El razonamiento defectible y sus fronteras epistémico-metodológicas[1]

Luis A. PÉREZ-MIRANDA
(Universidad del País Vasco)

Resumen: En este artículo hacemos un estudio analítico de la teoría del razonamiento defectible (o no monótono) atendiendo tanto a cuestiones de orden epistémico como metodológico. El razonamiento defectible es un tipo de razonamiento que está basado sobre la construcción de argumentos no lineales. Esto es, a lo largo de un proceso de razonamiento la introducción de nuevos supuestos en los argumentos puede llegar a retractar las conclusiones inferidas. Además, el razonamiento del sentido común que trabaja con deducciones es bidireccional, en el sentido que no sólo opera desde las premisas a la conclusión (hacia adelante), sino también desde la conclusión a las premisas (hacia atrás). Los sistemas de razonamiento que combinan ambos tipos de razonamiento se dice que son dirigidos a intereses. Que un agente esté interesado en conocer el status de una proposición determinada exige que éste sea capaz de discernir entre razonamiento teórico (razonamiento acerca de qué creencias poseer) y razonamiento práctico (razonamiento acerca de qué acciones llevar a

[1] Mis agradecimientos a los miembros del ILCLI (Institute for Logic, Cognition, Language, and Information, Universidad del País Vasco) por los comentarios recibidos. Por otra parte, hay señalar que este trabajo sólo ha podido realizarse gracias a una beca post-doctoral del Departamento de Educación, Universidades e Investigación (Gobierno Vasco/Eusko Jarlaritza).

cabo). Una cuestión en la modelización del razonamiento ordinario es establecer una distinción entre teorías de la justificación y teorías de la garantía. Aquellas proposiciones que estén soportadas por algún argumento no derrotado en algún estadio de la computación para una situación epistémica concreta, se dirá que están justificadas, sin estar necesariamente garantizadas. Las proposiciones garantizadas son aquellas que deberían estar justificadas 'a largo plazo', en el caso de que el sistema de razonamiento fuera capaz de llevar a cabo todo el razonamiento relevante posible. Sin embargo, veremos que una vez que una proposición resulta estar justificada, es razonable aceptarla provisionalmente y actuar sobre ella.

1. Introducción

Hay muchas situaciones en las que es apropiado para sistemas inteligentes aumentar sus creencias con otras nuevas que no se siguen *"lógicamente"* a partir de las que son explícitas. La presión de los acontecimientos nos fuerza en ocasiones a tomar decisiones y posteriormente a actuar sobre esas decisiones con anterioridad a que todos los hechos relevantes estén a nuestra disposición. Quizás sería útil para este tipo de sistemas que fueran capaces de asumir que las creencias que poseen en una situación epistémica determinada acerca de una cuestión fueran todas las creencias importantes acerca de esa cuestión. Incluso en las situaciones en las que los elementos pragmáticos juegan un considerable papel en la comunicación, debemos asumir o llevar a cabo inferencias que ascribimos al hablante sin estar del todo seguros de que coinciden con sus pretensiones reales. En estos casos asumimos por defecto un determinado conjunto de posibles interpretaciones.

Además, cualquier intento de recoger todo el conocimiento acerca del mundo real mediante un conjunto finito de enunciados es fundamentalmente imposible. A medida que aumenta nuestro conocimiento, nuestra conceptualización de un área específica cambia. Cualquier conceptualización, propuesta para cierto propósito, está sujeta a revisión. Pensemos en el siguiente enunciado acerca de las panteras: "*todas las panteras son negras*". Este enunciado podría ser utilizado para ciertos propósitos limitados, pero si intentamos aplicarlo más generalmente, podríamos vernos enfrentados al hecho de que supuestamente algunas panteras de zoológico (llamémoslas zoo-panteras), aun siendo panteras, sin embargo, no son negras. Incluso el enunciado

$\forall x$ Pantera (x) $\wedge$ $\neg$Zoo-pantera (x) $\supset$ Negra (x) no recoge adecuadamente la situación del mundo real, puesto que podemos imaginar otros tipos de panteras, albinas, por ejemplo, que no son negras. La lista de una *calificación* semejante es demasiado larga si no interminable, llevándonos, quizás, a la desesperación cuando usamos un lenguaje para la representación del conocimiento. Este problema a menudo es denominado *problema de la calificación*. La mayor parte de los enunciados cuantificados universalmente tendrán que incluir un número infinito de calificaciones si deben ser interpretados como enunciados correctos acerca del mundo. Pensemos, por ejemplo, en la definición del significado del concepto *pantera*. Si bien resulta imposible dar una definición de su significado, no parece inviable dar con una identificación del mismo. De hecho, en nuestra vida diaria, los humanos utilizamos enunciados que asumimos como verdaderos. Así pues, lo que parece que necesitaríamos sería algún tipo de *reglas inferenciales* que nos justificaran la adopción de creencias sobre la base de las cuales poder actuar temporalmente, asunciones que serían posteriormente revisadas según las calificaciones adicionales se hicieran importantes.

Es así que han aparecido nuevas metodologías para el análisis del razonamiento ordinario que han desembocado en la utilización de lógicas no-monótonas. El razonamiento del sentido común es una empresa compleja que no sólo lleva a la adopción de nuevas creencias a partir de las ya disponibles mediante algún proceso de transformación, sino que, en ocasiones, también nos conduce a la retractación de creencias previas en función de la nueva información recibida. En este último sentido decimos que el razonamiento es *defectible* o no monótono. En IA (Inteligencia Artificial) se han venido desarrollando un número considerable de técnicas basadas en la lógica para conseguir las consecuencias no monótonas deseadas. Entre las más significativas podríamos destacar las siguientes: *asunción del mundo cerrado* (Reiter, 1978); *circunscripción* (McCarthy, 1980); *razonamiento por defecto* (Reiter, 1980); *lógica autoepistémica* (Moore, 1985); *razonamiento defectible* (Pollock, 1991, 92). Sin embargo, si a la hora de construir teorías del razonamiento tenemos en cuenta ciertos criterios tanto cognitivo-computacionales (concesiones mutuas entre tratabilidad y expresividad) como epistémicos (adecuación de las teorías), ciertamente la teoría del razonamiento defectible parece ser la más adecuada. En este sentido, pretendemos mostrar que se está produciendo un giro importante en las investigaciones de procesos cognitivos como el del razonamiento. No sólo se hacen análisis conceptuales de los problemas, sino que además se buscan formalizaciones que recojan adecuada-

mente dichos procesos y resulten efectivas con vistas a su posterior implementación. Se puede afirmar que, a partir de las investigaciones especialmente en IA y Ciencia Cognitiva, se ha producido un giro copernicano en el tratamiento de problemas como el del razonamiento ordinario —la construcción de razonadores automatizados y planificadores es un ejemplo de ello.

Paralelamente, si estimamos necesaria la separación entre razonamiento teórico (razonamiento acerca de qué creencias tener) y razonamiento práctico (razonamiento acerca de qué acciones llevar a cabo), deberemos poner en entredicho el denominado *razonamiento crédulo* que ha primado en IA. Generalmente en los formalismos no monótonos se asume cierta información (que completa la teoría) que no entra en contradicción con la previamente disponible en la base de datos. No obstante, el razonamiento humano no parece funcionar de ese modo, sino que es constantemente sensible a la introducción de nuevos supuestos en los argumentos (*razonamiento suposicional*), pudiendo éstos llegar a alterar la información en cada una de las líneas de las cadenas deductivas que definen esos argumentos. Es más, la adecuación de la información disponible por el agente al contexto en el que éste se desenvuelve implica una revisión continuada de esa información, y, en consecuencia, de las conclusiones argumentadas.

Otra cuestión importante a destacar es que el razonamiento del sentido común que trabaja con deducciones es bidireccional, en el sentido de que no va sólo de las premisas a la conclusión (*razonamiento hacia adelante*), sino que también lo hace desde la conclusión a las premisas. Es decir, de ciertos intereses a otros intereses; construyendo una cadena de deducciones que nos permita inferir la conclusión de partida (*razonamiento hacia atrás*). Un sistema de razonamiento que combina ambos tipos de razonamientos se dice que es *dirigido a intereses*.

En los siguientes apartados nos proponemos desarrollar las ideas más genuinas del razonamiento defectible, con el objetivo de mostrar que algunas de las técnicas utilizadas por los formalismos no monótonos habituales deberían ser reformuladas, con vistas a construir una teoría que recoja algunas de las características del razonamiento del sentido común. Asimismo, indicamos la metodología a seguir en la construcción de un razonador defectible automatizado. Se trata también de poner de manifiesto la necesidad de extender estos formalismos para el caso del razonamiento práctico.

2. Razones y Derrotadores de Argumentos

Para la teoría del razonamiento defectible el razonamiento procede mediante la construcción de argumentos, donde las *razones* proporcionan los lazos atómicos para esos argumentos (Pollock, 1987). Algunas de las razones son *conclusivas*, es decir, implicar lógicamente su conclusión. Sin embargo, la experiencia de usar razones de este tipo en epistemología indica que en la práctica no todas las razones son conclusivas, a diferencia de lo que ocurre en el razonamiento deductivo de la lógica formal clásica (Pollock, 1986). Las denominadas razones *prima facie* crean una presunción en favor de su conclusión, pero en estos casos siempre cabe la posibilidad de su retractación con la incorporación de nueva información en el sistema de razonamiento. Este último tipo de razones no conclusivas soporta a sus conclusiones manteniendo presente la idea de *defectibilidad*. Una buena razón teórica soporta, aunque no necesita implicar, la racionalidad de creer la proposición (o proposiciones) para la cual es una buena razón; una razón práctica soporta, aunque no necesita implicar, la racionalidad de la acción (o acciones) para la que es una buena razón (Cf. Audi, 1991). Veamos, pues, más en detalle cuáles son las diferencias entre ambos tipos de razones:

Una razón *prima facie* se define como el par ordenado $\langle \Gamma, p \rangle$ donde Γ es el conjunto finito de premisas y p es la conclusión. Cuando unas razones entran en conflicto con otras razones poniendo en entredicho las conclusiones soportadas por esas otras razones, decimos que las primeras actúan como derrotadores de las segundas. El tipo más simple de derrotador para una razón *prima facie* $\langle \Gamma, p \rangle$ es una razón para negar la conclusión. Si $\langle \Gamma, p \rangle$ es una razón *prima facie*, $\langle \Lambda, q \rangle$ es un tipo de derrotador que denominamos *rebatidor* para $\langle \Gamma, p \rangle$ si y sólo si $\langle \Lambda, q \rangle$ es una razón y $q = \neg p$[2]. Aunque no es un punto tratado en este artículo, para que una razón *prima facie* derrote de modo *efectivo* a otra razón *prima facie*, la *fuerza* de la primera tiene que ser mayor que la fuerza de la segunda[3].

Los derrotadores que no son del tipo *rebatidor* atacan a una razón *prima facie* sin atacar a su conclusión. A éstos se les denomina saldadores. Los derrotadores de este tipo ponen en entredicho la conexión entre las premisas

[2] Las razones para las que los únicos derrotadores son derrotadores del tipo "rebatidor" son análogas a los defectos normales de las lógicas con defectos de Reiter.

[3] Un trabajo que clarifica la necesidad de introducir la noción de fuerza de un argumento en los estudios del razonamiento ordinario lo encontramos en Pollock (1991).

y la conclusión. Sirvámonos de un ejemplo: *'x parece rojo'* es una razón *prima facie* para *'x es rojo'*. Pero si sabemos no sólo que *'x parece rojo'*, sino también que x está siendo iluminado por las luces rojas de un automóvil, y las luces rojas de un automóvil pueden hacer parecer a las cosas de color rojo cuando de hecho no lo son, entonces resulta irracional inferir que x es de color rojo. En consecuencia, *'x es iluminado por las luces rojas y las luces rojas pueden hacer aparecer a las cosas rojas cuando de hecho no lo son'* es un derrotador, pero no es una razón para pensar que x no es rojo, por lo que no es un derrotador del tipo *rebatidor*. Sin embargo, ataca a la conexión entre *'x parece de color rojo'* y *'x es rojo'*, dándonos una razón para dudar de la proposición *'x no debería parecer rojo al menos que fuera rojo'* (o lo que es lo mismo, una razón para creer que x puede parecer rojo cuando de hecho no lo es). No obstante, en los formalismos no monótonos usuales no se hace uso de derrotadores del tipo *saldador*. En consecuencia, gran parte de las relaciones existentes entre los argumentos del razonamiento defectible quedan sin recoger en los enfoques computacionales clásicos.

El tipo de derrotador que hemos denominado *saldador* se define formalmente del modo siguiente: 'P no debería ser verdadero al menos que Q lo fuera' es un tipo de condicional, que simbolizamos P»Q. Esto quiere decir que $\langle \Gamma, p \rangle$ es una razón *prima facie*, entonces cualquier razón para denegar $\lceil \Pi\Gamma, p \rceil$, donde $\Pi\Gamma$ representa la conjunción de las premisas de Γ, es un derrotador del tipo *saldador*. Si $\langle \Gamma, p \rangle$ es una razón *prima facie*, $\langle \Lambda, q \rangle$ es un derrotador del tipo *saldador* para $\langle \Gamma, p \rangle$ si y sólo si $\langle \Lambda, q \rangle$ es una razón y $q = \lceil \neg(\Pi\Gamma \text{»} p) \rceil$. La distinción entre estos dos tipos de derrotadores de argumentos nos pone claramente de manifiesto que nos encontramos frente a una teoría formal del razonamiento no monótono bastante distanciada de los enfoques usuales en IA.

3. Estructura de un Argumento en el Razonamiento Defectible

En este apartado se define el concepto de *argumento* en el ámbito del razonamiento defectible. En la literatura lógico-filosófica el concepto de argumento tiene en ocasiones un carácter más complejo que el aquí desarrollado. Por ejemplo, recuérdese la definición peirciana de argumento según la cual éste queda dividido en tres clases: inductivo, abductivo, y deductivo (Cf. Peirce, 1987). Hay que tener en cuenta que la línea divisoria entre el razonamiento deductivo de la lógica formal clásica y el razonamiento defectible no

sólo viene dada por la utilización de unas reglas o esquemas de inferencia específicos, además de las habituales, sino también por la propia idea de la retractabilidad de las inferencias alcanzadas.

En el contexto del razonamiento defectible, un razonamiento comienza con premisas que constituyen la entrada de información en el razonador. El sistema entonces lleva a cabo inferencias tanto conclusivas como defectibles a partir de la información recogida en las premisas sirviéndose de *esquemas* de inferencia. Las razones se combinan de varias maneras dando lugar a los argumentos. Podría decirse que el razonamiento consiste en la construcción de argumentos de tal modo que podamos adoptar alguna actitud cognitiva (de creencia, por ejemplo) hacia una proposición dada, en la que podríamos estar interesados por alguna cuestión práctica. En este caso, se trataría de construir un argumento en el que la proposición que estamos interesados en demostrar no se viera afectada por ningún derrotador. Así pues, cuando hablamos de sistemas de razonamiento nos referimos a razonadores *dirigidos a intereses* donde las proposiciones a demostrar se establecen previamente al nivel del razonamiento práctico.

El razonamiento defectible se lleva a cabo tanto mediante la construcción de argumentos *lineales* como *suposicionales*. Los argumentos *lineales*, que son los más simples, se construyen mediante la derivación continuada de conclusiones a partir de creencias previas que constituyen las razones para las conclusiones. O si se prefiere, estos argumentos constituyen secuencias finitas de proposiciones cada una de las cuales es o bien un miembro de la entrada (información recibida en forma de *inputs*) o inferible a partir de miembros previos de la secuencia de acuerdo con algún esquema de razonamiento. Enfoques que focalizan su atención en argumentos lineales podemos encontrarlos en los trabajos de Loui, 1987, y Lin & Shoham, 1990, entre otros. Los argumentos lineales pueden ser vistos como argumentos en los cuales el conjunto de suposiciones es siempre vacío. Sin embargo, el razonamiento deductivo puede llevar a conocimiento *a priori* de 'verdades de razón'. O si se prefiere, a la utilización de algunas de las verdades lógicas. Por ejemplo, podemos obtener conclusiones como $[(p\&q) \supset] p$ o $(p \vee \neg p)$ que no dependen de las premisas. A diferencia de lo que ocurre con los argumentos de carácter lineal, se emplean reglas que permiten introducir "argumentos subsidiarios" dentro del argumento principal y derivar conclusiones dentro del argumento principal que están en relación con las conclusiones derivadas en los argumentos subsidiarios. Lo que hace esto posible es el *razonamiento suposicional*. En éste 'suponemos' algo (que es el caso) que no hemos inferido a par-

tir de alguna de las premisas de partida, extraemos conclusiones a partir de la suposición, y entonces *'descargamos'* la suposición para obtener una conclusión que ya no depende por más tiempo de la suposición. Dos ejemplos característicos del razonamiento suposicional son la *condicionalización* y la *reducción al absurdo*[4].

En el razonamiento suposicional, no podemos pensar en los argumentos como secuencias finitas de proposiciones, puesto que cada línea de un argumento puede depender a su vez de otros supuestos. Las líneas de los argumentos pueden definirse como triplos ordenados ⟨X,p,β⟩ donde X es el conjunto de proposiciones que comprende lo que está supuesto en relación a esa línea, p es la proposición obtenida en esa línea, y β describe la *base* para la línea. β será tomada como un par ordenado ⟨λ, R⟩ donde R es la regla de inferencia usada para obtener la línea y λ indica el número de las líneas a partir de las cuales la línea presente es inferida mediante R. La condicionalización arriba mencionada funciona como una 'regla de descarga' que nos permite manipular los conjuntos de suposiciones. Esta podría ser formulada del modo siguiente: A partir de las líneas de argumento ⟨X U {p}, q, β⟩, inferimos ⟨⟨X, (p⊃q), ⟨{i}, *condicionalización*⟩⟩. La condicionalización es pues aquí entendida como una regla de inferencia que aplicada sobre {i} nos permite obtener la proposición (p⊃q). Por ejemplo, si tenemos como punto de partida la proposición p y estamos interesados en probar ((p&q)∨(p&¬q)), podríamos construir un argumento basado sobre condicionalización cuya última línea sostuviera la proposición en cuestión. Estos son los pasos a seguir: (1) Suponemos la negación de una de las partes de la disyunción, por ejemplo, ¬(p&q); (2) por De Morgan, a partir de 1, obtenemos (¬p∨¬q); (3) tomamos la proposición de partida p; (4) obtenemos ¬q mediante la aplicación del silogismo disyuntivo en las líneas 3 y 2; (5) obtenemos (p&¬q) mediante adjunción de 3 y 4; (6) es este punto aplicamos *condicionalización* a partir de 5 lo que nos permite obtener en esta línea del argumento (¬(p&q)⊃(p&¬q); (7) finalmente, sirviéndonos de la regla disyuntiva sobre 6 llegamos a la conclusión deseada ((p&q)∨(p&¬q)).

Decimos que un argumento σ *soporta* a la proposición p relativa a la suposición X si y sólo si una de sus líneas tiene la forma ⟨X,p,β⟩. σ *soporta* a p si y sólo si σ soporta a p relativa a la suposición vacía. La conclusión de un argumento es la última línea. Los argumentos derrotan a otros argumen-

[4] Véase Pollock (1990).

tos soportando tanto derrotadores del tipo *rebatidor* como *saldador* para alguno de sus pasos defectibles[5]. Cabe destacar que, a diferencia de la mayor parte del trabajo sobre lógica no monótona en IA que emplea la semántica modelista, la teoría del Razonamiento Defectible emplea una semántica basada en argumentos.

4. Justificación y Garantía

Uno de los problemas a los que nos enfrentamos a la hora de diseñar un razonador defectible es cómo evaluar el razonamiento que el sistema realiza. Para responder a la pregunta de qué es lo que entendemos por un sistema que lleva a cabo razonamiento 'correcto', resulta útil distinguir entre teorías del razonamiento y teorías de la garantía (Pollock, 1991). Las teorías del razonamiento son básicamente teorías procedimentales. Están comprometidas con lo que un razonador debería llevar a cabo de inmediato cuando éste se encuentra en una situación epistémica determinada. En cada estadio del razonamiento, si el razonamiento es correcto, entonces una creencia sostenida sobre la base de ese razonamiento está justificada, incluso si cabe la posibilidad de que un razonamiento posterior obligue a su retractación. La denominada *justificación epistémica* es una noción procedimental que está compuesta por reglas correctas para cambios de creencia que han sido seguidas por el sistema hasta el momento presente en conexión con la creencia que está siendo evaluada.

En contraposición, la garantía es lo que el sistema de razonamiento está en última instancia tratando de alcanzar. Una proposición está garantizada en una situación epistémica particular sí y sólo si, partiendo de esa situación epistémica, un razonador ideal (sin restricciones de recursos espacio-temporales), se ve en último término llevado a adoptarla como creencia. Las proposiciones garantizadas son aquellas que deberían estar justificadas *a largo*

[5] Si hacemos la simplificación de que las fuerzas de los argumentos son equivalentes, esta idea puede ser descrita del siguiente modo: un argumento σ rebate (utilizando un derrotador del tipo *rebatidor*) a un argumento η si:

(1) la última línea de tiene la forma $<Y,q,<\alpha,$razón$>>$ donde las proposiciones sostenida sobre las líneas en constituyen una razón prima facie para q; y

(2) la última línea de σ tiene la forma $<X,\neg q,\beta>$.

Análogamente se sigue para derrotadores del tipo que hemos denominado *saldador*.

plazo, esto es, en el caso de que el sistema fuera capaz de llevar a cabo todo el razonamiento relevante posible[6].

La *perspectiva procedimental* propone analizar el razonamiento defectible dando una correcta descripción de cómo éste trabaja dejando las consideraciones semánticas en suspenso. Si entendemos el término *semántica* de modo que sólo incluya la semántica modelista, entonces puede darse el caso de teorías de la garantía que no sean semánticas. De hecho, podemos afirmar que la teoría del razonamiento defectible que estamos analizando es en su fundamento una teoría de la garantía y, no obstante, no presenta una semántica en el sentido de la teoría de modelos.

Antes de introducirnos seriamente en una investigación de orden semántico, deberíamos cerciorarnos de que la teoría descrita por la semántica es realista desde un punto de vista epistémico, y esto nos exige atender a las particularidades de cómo funciona realmente el razonamiento defectible. Por ejemplo, supongamos que sabemos que la mayoría de las panteras son negras. Ahora *supongamos* que sabemos que en una de las jaulas cubiertas del zoo que estamos visitando hay un felino de gran tamaño y que tratamos de identificarlo mediante los ruidos que produce. Es perfectamente plausible razonar diciendo que si es una pantera entonces es negra. Mediante condicionalización *suponemos* que el animal es una pantera, inferimos *defectiblemente* que es negra, y entonces descargamos la suposición para concluir que si es una pantera entonces es negra. Sin embargo, ningún tipo de sistema de IA no monótono puede acomodar este tipo de inferencias simples. Alguien podría argumentar que el esquema del razonamiento utilizado se corresponde simplemente con el del modus ponens en el razonamiento deductivo de la lógica formal clásica, pero lo interesante en el ejemplo no es la estructura en sí de esta regla de inferencia, sino más bien el modo en el que hacemos uso de ella en el razonamiento suposicional. Hay que tener en cuenta que este tipo de conclusiones, en cuanto que forman parte de argumentos defectibles, siempre

[6] Nótese que, una proposición puede estar justificada sin estar garantizada, puesto que aunque el sistema puede haber inferido algo correcto hasta el momento presente, y haberle llevado esto a la adopción de la creencia, puede haber más razonamiento que espera ser realizado que podría, sin embargo, obligarnos a retractar la proposición previamente adoptada como creencia. Similarmente, una proposición puede estar garantizada sin estar justificada, puesto que aunque el razonamiento basado en la información presente puede haber fallado a la hora de encontrar razones adecuadas para adoptar la conclusión, razonamiento posterior puede proporcionar tales razones. De igual modo, razonamiento basado en la información presente puede imponer la adopción de retractadores o derrotadores, los cuales pueden a su vez ser retractados por razonamiento que sea realizado más adelante.

están sujetas a revisión. Esto no impide, sin embargo, construir argumentos de carácter suposicional, que tendrán como mínimo la misma provisionalidad que la información sobre la que descansan.

La interrelación entre argumentos es una cuestión fundamental en la teoría del razonamiento defectible. Una caracterización de lo que debería ser creído dados todos los argumentos relevantes es una caracterización del conjunto de proposiciones garantizadas. Dar cuenta de ello es relativamente sencillo si tomamos como primitiva la noción de retractación entre argumentos. Supongamos que tenemos un argumento α que soporta una conclusión P, y un argumento β que derrota a α. Si éstos son los únicos argumentos relevantes, entonces P no está garantizada. Pero ahora supongamos que adquirimos un tercer argumento γ que derrota a β. La adición de γ debería tener el efecto de restaurar a α, dejando a P garantizada.

Pollock, (1991a) recoge esta idea de interrelación entre los argumentos mediante la siguiente definición:

> Todos los argumentos son argumentos de nivel 0.
> Un argumento es un argumento de nivel n+1 si y sólo si es un argumento de nivel 0 y no es derrotado por ningún argumento de nivel n.
> Un argumento está *dentro* en el nivel n si y sólo si es un argumento de nivel n.
> En otro caso está *fuera*.

Consideremos el caso de que tenemos un argumento α que soporta la conclusión "el objeto es de color rojo", y otro argumento β que derrota a α. En este caso, diremos que la conclusión no está garantizada. Obviamente, para que el argumento fuera nuevamente reestablecido y, por tanto, quedara garantizado nos bastaría con encontrar un argumento que a su vez derrotara a β y no fuera derrotado por ningún otro argumento. En realidad, por la propia naturaleza del razonamiento ordinario, sólo tiene sentido hablar de argumentos últimamente no derrotados para el caso de situaciones epistémicas específicas. Un argumento está *últimamente no derrotado* si y sólo si hay un m tal que para cada $n \geq m$, el argumento está dentro en el nivel n. Así, decimos que una proposición está garantizada si y sólo si es soportada por algún argumento últimamente no derrotado.

Las limitaciones de recursos computacionales obligan a centrarse más en la justificación que en la garantía. Esto nos lleva a la cuestión de cómo construir un sistema procedimental que razone correctamente salvando esas limitaciones. Si pretendemos que el sistema implemente un procedimiento efec-

tivo para determinar la garantía, entonces se trata de una tarea imposible de realizar. Por la tesis de Church, el conjunto de teoremas del cálculo de predicados no es decidible. De este modo, ningún sistema puede computar la garantía en el sentido señalado. El conjunto de fórmulas no válidas del cálculo de predicados no es recursivamente enumerable (r.e.). De aquí se sigue que, por ejemplo, en la lógica con defectos (*default logic*), una teoría de primer orden que incluya *defectos normales* puede tener un conjunto de teoremas que no es recursivamente enumerable, y por tanto no puede haber un procedimiento efectivo para generar ese conjunto de teoremas. La misma conclusión se aplica a todas las teorías del razonamiento defectible y a todas las lógicas no monótonas (con la excepción de algunos fragmentos)[7].

Si por motivos obvios la exigencia que se le pide a un sistema de razonamiento automático no es la computación de la garantía, lo que sí podemos demandarle es que el conjunto de creencias se aproxime al conjunto de proposiciones en el límite. El *desideratum* preciso para un razonador automático es que la justificación se aproxime a la garantía del modo siguiente:

> Las reglas para el razonamiento deberían ser tales que:
> (1) si una proposición p está garantizada, entonces el sistema alcanzará eventualmente un punto donde p es adoptada y permanece adoptada.
> (2) si p no está garantizada entonces el sistema alcanzará eventualmente un punto donde p no es adoptada y permanece inadoptada.

En otras palabras, la tarea del razonador no es computar la garantía, sino generar conjuntos sucesivos de creencias que se aproximen a la garantía cada vez más estrechamente. Para llevar a cabo esa tarea se hace imprescindible la introducción del concepto de *conjunto defectiblemente enumerable*[8]. La diferencia intuitiva entre conjuntos defectiblemente enumerables y conjuntos recursivamente enumerables es que estos últimos pueden ir aproximándose sistemáticamente desde abajo, mientras que los defectiblemente enumerables pueden hacerlo simultáneamente tanto desde arriba como desde abajo (Cf.

[7] Algunos problemas computacionales relacionados con esta cuestión y posibles soluciones a los mismos son analizados en Levesque (1993).

[8] Un conjunto A es *defectiblemente enumerable* si y sólo si existe una función f efectivamente computable y un conjunto recursivo A_0 tal que si definimos $A_i+1=f(A_i)$ entonces: (1) $(\forall x)$ si $x\in A$ entonces $(\exists n)$ $(\forall m>n)$ $x\in A_m$; (2) $(\forall x)$ si $x\notin A$ entonces $(\exists n)$ $(\forall m>n)$ $x\notin A_m$. Decimos que el par (A_0,f) es una aproximación d.e. (defectiblemente enumerable) de A.

Pollock, 1991)[9]. Esto quiere decir que si A es defectiblemente enumerable los conjuntos A_i no necesitan ser subconjuntos de A. Estos pueden de hecho aproximarse a A simultáneamente en ambas direcciones, en el sentido de que pueden contener elementos que no están contenidos en A. Si bien dichos elementos irán paulatinamente desapareciendo de los conjuntos A_i conforme nos aproximemos al conjunto A, no es necesario que exista un punto a partir del cual esos elementos hayan desaparecido del todo.

El propósito subyacente al razonamiento y a la garantía es por tanto que el conjunto de proposiciones garantizadas sea defectiblemente enumerable, y las reglas para el razonamiento sean reglas que vayan sucesivamente aproximándose a la garantía de la manera descrita, es decir, que sean reglas para construir una aproximación defectiblemente enumerable. Una consecuencia importante de este propósito es que no podemos esperar que un sistema de razonamiento automatizado se detenga en todos los casos en los que analizamos el status de la conclusión de un argumento. Este razonador así constituido puede informarnos que hasta tal punto una cierta conclusión está justificada, pero puede tener que continuar indeterminadamente en una búsqueda posiblemente sin resultados para argumentos de carácter defectible.

Sin embargo, una vez que una conclusión está justificada, es razonable aceptarla provisionalmente y actuar sobre ella. Una confusión común en las teorías de IA del razonamiento defectible ha sido que antes de que sea razonable actuar sobre la conclusión de algún razonamiento defectible, debe ser establecido que no haya retractadores o derrotadores verdaderos — esto es lo que definimos en la introducción como *razonamiento crédulo*. Si atendemos a la escasez de recursos (memorísticos, inferenciales, espacio-temporales) con la que tenemos que vernos los agentes racionales-conductuales cuando manipulamos conocimiento, debería sernos suficiente que el conocimiento sobre el que posiblemente vayamos a actuar esté racionalmente justificado para esas circunstancias, con independencia de que hayamos optimizado al máximo o no la información.

Indirectamente lo que aquí se está poniendo de manifiesto es la necesidad de establecer una interrelación más estrecha entre el razonamiento teóri-

[9] Si A es recursivamente enumerable (r.e.), entonces existe una secuencia efectivamente computable de conjuntos A_i mediante f tal que (1) $(\forall x)$ si $x \in A$ entonces $(\exists n)(\forall m>n)\ x \in A_m$; (2) $(\forall x)$ si $x \notin A$ entonces $\forall m\ x \notin A_m$. Los conjuntos A_i se aproximan a A desde abajo en el sentido de que todos son subconjuntos de A y crecen monótonamente, aproximándose a A en el límite (Cf. Cutland, 1980).

co y el razonamiento práctico. Podría ser racional servirse de un proceso de razonamiento que, a pesar de resultar satisfactorio, no maximizara la utilidad esperada por el agente, pongamos por caso. Incluso si esta clase de procesos incluyen simplificaciones, heurísticos, o rutinas que violan los criterios normativos del modelo de la teoría de la decisión, el cálculo de probabilidades, etc., no es irracional basar el razonamiento sobre ese tipo de métodos de aproximación. Si entendemos el razonamiento como un proceso deliberativo dirigido a fines que dependen tanto de las actitudes proposicionales de los agentes como de sus recursos cognitivo-computacionales, entonces no resulta extraño que un razonador automatizado que trata al menos de modelizar el razonamiento humano se vea sometido a ciertas restricciones tanto epistémicas como computacionales.

En conexión con la idea de razonamiento defectible aquí defendida las más destacables son las de Nute & Lewis, 1986; Nute, 1990; Levesque, 1990; Ginsberg, 1989; Baker & Ginsberg, 1989; y Geffner, 1990). Los razonadores utilizados por estas teorías están basados sobre una amplia variedad de enfoques para el razonamiento defectible (circunscripción, lógica con defectos, etc.). A pesar de su variedad, parece que es posible encontrar un patrón común que nos permita compararlos entre sí. Esto es debido a que estas teorías han pretendido construir un razonador análogo al tradicional probador de teoremas. Estos razonadores de teoremas construyen conjuntos de conclusiones r. e. (recursivamente enumerables). Teniendo en cuenta que las extensiones de la lógica de primer orden subyacentes a aquellas teorías no son tan siquiera semidecidibles, este tipo de razonadores no monótonos usualmente se ha centrado en lógicas expresivamente muy débiles.

En contraposición, el razonador automatizado OSCAR (Pollock, 1995) al proporcionar una aproximación basada en el concepto de *defectibilidad enumerable dirigida a intereses (d.e.i.)* para la garantía, más que una enumeración recursiva de las conclusiones garantizadas, es aplicable para todo el cálculo de predicados, así como adecuado en los términos descritos para el razonamiento defectible[10]. No obstante, el modelo propuesto sigue siendo demasiado idealizado y no da cuenta de muchas de las situaciones empíricas interesantes que van más allá del razonamiento realizado sobre la base de lógicas de primer orden. Sin duda, se trata de un primer paso que debe completarse con la construcción de arquitecturas cognitivas que abarquen también proce-

[10] El razonamiento suposicional dirigido a intereses aparece desarrollado en Pollock (1990).

sos de razonamiento práctico y que incluyan, por ejemplo, pro-actitudes como las intenciones (Cf. Pérez Miranda, 1994). Dicho de otro modo, un estudio del razonamiento ordinario no debería reducirse a la evaluación o justificación de creencias mediante la construcción de argumentos, sino que también debería ocuparse de los procesos de decisión acerca de qué acciones debería llevar a cabo un agente racional atendiendo a sus necesidades, objetivos y planes (Cf. Pérez Miranda, 1996). Esto último implica hacer frente a la cuestión de las representaciones mentales del agente acerca del mundo y a la cuestión de las actitudes proposicionales.

Los criterios de carácter epistémico para la justificación de la creencia aquí desarrollados pueden ser aplicables para el caso de la selección de objetivos por parte de un agente. Por ejemplo, las razones que justifiquen la adopción de los objetivos de un agente que planifica deben ser razones no derrotadas por ninguna otra razón relevante en curso. Supongamos que 'x es deseado por el agente A' es una razón *prima facie* para 'la adopción de x como objetivo por parte de A', pero que en la práctica A sabe que 'el deseo de alcanzar x no guarda relación con los planes pretendidos por A'. Este último hecho puede ser considerado como una razón que ataca a la conexión entre el deseo del agente y la adopción del objetivo del agente sin necesidad de atacar directamente a la conclusión. Es así que consideramos el estado mental del agente no como una razón para la adopción de un objetivo, sino como una razón para desestimar cualesquiera otras razones para adoptar un objetivo que no sea, en este caso, el de seguir el plan pretendido. Las razones en cuestión no son estrictamente razones que se contradigan, sino que se trata más bien de un conflicto entre razones de primer y de segundo orden.

5. Defectibilidad Colectiva y Provisional y Argumentos Autodefectibles

La interacción entre razones de argumentos nos lleva a ocuparnos, en primer lugar, del problema de la defectibilidad colectiva y provisional y, en segundo lugar, del problema de los argumentos autodefectibles. Supongamos que tenemos dos amigos Mikel y Agustín, a los cuales les damos el mismo grado de confianza, Mikel se aproxima a ti y te dice, *'Han subido las tasas municipales'*. Agustín entonces afirma, *'No le creas. Siguen igual de precio'*. Sin más evidencia que la presente, ¿qué es lo que se debería creer? Es obvio que deberíamos retener nuestra creencia acerca de las tasas municipales hasta que obtuviéramos más información al respecto. Esto es una ilustración del

fenómeno de la *defectibilidad colectiva*. Consideremos un escenario en el que el conjunto de premisas es {p,q}, y ⟨{p}, r⟩ y ⟨{q}, ¬r⟩ son razones *prima facie* de la misma fuerza[11]. Entonces podemos construir dos argumentos sencillos:

α: p⇒¬r
β: q⇒r

Se sigue a partir de nuestro análisis que un argumento derrota o retracta al otro (se derrotan recíprocamente). De acuerdo con esto y utilizando la definición que recoge la interrelación entre argumentos dada en el apartado cuarto, están *dentro* en el nivel 0, *fuera* en el nivel 1, *dentro* nuevamente en el nivel 2, *fuera* en el nivel 3, y así sucesivamente. Por tanto, ninguno de ellos queda definitivamente no derrotado, y en consecuencia ni r ni ¬r están garantizados en esa situación epistémica.

La defectibilidad colectiva opera de acuerdo con el siguiente principio según el cual, Si Σ es un conjunto de argumentos tales que (1) cada argumento en Σ es derrotado o retractado por algún otro argumento, y (2) ningún argumento en Σ es derrotado por algún otro argumento que no esté en Σ, entonces ningún argumento en Σ es definitivamente derrotado. Esto es debido a que cada argumento en Σ estará *dentro* en cada nivel par, pero *fuera* en cada nivel impar. En función de lo establecido anteriormente podemos definir:

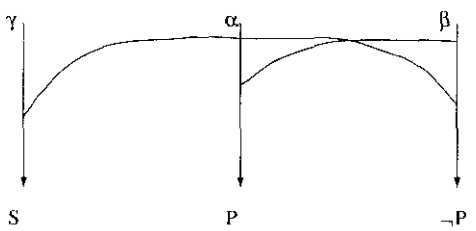
Un argumento σ es últimamente no derrotado si y sólo si hay un nivel ν tal que σ está *dentro* en todos los niveles más altos.
Un argumento σ es *provisionalmente derrotado* si y sólo si no hay un nivel ν tal que σ está dentro en todos los niveles más altos o fuera en todos los niveles más altos.

Los argumentos derrotados colectivamente son derrotados provisionalmente, pero resulta que un argumento puede ser provisionalmente derrotado sin tomar parte en el caso de defectibilidad colectiva con otros argumentos. Aunque un argumento últimamente derrotado no puede derrotar a otros argu-

[11] Un sistema formal de argumentación defectible permite pruebas que varían en fuerza conclusiva. *'The idea is that, in case of conflicting proofs, the best proof remains in force to deliver the authoritative conclusion*" (Wreeswijk, 1991, p.526). Un estudio acerca de cómo pueden medirse las fuerzas de las razones lo encontramos en Pollock (1991).

mentos, argumentos derrotados provisionalmente podrían incluso derrotar a otros argumentos de un modo provisional. Para ilustrarlo, supongamos que α y β se derrotan recíprocamente. En este caso, α está *dentro* en el nivel par, y *fuera* en cada nivel impar. Ahora supongamos que α soporta a un derrotador para un tercer argumento γ. Esto tendrá el efecto de que γ estará *fuera* en cada nivel par, y volverá a estar *dentro* en cada nivel impar, de modo que γ es también derrotado provisionalmente, si bien no puede de hecho derrotar a los argumentos que lo derrotan provisionalmente. Este sería el diagrama de la situación:

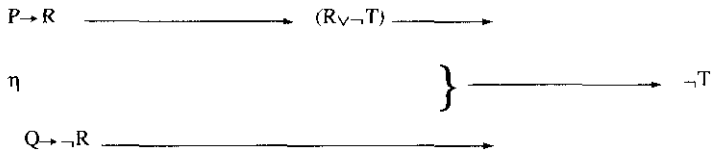
Σ (conjunto de argumentos)



Otro problema es que si la única fuente de defectibilidad entre los argumentos fueran las relaciones que se derivan de la utilización de los derrotadores del tipo rebatidor y saldador las manipulaciones puramente formales deberían producir argumentos que virtualmente derrotarían a cualquier argumento defectible, con el resultado negativo de que toda la estructura del razonamiento defectible debería colapsar. A este tipo de argumentos se les denomina *autodefectibles*. Por ejemplo, supongamos que P es una razón *prima facie* para ¬R, Q es igualmente una razón *prima facie* para R, S es una razón *prima facie* para T, y tenemos como *entrada* ={P,Q,S}. Entonces podemos construir los tres argumentos siguientes:

α: P → R β: Q → R σ: S → T

α y β se derrotan colectivamente, pero σ debería ser independiente de α y β, además de no derrotado definitivamente. El problema es que podemos construir un cuarto argumento (donde las inferencias deductivas son indicadas mediante un trazo continuo):



η usa una estrategia estándar para derivar una conclusión arbitraria a partir de una contradicción[12]. El problema es ahora que rebate η a σ. El mismo η es derrotado por α y β. Pero esto sólo tiene el resultado de que η es derrotado provisionalmente, y un argumento derrotado provisionalmente puede todavía derrotar a otro argumento. η está *fuera* en cada nivel par, pero continúa *dentro* en cada nivel impar. En consecuencia, fuerza a σ a estar *fuera* en cada nivel impar, y por tanto σ es también derrotado provisionalmente. Cuando en realidad σ no debería ser provisionalmente derrotado, sino que debería ser últimamente no derrotado. Este problema parece indicar que para manejar correctamente la idea de defectibilidad hace falta un desarrollo más completo de esta teoría del razonamiento defectible. Entre otras cosas, la definición de estar *dentro* en un nivel debería ser modificada de modo que los argumentos autodefectibles fueran excluidos en cada nivel. Puesto que η es un argumento autodefectible debería estar fuera en todos y cada uno de los niveles y no sólo en los niveles pares. De este modo, η no debería ser considerado como un argumento provisionalmente derrotado, sino como un argumento que se derrota internamente a sí mismo. Argumentos de este tipo no deberían tan siquiera contar como argumentos de nivel 0, evitando así su presencia en niveles más altos.

6. Defectibilidad

Podemos pensar en el razonamiento teórico como formando parte de tres procesos: (1) construyendo argumentos que soportan las conclusiones en las que estamos interesados, y adoptando creencias sobre la base de los argu-

[12] El argumento η, que no constituye sino una aplicación del principio clásico *de Ex contradictione quodlibet*, sirve precisamente para poner en entredicho la definición de estar fuera o dentro de un nivel para un argumento. Algunas de las posibles soluciones a estas cuestiones parecen hacer uso de la teoría de grafos (véase, Pollock, 1995).

mentos que encontramos; (2) retractando argumentos y creencias en función de los derrotadores; (3) restableciendo argumentos y readoptando creencias cuando los propios derrotadores resulten ser derrotados o retractados. Debido a la naturaleza del razonamiento defectible el sistema de razonamiento puede entrar en un bucle y, en consecuencia, no detenerse. Cualquier proposición puede ser adoptada, retractada, y restablecida muchas veces. Sin embargo, el razonador puede no saber nunca que una proposición dada ha alcanzado un estado estable. Y éste parece ser precisamente el modo en el que los humanos razonamos. El sentido en que el razonamiento humano correcto resulta defectible es que vemos a tal razonamiento como *'inocente hasta que se demuestre lo contrario'*. Una vez que una conclusión resulta estar justificada, es razonable aceptarla provisionalmente y actuar sobre ella. En el caso del razonamiento práctico este interesante recurso se ve acompañado por la utilización de heurísticos (como puede ser la *sobrecarga* (Pollack, 1992) o los módulos *quick & dirty* (Pollock, 1993), que nos permiten un ahorro importante de los medios cognitivos a nuestro alcance.

Contrariamente, las teorías del razonamiento no monótono en IA se han basado usualmente sobre una noción más rígida de defectibilidad de acuerdo con la cual una conclusión defectible es aceptable si y sólo si ha sido establecido que carece absolutamente de fallos. Esto último pretende probar que la conclusión está garantizada, lo cual exige asumir condiciones excesivamente fuertes de estabilidad en la información que manipulamos. Se supone que durante el razonamiento presente la información no va a sufrir fluctuaciones de ningún tipo[13]. Esta idea ha hecho parecer misterioso cómo el razonamiento no monótono podría ser implementable para el caso de un agente finito. Como ya hemos apuntado la solución consiste en adoptar el principio de defectibilidad de *'inocencia hasta que se demuestre lo contrario'*, y permitir a un agente racional actuar sobre sus conclusiones defectibles incluso aunque no se haya establecido conclusivamente el hecho de la no existencia de más derrotadores para las mismas. A los razonadores construidos sobre la base de mencionado principio se les denomina razonadores *interrumpibles*: A pesar de que el razonamiento pudiera ser interrumpido en algún punto de la computación, sin embargo es razonable actuar sobre conclusiones alcanzadas en ese punto.

Teniendo en cuenta las concesiones mutuas entre expresividad y tratabi-

[13] Bratman (1991) ofrece un estudio interesante sobre la estabilidad de la información en el ámbito de las actitudes proposicionales.

lidad, de los dos modos posibles de enfrentarse a este problema propuestos por Levesque & Brachman (1987), como son el de restringir la expresividad del lenguaje representacional dado, por ejemplo, restringiendo la base de conocimiento a un conjunto de enunciados atómicos junto con una colección de enunciados cuantificados universalmente que expresan asunciones del mundo cerrado de varios tipos[14]; y el de aceptar aproximaciones como soluciones, por ejemplo limitando el poder inferencial, OSCAR nos muestra cómo puede ser construido un razonador automatizado para el razonamiento defectible siguiendo esta segunda técnica.

7. Un Razonador Defectible Dirigido a Intereses

En este apartado describimos brevemente los pasos a seguir en la construcción de un razonador defectible de propósito general que es completo para la lógica de primer orden y probablemente adecuado para la concepción basada en argumentos del razonamiento defectible. El razonador defectible difiere de los clásicos probadores de teoremas precisamente en el hecho de que utiliza un concepto de defectibilidad enumerable dirigido a intereses que le permite un mejor equilibrio entre expresividad y tratabilidad (Cf. Pollock, 1991). Cualquier razonador automático que sea práctico y eficiente utiliza alguna estructura de control que le permite localizar su atención sobre argumentos que están íntimamente más conectados con las conclusiones que se está deseando establecer. El razonador deductivo dirigido a intereses OSCAR (Pollock, 1990, 1995), a diferencia de los usuales probadores de teoremas basados en resolución y skolemización, puede en principio dar cuenta del razonamiento defectible o no monótono. Veamos, pues, cuales son sus características principales un poco más en detalle.

Es dirigido a intereses en el sentido de que cuando intenta inferir una conclusión específica a partir de un conjunto de premisas, el razonador no sólo procede hacia adelante desde las premisas a la conclusión, sino que también lo hace hacia atrás desde la conclusión. Esto último es precisamente lo que se denomina reducción de propósito. El razonador comienza con un conjunto de premisas de entrada (inputs) y un conjunto último de conclusiones en las que está últimamente interesado denominado *'ultimate'*. El razonador guarda dos tipos de datos —adopciones e intereses. El primer tipo comprende las propo-

[14] Las llamadas bases de conocimiento *vividas* (Levesque, 1993).

siciones creídas en cualquier momento, y el segundo queda constituido por aquellas inferencias que el razonador está intentando inferir en cualquier estadio del razonamiento. El razonador razona hacia adelante desde adopciones a nuevas conclusiones que son entonces insertadas junto con las otras adopciones, y hacia atrás desde intereses a nuevos intereses. El conocimiento básico que utiliza el razonador para adoptar nuevos intereses es guardado en el conjunto denominado *'forset'*.

Paralelamente se hace una distinción entre aquellos esquemas de razón que son usados en el razonamiento hacia adelante y aquellos que son usados en el razonamiento hacia atrás. De hecho, los esquemas de razones individuales no pueden ser usados indiscriminadamente en un tipo u otro de razonamiento. Las tres reglas básicas que gobiernan al razonamiento dirigido a intereses son concretamente: *R-deduce, Interest-adoption; Discharge-interest*[15]. A lo que hay que añadir las reglas estructurales que OSCAR incorpora como la condicionalización o la reducción al absurdo que gobiernan el razonamiento suposicional.

Lo que se espera de un razonador de estas características es que descubra aquellas proposiciones epistémicamente garantizadas que nosotros asignamos como interesantes. Un razonador (que cumple con las condiciones de la defectibilidad enumerable) es adecuado relativo a un conjunto de argumentos A sí y sólo si, para cualquier entrada y cualquier proposición P en el conjunto *'ultimate'*: (1) si P está garantizada relativa a A entonces hay algún n tal que después de n ciclos de razonamiento, P es marcada como no derrotada y esta marca no es subsiguientemente cambiada; (2) si P no está garantizada relativa a A entonces hay algún n tal que después de n ciclos de razonamiento, P es marcada como derrotada y esta marca no es subsiguientemente cambiada.

Las inferencias relevantes son aquellas involucradas en la construcción de argumentos que soportan proposiciones en el conjunto *'ultimate'*, así

[15] *R-deduce*: si X es una razón para q, y adoptas algún miembro de X y todos los otros miembros han sido ya adoptados, entonces adopta q; *Interest-adoption*: si $\langle \Gamma, p \rangle$ es una instancia de un esquema de razón hacia atrás, y para alguna suposición X, el sistema adopta interés en p relativo a X, entonces adopta interés en los miembros de Γ relativos a X y guarda la base del interés en *forset*. Si todos los miembros de Γ han sido ya adoptados relativos a X, entonces adopta p relativo a X; *Discharge-interest*: si estás interesado en los miembros de X como un modo de alcanzar q, y adoptas algún miembro de X y los otros ya han sido adoptados, entonces adopta q relativo a X. Ejemplos sustanciales del funcionamiento de estas tres reglas los encontramos en Pollock (1990).

como todos los argumentos que soporten derrotadores para esos argumentos, y a su vez derrotadores para los argumentos derrotadores de los primeros argumentos, y así sucesivamente. Para que este proceso sea efectivo, el razonador debe ser capaz de encontrar todos los argumentos relevantes para algún enunciado en el que está interesado en 'demostrar', y además tendrá que adoptar intereses en posibles derrotadores para cada inferencia que lleve a cabo. Cumplidos estos dos requisitos, si hay un argumento que soporta la conclusión deseada, éste será encontrado por el razonador. Para evitar las posibles variantes infinitas de los argumentos derivadas tanto de inferencias parasitarias como de pasos redundantes, el razonador deber ser interés-completo relativo al conjunto de argumentos A. Un razonador se dirá que es interés-completo relativo a A si y sólo si, para cualquier *entrada* y cualquier P y cualquier argumento en A que infiera P desde la *entrada*, si al razonador le es dada la *entrada* como conjunto de premisas y adopta intereses en P, entonces el razonador construirá un argumento que infiera P desde la *entrada* de tal modo que cualquier inferencia que ocurra en β también ocurre en α, y lo hace relativo a la misma suposición o a la suposición más inclusiva. De este modo se consigue que si el razonador *saltara* en un estadio de la computación, no quedaría por probar nada de interés que no hubiera sido probado con anterioridad a ese estadio, de modo que los miembros del conjunto 'ultimate' están garantizados si y sólo si están justificados en el estadio en el que el razonador salta.

Habría que añadir que OSCAR no es sólo un razonador dirigido a intereses, sino que además tiene la propiedad de ser *interrumpible*. Si tenemos en cuenta que un razonador que se enfrenta con problemas de cierta complejidad puede no detenerse nunca, no obstante, podemos exigir a dicho razonador que tenga la capacidad de llevar a cabo acciones en momentos específicos, incluso aunque el proceso de razonamiento no haya todavía terminado. A un razonador que cumpla esta condición se le denomina *interrumpible*.

8. Algunas Consideraciones Finales

Gran parte de los enfoques usuales sobre razonamiento defectible o no monótono prescinden de los resultados obtenidos en epistemología a la hora de construir sus teorías. En consecuencia dejan de lado el tratamiento de problemas ineludibles para el razonamiento del sentido común. Por ejemplo, no dan cuenta de los derrotadores de argumentos que aquí denominamos *salda-*

dores; construyen razonadores (probadores de teoremas) basados en el concepto de recursividad enumerable (que resultan insuficientes desde el punto de vista de una teoría que establezca una clara separación entre razonamiento teórico y práctico, y además pretenda recoger el razonamiento defectible); u olvidan el razonamiento suposicional. Desde estas limitaciones no es difícil apreciar que la metodología a seguir en la modelización de algunas de las propiedades intrínsecas al razonamiento defectible del sentido común debería estar más abierta a los resultados obtenidos en disciplinas como la epistemología o la psicología cognitiva, de modo que las modelizaciones no sólo se acomodaran más a la realidad, sino que a la postre fueran más eficientes. La teoría del razonamiento defectible arriba discutida constituye uno de los escasos ejemplos de cómo es posible casar efectividad computacional y adecuación epistémica en una misma teoría.

Asimismo, hemos visto que nuestro razonamiento implica una estructura de control que le hace estar más directamente relacionado con el tipo de problemas que tratamos de resolver; lo que, por otra parte, le permite ser más eficiente. No obstante, esta teoría del razonamiento dirigido a fines debe extenderse al caso del razonamiento práctico. Esto implica entender a los agentes racionales como sistemas que seleccionan objetivos, planifican y hacen uso de planes. La relación existente entre planes y objetivos es más complicada de lo que en principio se puede pensar. La valoración y enjuiciamiento de los objetivos con vistas a su selección final depende de una serie de evaluaciones que pueden ser computadas y asociadas con esos objetivos: la importancia de los costes y beneficios que conlleva la satisfacción del objetivo; urgencia e importancia de los objetivos; la razón que subyace al objetivo; su estado con respecto al plan global establecido al nivel del metarazonamiento, etc.

Además, el mero hecho de que se disponga de un plan exitoso para lograr un objetivo no es una razón suficiente para considerarlo un buen plan. Si la realización del plan implica una serie de consecuencias no deseadas para el agente, el plan puede ser abandonado e incluso, en una situación extrema, el agente puede verse forzado a desprenderse del objetivo perseguido. La construcción y evaluación de planes es un proceso costoso que consume tiempo. Un agente racional ideal, sin restricciones espacio-temporales, podría generar sistemáticamente todos los planes posibles y seleccionar para su ejecución aquellos que tengan valores esperados más altos. Sin embargo, de modo similar a lo que sucede en el caso del razonamiento teórico, los agentes sujetos a limitación de recursos deben centrar su atención de un modo más eficiente

sobre planes que resulten relevantes para la situación en la que se encuentren. No disponen del privilegio de poder elegir entre una serie infinita de planes, de los cuales la mayor parte es irrelevante y, por tanto, carente de utilidad alguna para sus propósitos. La moraleja de esto último es que los agentes están limitados en cuanto a recursos y necesitan planificar para reducir el gasto de esos recursos en los procesos de razonamiento y coordinar así su modo de actuar en relación a situaciones tanto presentes como futuras. En este sentido, una parte considerable de las técnicas de trabajo utilizadas en la construcción de razonadores defectibles puede ser igualmente aprovechable para el caso del razonamiento práctico.

Bibliografía

Audi, R. (1991), *Practical Reasoning*. London: Routledge.

Baker, A. and M. GINSBERG (1989), A Theorem Prover for Priorized Circumscription, In *Proceedings IJCAI-89*, 463-467, Detroit.

Bratman, M. E. (1991), *Planning and the Stability of Intention*. Report. No. CSLI-91-159, Stanford, CA.

Cutland, N. J. (1980), *Computability*. Cambridge, Cambridge University Press.

Geffner, H. (1990), Conditional Entailment: Closing the Gap between Defaults and Conditionals. In: *Proceedings of the Third International Workshop on Non-Monotonic Reasoning*, 58-72.

Ginsberg, M. L. (1989), A Circumscriptive Theorem Prover. *Artificial Intelligence 39*, 209-230.

Levesque, H. (1990), All I Now: A Study in Autoepistemic Logic. *Artificial Intelligence 42*, 263-310.

Levesque, H. J. (1993), Is Reasoning too Hard? In: N. Asher, J. Ezquerro, and K. Korta (eds.). *Proceedings of the Third International Colloquium on Cognitive Science*. ICCS-93.

Levesque, H. J. & R J. Brachman (1987), Expressiveness and Tractability in Knowledge Representation and Reasoning. *Computational Intelligence 3*, 78-93.

Lin, F. & Y. Shoham (1990), *Argument Systems: A Uniform Basis for Nonmonotonic Reasoning*, Research Report, Stanford University, Department of Computer Science.

Loui, R. P. (1987), Defeat Among Arguments: A System of Defeasible

Inference. *Computational Inference 3*, 100-106.

McCarthy, J. (1980), Circumscription — A form of Non-monotonic Reasoning. *Artificial Intelligence 13*, 21 - 48.

Moore, R. C. (1985), Semantical Considerations on Nonmonotonic Logic. *Artificial Intelligence 25*(1), 75-94.

Nute, D. (1990), Basic Defeasible Logic. In: L.Fariñas del Cerro and M.Pentonnen (eds.). *Intentional Logics for Programming*.

Nute, D. & M. LEWIS (1986), *A Users Manual for d-Prolog*. ACMC Research Report 01-0016. Athens: University of Georgia.

Peirce, C. S. (1987), *Obra Lógica Semiótica*. (Versión castellana de R. Alcalde y M. Prelooker de algunos de los textos de C.S.Peirce: *Selected Writings*, *Collected Papers*, etc.). Madrid: Taurus.

Pollack, M. (1992), The Use of Plans. *Artificial Intelligence 57*, 43-68.

Pollock, J. (1986), *Contemporary Theories of Knowledge*. Rowman and Littlefield.

Pollock, J. (1987), Defeasible reasoning. *Cognitive Science 11*, 481-518.

Pollock, J. (1989), *How to Build a Person*. Cambridge, MA: MIT Press.

Pollock, J. (1989a), OSCAR: a general theory of rationality. *Artificial Intelligence 1*, 209-226.

Pollock, J. (1990), Interest Driven Suppositional reasoning. *Journal of Automated Reasoning 6*, 419-461.

Pollock, J. (1991), A Theory of Defeasible Reasoning. *International Journal of Intelligent Systems 6*, 33-54.

Pollock, J. (1991a), Self-defeating arguments. *Minds and Machines 1*, 367-392.

Pollock, J. (1992), *How to Reason Defeasibly. The Oscar Proyect*. Technical Report. University of Arizona.

Pollock, J. (1993), The Phylogeny of Rationality. *Cognitive Science 17*, 588-622.

Pollock, J. (1995), *Cognitive Carpentry: a Blueprint for How to Build a Person*. Cambridge, Mass., The MIT Press.

Pérez-Miranda, L. A. (1994*), The Role of the Planining-Intention Pair in theories of Practical Reasoning*. Report No. ILCLI-94-CS-1. Diciembre 1994.

Pérez-Miranda, L. A. (1996), *Deciding, Planinng, and Practical Reasoning: Elements towards a Cognitive Architecture*. Cognitive Science Rescarch Report RP-96-6, School of Computer Science and Cognitive Science Research Center, University of Birmingham.

Reiter, R. (1978), On Closed World Data Bases. In: Proc. *Second Symp. on Theoretical Issues in Natural Language Processing*. Urbana, Illinois.

Reiter, R. (1980), A Logic for Default Reasoning. *Artificial Intelligence 13*, 81-132.

Wreeswijk, G. (1991), The Feasibility of Deafeat in Defeasible Reasoning. In: J.Allen, R.Fikes, E.Sandewall (eds.), *Principles of Knowledge Representation and Reasoning. Proccedings of the Second International Conference*. San Mateo, CA: Morgan Kaufmann.