi, C_I)$. Por la observación 2, ya sea que $\phi \in C_I$ ó $\phi \notin C_I$, $learn(\phi, abandon(\phi^c, C_I) \subseteq C_I$, es decir, es un subconjunto de $learn(\phi, C_I)$. Caso 4: Si $\phi^c \notin C_I$ entonces $learn(\phi, C_I) \subseteq learn(\phi, abandon(\phi^c, C_I))$. Por la observación 2, si $\phi \in C_I$ entonces $learn(\phi, abandon(\phi^c, C_I)) = C_I$ y si $\phi \notin C_I$ entonces $learn(\phi, abandon(\phi^c, C_I)) = C_I \cup \phi$, esto es, $learn(\phi, C_I) \subseteq learn(\phi, abandon(\phi^c, C_I))$. Caso 5: Si $\vdash \neg\phi$, entonces para cualquier configuración agente, $\neg\phi \in C_I$. Si aplicamos $learn(\phi, abandon(\phi^c, C_I))$, obtenemos C_I t.q. $\phi \in C_I$, lo cual es una contradicción. Caso 6: Si $\phi \Leftrightarrow \psi$, $learn(\phi, abandon(\phi^c, C_I)) = learn(\psi, abandon(\psi^c, C_I))$. ■

Capítulo 11

Proposición 8 *($Subalternas_1$) Si $\vdash \phi$ entonces $\mid\!\sim \phi$.*

Prueba. Supongamos que $\vdash \phi$ pero no $\mid\!\sim \phi$, es decir, $\sim\!\mid \phi$. Entonces, dado $\vdash \phi$ tenemos dos casos generales. Caso 1: dada la suposición inicial de que $\vdash \phi$, por Definición 48 ítem 1.1, tenemos que $\Box$A(INT(ϕ)). Ahora, por la segunda suposición, de que $\sim\!\mid \phi$, por la Definición 48 ítem 4.1, llegamos a $\dashv \phi$. Y por tanto, $\Diamond$E(INT($\neg\phi$)), y entonces derivamos $\neg\phi$; sin embargo, dada la suposición inicial, también obtenemos ϕ, lo cual es una contradicción.

Caso 2: dada la suposición de que $\vdash \phi$, por Definición 48 ítem 1.2, tenemos que $\exists\phi|[g]| \in F_{ps}$: BEL($ctx(\phi)$) $\wedge\, \forall\psi|[g^c]| \in body(\phi) \vdash \psi|[g^c]|$. Ahora, por la segunda suposición tenemos que $\sim\!\mid \phi$, y entonces $\dashv \phi$, por lo que obtenemos

$\Diamond$E($\forall\phi|[g]| \in F_{ps} : \neg$BEL$(ctx(\phi)) \vee \exists\psi|[g^c]| \in body(\phi) \dashv \psi$), y por tanto podemos obtener $\forall\phi|[g]| \in F_{ps} : \neg$BEL$(ctx(\phi)) \vee \exists\psi|[g^c]| \in body(\phi) \dashv \psi$) lo cual es $\neg(\exists\phi|[g]| \in F_{ps}$: BEL$(ctx(\phi)) \wedge \forall\psi|[g^c]| \in body(\phi) \vdash \psi|[g^c]|)$. ■

Proposición 9 *($Contradictorias_1$) No existe ϕ t.q. $\vdash \phi$ y $\dashv \phi$.*

Prueba. Asumamos que existe un ϕ t.q. $\vdash \phi$ y $\dashv \phi$. Si $\dashv \phi$ entonces, por Definición 48 ítem 3.1, $\Diamond$E(INT$(\neg\phi)$). Entonces podemos obtener $\neg\phi$. Sin embargo, puesto que $\vdash \phi$ también se sigue que ϕ, lo cual es absurdo. ■

Proposición 10 *(Contrarias) No existe ϕ t.q. $\vdash \phi$ y $\sim\!\!\!|\ \phi$.*

Prueba. Supongamos que existe un ϕ t.q. $\vdash \phi$ y $\sim\!\!\!|\ \phi$. Por la Proposición 8, se sigue que $|\!\sim \phi$, pero eso contradice la suposición $\sim\!\!\!|\ \phi$ por el Corolario 2. ■

Proposición 11 *(Subcontrarias) Para toda ϕ, $|\!\sim \phi$ ó $\dashv \phi$.*

Prueba. Asumamos que no es el caso que para toda ϕ, $|\!\sim \phi$ ó $\dashv \phi$. Entonces hay un ϕ t.q. $\sim\!\!\!|\ \phi$ y $\vdash \phi$. Tomando $\sim\!\!\!|\ \phi$ se sigue del Corolario 1 que $\dashv \phi$. Por la Proposición 9 llegamos a una contradicción con $\vdash \phi$. ■

Proposición 12 *Las siguientes relaciones se dan:*

$$a)\ Si \vdash \phi\ entonces \models \phi \qquad b)\ Si\ |\!\sim \phi\ entonces\ |\!\approx \phi$$

Prueba. Caso base: Tomamos Δ_i como una secuencia con $i = 1$. Caso *a)* Asumiendo $\vdash \phi$, tenemos dos subcasos. El primer subcaso está dado por la Definición 48 ítem 1.1. Por tanto tenemos que $\Box$A(INT(ϕ)). Esto implica, por Definicion 39 ítems P4 y S5 y la Definición 38, que para todo camino y todo estado $\phi \in C_I \vee C_E$. Podemos representar esta expresión, por medio de una traducción, en términos de corridas. Dado que los caminos son secuencias de estados y los estados son configuraciones agente tenemos que $\forall K_\Gamma^\beta \models \phi$, lo cual implica $\models \phi$. El segundo subcaso está dado por la Definición 48 ítem 1.2, lo cual en términos de corridas significa que para toda corrida $\exists\phi|[g]| \in F_{ps}$: BEL$(ctx(\phi)) \wedge \forall\psi|[g']| \in body(\phi) \vdash \psi|[g']|$. Dado que Δ_1 es un único paso, $body(\phi) = \top$ y para toda corrida BEL$(ctx(\phi)))$, es decir, $ctx(\phi) \in F_{bs}$. Entonces $\forall K_\Gamma^\beta \models \phi$ lo cual, al igual que arriba, implica $\models \phi$.

Caso *b)* Supongamos que $|\!\sim \phi$. Entonces tenemos dos subcasos. El primero está dado por la Definición 48 ítem 2.1. Por ello, tenemos que $\vdash \phi$ lo cual, como

arriba, ya implica que $\models \phi$. Por otro lado, por el ítem 2.2, tenemos $\dashv \neg\phi$ y dos alternativas. La primera alternativa, ítem 2.2.1, es $\Diamond$E(INT(ϕ) U $\neg$BEL($ctx(\phi)$)). Así, podemos reducir esta expresión por medio de la Definición 39 ítems P3 y S4, a una traducción en términos de corridas: $\exists K_\Gamma^\beta \models \phi$ U $\neg$BEL($ctx(\phi)$), lo cual implica $\not\approx \phi$. La segunda alternativa viene del ítem 2.2.2, $\Diamond$E($\exists\phi|[g]| \in ps$: BEL($ctx(\phi)$) $\wedge \forall\psi|[g']| \in body(\phi) \not\sim \psi|[g']|$) lo cual en términos de corridas significa que para alguna corrida $\exists\phi|[g]| \in ps$: BEL($ctx(\phi)$) $\wedge \forall\psi|[g']| \in body(\phi) \not\sim \psi|[g']|$, pero Δ_1 es un paso único, y por tanto $body(\phi) = \top$. Por tanto, hay una corrida en la que $\exists\phi|[g]| \in ps$: BEL($ctx(\phi)$), es decir, ($\exists K_\Gamma^\beta \models (\phi$ U $\neg$BEL($ctx(\phi)$)) usando el caso débil de la Definición 48 P5. Luego, por adición, ($\exists K_\Gamma^\beta \models (\phi$ U $\neg$BEL($ctx(\phi)$))v $\models \phi$, por lo que $\not\approx \phi$.

Caso inductivo. Caso *a)* Supongamos que para $n \leq k$, si $\Delta_n \vdash \phi$ entonces $\Delta \models \phi$. Y supongamos que Δ_{n+1}. Más aún, supongamos que $\Delta_n \vdash \phi$, entonces tenemos dos alternativas. La primera, por Definición 48 ítem 1.1, implica que tenemos una intención ϕ t.q. $ctx(\phi) = body(\phi) = \top$. Dado que $body(\phi)$ es vacío, trivialmente es el caso en n, y por la hipótesis de inducción, $body(\phi) \subseteq \Delta_{n+1}$, y por tanto $\models \phi$. En segundo lugar, por la Definición 48 ítem 1.2, para toda corrida $\exists\phi|[g]| \in ps$: BEL($ctx(\phi)$) $\wedge \forall\psi|[g']| \in body(\phi) \vdash \psi|[g']|$. Por tanto, para toda corrida n, $\forall\psi|[g']| \in body(\phi) \vdash \psi|[g']|$, y por la hipótesis de inducción, $body(\phi) \subseteq \Delta_{n+1}$, es decir, $\Delta \vdash \psi|[g']|$. Por tanto, $\models \phi$.

Caso *b)* Supongamos que $n \leq k$, si $\Delta_n \not\sim \phi$ entonces $\Delta \not\approx \phi$. Y supongamos que Δ_{n+1}. Asumamos que $\Delta_n \not\sim \phi$. Tenemos dos alternativas. La primera viene de la Definición 48 ítem 2.1, es decir, $\vdash \phi$, que ya implica $\models \phi$. La segunda alternativa viene del ítem 2.2, $\Delta \dashv \neg\phi$ con dos subcasos: $\Diamond$E(INT(ϕ) U $\neg$BEL($ctx(\phi)$)) ó $\Diamond$E($\exists\phi|[g]| \in ps$: BEL($ctx(\phi)$) $\wedge \forall\psi|[g']| \in body(\phi) \not\sim \psi|[g']|$). Si consideramos el primer subcaso hay corridas n que cumplen con la definición de $\not\approx \phi$. En el caso restante tenemos que $\forall\psi|[g']| \in body(\phi) \not\sim \psi|[g']|$, dado que $body(\phi) \subseteq \Delta_n$, por la hipótesis de inducción $\Delta \not\sim \psi|[g^c]|$, y entonces, $\Delta_{n+1} \not\sim \phi$, es decir, $\not\approx \phi$. ■

Proposición 13 *($Supraclasicalidad$) Si $\Delta \vdash \phi$, entonces $\Delta \not\sim \phi$.*

Prueba. Similar a la Proposición 8. ■

Proposición 14 *(Preservación de Consistencia) Si* $\Delta \mid\!\sim \bot$, $\Delta \vdash \bot$.

Prueba. Consideremos la forma de la intención $\bot$. Dicha intención es la intención de la forma $\phi \wedge \neg\phi$, que es, por tanto, imposible de intentar, esto es, para cualquier corrida agente $\mid\!\sim \bot$ nunca es alcanzada. Por tanto $\Delta \mid\!\sim \bot$ es falsa, lo cual hace a la implicación verdadera. ■

Proposición 15 *(Corte Cauto) Si* $\Delta \mid\!\sim \phi$ *y* $\Delta, \phi \mid\!\sim \psi$ *entonces* $\Delta \mid\!\sim \psi$.

Prueba. Comencemos por transformar la proposición original en la siguiente: si $\Delta \nmid\!\sim \psi$ entonces no es el caso que $\Delta \mid\!\sim \phi$ y $\Delta, \phi \mid\!\sim \psi$. Esta proposición puede transformarse de nuevo: si $\Delta \nmid\!\sim \psi$ entonces $\Delta \nmid\!\sim \phi$ ó $\Delta, \phi \nmid\!\sim \psi$ de donde, usando el Corolario 1, podemos inferir: si $\Delta \dashv \psi$ entonces $\Delta \dashv \phi$ ó $\Delta, \phi \dashv \psi$. Ahora, supongamos que $\Delta \dashv \psi$ pero que no es el caso que $\Delta \dashv \phi$ ó $\Delta, \phi \dashv \psi$, es decir, que $\Delta \dashv \psi$ pero $\Delta \vdash \phi$ y $\Delta, \phi \vdash \psi$. Considerando la expresión $\Delta, \phi \vdash \psi$ tenemos dos alternativas: $\psi \in body(\phi)$ ó $\psi \notin body(\phi)$. En el primer caso, puesto que $\Delta \vdash \phi$ y dado que $\psi \in body(\phi)$ se sigue que $\vdash \psi$, pero eso contradice la suposición que $\Delta \dashv \psi$. En el caso restante, si $\Delta, \phi \vdash \psi$ pero $\psi \notin body(\phi)$, entonces $\Delta \vdash \psi$, lo cual contradice $\Delta \dashv \psi$. ■

Proposición 16 *(Monotonicidad Cauta) Si* $\Delta \mid\!\sim \psi$ *y* $\Delta \mid\!\sim \gamma$ *entonces* $\Delta, \psi \mid\!\sim \gamma$.

Prueba. Transformemos la proposición original: si $\Delta, \psi \nmid\!\sim \gamma$ entonces no es el caso que $\Delta \mid\!\sim \psi$ y $\Delta \mid\!\sim \gamma$. Así, si $\Delta, \psi \nmid\!\sim \gamma$ entonces $\Delta \nmid\!\sim \psi$ ó $\Delta \nmid\!\sim \gamma$, y por el Corolario 1, si $\Delta, \psi \dashv \gamma$ entonces $\Delta \dashv \psi$ ó $\Delta \dashv \gamma$. Ahora, supongamos que $\Delta, \psi \dashv \gamma$ pero no $\Delta \dashv \psi$ ó $\Delta \dashv \gamma$, es decir, que $\Delta, \psi \dashv \gamma$ y $\Delta \vdash \psi$ y $\Delta \vdash \gamma$. Considerando la forma de $\Delta, \psi \dashv \gamma$ tenemos dos alternativas: $\gamma \in body(\psi)$ ó $\gamma \notin body(\psi)$. En el primer caso, puesto que $\gamma \in body(\psi)$ y $\Delta \dashv \psi$, entonces $\dashv \gamma$, lo cual contradice la suposición de que $\Delta \vdash \gamma$. Si consideramos la segunda alternativa, $\Delta \dashv \gamma$, pero eso contradice la suposición de que $\Delta \vdash \gamma$. ■

Apéndice B

Publicaciones, ponencias y productos

Durante el desarrollo de este trabajo obtuvimos varios resultados. En este apéndice los dividimos en tres categorías: publicaciones, ponencias y otros productos.

Publicaciones

A través del quinto LAM[1] (por *International Workshop on Logics, Agents, and Mobility*) se publicó:

- J. M. Castro-Manzano. *Modelling intentional reasoning with temporal and defeasible logic*. In Michael Köhler-Bußmeier, editor, Joint Proceedings of the 5th International Workshop on Logics, Agents, and Mobility 2012, CEUR Workshop Proceedings, volume 853, Germany, 2012.

[1]En su quinta edición fue un evento satélite que se llevó a cabo en el marco del *33rd Conference on Theory and Application of Petri Nets and Concurrency*, en Hamburgo, Alemania; Universidad de Hamburgo. El LAM es un evento que se viene realizando desde el 2008 ininterrumpidamente y ha sido organizado por Berndt Müller y Michael Köhler-Bußmeier con el objetivo general de reunir investigadores activos en el área de lógica y otros métodos formales que pueden ser usados para describir sistemas móviles o dinámicos, siendo su principal enfoque el área de la lógica para agentes. Sus tópicos principales son la especificación y el razonamiento sobre agentes y sistemas agentes, las lógicas modales y temporales, el *model checking* y la programación lógica.

Mediante el MICAI[2] (por *Mexican International Conference on Artificial Intelligence*) se han publicado los siguientes trabajos:

- J. M. Castro-Manzano. *A defeasible logic of intention*. Advances in Artificial Intelligence. LNCS, vol. 7629, Springer, 2013.
- J. M. Castro-Manzano, Axel Arturo Barceló-Aspeitia and Alejandro Guerra-Hernández. *Consistency and soundness for a defeasible logic of intention*. Advances in soft computing algorithms, Research in Computing Science vol. 54, 2011.
- J. M. Castro-Manzano. *The argument from autonomy revisited*. In MICAI 2010: 9th Mexican International Conference on Artificial Intelligence, Special Session, Los Alamitos, 2010. IEEE Computer Society CPS.
- J. M. Castro-Manzano, Axel Arturo Barceló-Aspeitia and Alejandro Guerra-Hernández. *Intentional learning procedures as intention revision mechanisms*. In MICAI 2010: 9th Mexican International Conference on Artificial Intelligence, Special Session, Los Alamitos, 2010. IEEE Computer Society CPS.
- J. M. Castro-Manzano. *The revision of intentions*. In A. Gelbuch, M. González Mendoza and O. Herrera Alcántara, editors, MICAI 2009, Complementary Proceedings of MICAI 2009, 2009.

A través del LANMR[3] (por *Latin American Workshop on New Methods of Reasoning/Logic, Algorithms and Non-monotonic Reasoning*) hemos publicado los siguientes trabajos:

- J. M. Castro-Manzano. *Formal properties of intentional reasoning*. In M. Osorio, C. Zepeda, I. Olmos, C. Medina and J. Arrazola, editors, Proceedings of the Seventh Latin American Workshop on Non-Monotonic Reasoning 2012 (LANMR'12), CEUR Workshop Proceedings, volume 911, Germany, 2012.

[2]MICAI es el encuentro internacional más importante de hispanoamérica que cubre todas las áreas de la IA. Es organizado por la Sociedad Mexicana de Inteligencia Artificial (SMIA).

[3]El LANMR es un evento organizado por el GMLogyC (Grupo Mexicano de Lógica y Computación) en conjunto con la Benemérita Universidad Autónoma de Puebla (BUAP) y la Universidad de las Américas Puebla (UDLAP). Se viene realizando desde el 2004 y su objetivo es unir investigadores activos de diferentes instituciones en áreas de lógica, lenguajes formales, algoritmos, razonamiento no-monotónico y nuevos métodos de razonamiento para presentar trabajos teóricos innovadores y aplicaciones originales en esas áreas.

- J. M. Castro-Manzano, Axel Arturo Barceló-Aspeitia and Alejandro Guerra-Hernández. *Intentional reasoning as non-monotonic reasoning*. In M. Osorio, C. Zepeda, I. Olmos, C. Medina and J. Arrazola, editors, Proceedings of the Seventh Latin American Workshop on Non-Monotonic Reasoning 2011 (LANMR'11), CEUR Workshop Proceedings, volume 804, Germany, 2011.

- J. M. Castro-Manzano. *An introduction to intention revision: issues and problems*. In M. Osorio, C. Zepeda, I. Olmos, C. Medina and J. Arrazola, editors, Proceedings of the Fifth Latin American Workshop on Non-Monotonic Reasoning 2009 (LANMR'09), CEUR Workshop Proceedings, volume 533, Germany, 2009.

También hemos publicado algunos trabajos en el COMIA[4] (*Congreso Mexicano de Inteligencia Artificial*) y en la revista Factótum[5]:

- J. M. Castro-Manzano, Axel Arturo Barceló-Aspeitia y Alejandro Guerra-Hernández. *Razonamiento intencional y no-monotonicidad*. En M. González Mendoza y O. Herrera Alcántara (eds.), Avances Recientes en Sistemas Inteligentes, Editorial SMIA, México, 2011.

- J. M. Castro-Manzano, Axel Arturo Barceló-Aspeitia y Alejandro Guerra-Hernández. *Un marco no-monotónico para representar intenciones bajo una arquitectura BDI*. En M. González Mendoza y O. Herrera Alcántara (eds.), Avances en Sistemas Inteligentes en México, Editorial SMIA, México, 2010.

- J. M. Castro-Manzano. El argumento de la autonomía y el caso de los agentes adaptativos, *Factótum. Revista de Filosofía* (Salamanca, España), 7, 2010, pp 14-27.

[4]El COMIA es un evento organizado por la Sociedad Mexicana de Inteligencia Artificial (SMIA). Su objetivo es promover la investigación y enseñanza de la IA en universidades de nivel superior mexicanas, y se promueve como un foro científico para la presentación y publicación de trabajos de investigación derivados de tesis o proyectos, terminados o en proceso, en español.

[5]La revista de filosofía Factótum es una revista online editada por la Asociación Cultural Factótum (Salamanca, España). Los textos recibidos son evaluados mediante revisión ciega por pares y todos los contenidos son de libre acceso.

Ponencias

Asimismo, logramos presentar nuestros avances en el LACAP,[6] en el Primer y el Segundo Congreso de Alumnos de Posgrado de la UNAM y en el Coloquio del Posgrado en Filosofía de la Ciencia de la UNAM:

- J. M. Castro-Manzano. *Cambio racional de intenciones: no-monotonicidad*. En G. Soberán Chávez (coord.), Segundo Congreso de Alumnos de Posgrado: Memoria, UNAM, Mexico, 2012.

- J. M. Castro-Manzano. *Revisión de intenciones*. En H.H. Hernández Bringas y M. Quero (coord.), Primer Congreso de Alumnos de Posgrado: Memoria, UNAM, México, 2011.

- J. M. Castro-Manzano. *Intenciones, revisión y no-monotonicidad*. En Coloquio del Posgrado en Filosofía de la Ciencia "Avances contemporáneos en filosofía de la ciencia", Universidad Nacional Autónoma de México, México. Octubre 2011.

- J. M. Castro-Manzano. *Towards a grounded theory of intention revision*. En Simposio Latinoamericano de Computación y Filosofía LACAP. Instituto de Investigaciones en Matemáticas Aplicadas y Sistemas. Universidad Nacional Autónoma de México, México. Noviembre 2009.

[6] El LACAP (por *Latin America Computer and Philosophy*) es un foro afiliado a la IACAP (*International Association of Computing And Philosophy*). Su objetivo es reunir investigadores para discutir el giro computacional que está ocurriendo en la interacción de la filosofía y las ciencias de la computación.

Otros productos

Gracias al alcance y visibilidad de estos resultados fuimos invitados a ser árbitros del artículo:

- X. Fan. "Decision As Choice of Potential Intentions". *Web Intelligence and Agent Systems: An International Journal* 5 (2013) 1–5 IOS Press.

sometido para publicación en la revista *Web Intelligence and Agent Systems (WIAS): An International Journal*. El artículo ya ha sido publicado.[7]

Por último, tenemos el agrado de reportar que hemos conseguido los permisos de Michael Bratman y *CSLI Publications*[8] para traducir al español y publicar su obra principal:

- *Intentions, plans, and practical reason*.

como *Intenciones, planes y razón práctica*. La traducción, así como las gestiones para su publicación, están en proceso.

[7] *Web Intelligence and Agent Systems (WIAS): An International Journal* es la revista oficial del *Web Intelligence Consortium*, una organización internacional dedicada a promover la investigación científica. La revista busca alcanzar un balance entre tecnologías Web y tecnologías de agente. Sus objetivos giran en torno a la comprensión de los fundamentos computacionales, lógicos, cognitivos, físicos y sociales de la agencia. Es una revista arbitrada por pares publicada por IOS Press (ISSN: 1875-9289 (en línea), 1570-1263 (en papel)) con índice H de 14 y que se encuentra indexada en ACM Computing Reviews, ACM Guide to Computing Literature, Compendex Plus, SCOPUS, EBSCO's database, MasterFILE, INSPEC database, Zentralblatt MATH, CompuScience, Computer Abstracts, Computer & Communications Security Abstracts database.

[8] *CSLI Publications* reporta nuevos desarrollos en los estudios del lenguaje, la información, la lógica y la computación. Es la casa editorial del *Center for the Study of Language and Information* (CSLI) de la Stanford University.

Glosario

Aspecto formal Se dice de las propiedades metalógicas de un formalismo lógico, 25, 27, 28, 175, 179

Aspecto material Se dice de las capacidades de representación de un formalismo lógico, 25, 27, 28, 37, 171, 175, 178, 179, 184

Especificación Intersección entre las ciencias de la computación y la filosofía que se dedica a la caracterización de la agencia a través de arquitecturas abstractas, 28, 34–37, 44, 46, 50, 71, 72, 91, 106, 109, 127, 131, 132, 138, 139, 141, 165, 167–169

Lógica BDI Sistema lógico cuyos operadores principales son los operadores BDI, 27, 33, 35, 36, 72, 95–98, 105, 107, 109, 127, 131, 132, 188

Lógicas derrotables Sistemas lógicos no-estándar que no garantizan que los consecuentes se den siempre que se den los antecedentes, 25, 27, 28, 91, 133, 171

Modelo BDI Modelo cognitivo cuyos elementos son las creencias, los deseos y las intenciones (por *Beliefs*, *Desires* e *Intentions*), 22, 23, 26, 27, 29, 31–33, 35, 37, 41, 44, 45, 47, 50, 55, 58, 60, 62, 65, 67, 68, 71, 74, 75, 79, 82, 87, 89–91, 106, 107, 109, 119, 128, 141, 152, 153, 156, 162, 163, 165, 170, 174, 175, 177, 178, 183–188

Modelo bratmaniano Modelo BDI de agencia racional que *i)* sigue las líneas generales de la teoría de razonamiento práctico de Bratman, *ii)* usa la arquitectura BDI para representar estructuras de datos y *iii)* configura una noción de consecuencia lógica a partir de las relaciones entre estados intencionales, 25, 26, 28, 29, 35–37, 67, 68, 80, 83, 86, 90–92, 105–107, 119, 128, 131, 141, 152, 153, 162, 163, 165, 170, 171, 174, 175, 177–179, 181, 183–188

No-monotonicidad Propiedad de la relación de consecuencia que indica que es posible desplazar conclusiones previas a partir de información nueva; en otras palabras, que el poder inferencial puede decrecer con la aparición de nueva información, 22, 25–27, 29, 32, 35, 37, 67, 70, 71, 82, 83, 86, 89–91, 107, 128, 141, 151, 152, 163, 170, 171, 173–175, 177–179, 183–185, 188

Problema de persistencia Problema cuya meta es generar un modelo de las condiciones bajo las cuales las intenciones se mantienen o se modifican durante el tiempo, 24, 25, 29, 31, 187

Razonamiento intencional Razonamiento que usa creencias e intenciones, 22, 23, 25–31, 34, 35, 37, 41, 67, 72, 75, 80–82, 86, 89, 90, 106, 107, 128, 141, 145, 152, 163, 170, 171, 175, 177–179, 182–187

Revisión Proceso de cambio de estructuras de datos en el que nueva información inconsistente con un conjunto C de proposiciones conlleva al cambio de algunas de las proposiciones originarias de C, 22, 26–30, 32, 33, 35, 37, 67, 71, 83–90, 104, 107, 128, 132, 133, 140, 141, 145, 148, 151–157, 159–163, 165, 167–171, 177, 178, 183, 185, 186, 188

Nomenclatura

$+$ Adición de metas

$-$ Borrado de metas

$?$ Meta para probar (*Test goal*)

α_i Acciones

$\Box$ Siempre

$\mathsf{BEL}_{\langle ag,C\rangle}$ Creer en una configuración agente

$\mathsf{DES}_{\langle ag,C\rangle}$ Desear en una configuración agente

$\mathsf{INT}_{\langle ag,C\rangle}$ Intentar en una configuración agente

BEL Creer

$\bigcup$ Gran unión

$\bot$ *Falsum*

$\cap$ Intersección de conjuntos

$\circ$ Composición de funciones

$\cup$ Unión de conjuntos

DES Desear

$\Diamond$ En algún momento (*Eventually*)

$\exists$ Cuantificador existencial

$\forall$ Cuantificador universal

$\Gamma, \Delta \ldots$ Conjunto de términos

$\leftarrow$ Asignación de valores

A Inevitable

INT Intentar

$\Leftrightarrow$ Equivalencia

$\longrightarrow$ Mapeo, función

$\models, \nmodels$ Relación de consecuencia semántica fuerte

$\mathrel{|\!\sim}, \mathrel{\not|\!\sim}$ Relación de inferencia débil

$\mathrel{|\!\approx}, \mathrel{\not|\!\approx}$ Relación de consecuencia semántica débil

$\neg$ Negación

$\bigcirc$ Siguiente

$\odot$ Función de revisión

$\ominus$ Función de contracción

v Meta-disyuncion

$\oplus$ Función de expansión

E Opcional

$\phi, \psi \ldots$ Variables proposicionales

$\Rightarrow$ Implicación

$\subseteq$ Inclusión de conjuntos

$\succ$ Relación de orden

$\times$ Producto cartesiano

$\top$ *Verum*

U	Hasta
$\vdash, \dashv$	Relación de inferencia fuerte
$\vee$	Disyunción
$\wedge$	Conjunción
$\wp$	Conjunto potencia
a_i	Fórmulas atómicas de primer orden para denotar acción $A(t_1, \ldots, t_n)$
ag	Símbolo de agente
at_i	Fórmulas atómicas de primer orden para denotar predicación $P(t_1, \ldots, t_n)$
b_i	Creencias
bs	Base de creencias
c_i	Caminos
Cn	Relación de consecuencia lógica
ctx_i	Contextos
e_i	Eventos
g_i	Metas
h_i	Cuerpos
p_i	Planes
ps	Base de planes
s_i	Estados
t_i	Términos
te_i	Eventos disparadores
u_i	Actualizaciones
w_i	Mundos posibles