UNIVERSIDAD NACIONAL DEL SUR

TESIS DE DOCTOR EN FILOSOFÍA

Restablecimiento y especificidad en Sistemas Argumentativos

Claudio Andrés Alessio

BAHÍA BLANCA ARGENTINA

2015

Prefacio

Esta Tesis se presenta como parte de los requisitos para optar al grado Académico de Doctor en Filosofía, de la Universidad Nacional del Sur y no ha sido presentada previamente para la obtención de otro título en esta Universidad u otra. La misma contiene los resultados obtenidos en investigaciones llevadas a cabo en el ámbito del Departamento de Humanidades durante el período comprendido entre el 21 de abril de 2009 y el ……. de …….. de ……., bajo la dirección del Dr. Gustavo Adrian Bodanza.

[Firma del Alumno]

UNIVERSIDAD NACIONAL DEL SUR
Secretaría General de Posgrado y Educación Continua

La presente tesis ha sido aprobada el …./…./…… , mereciendo la calificación de .....(……)

Resumen

El restablecimiento es un principio de los sistemas argumentativos que permite considerar un argumento derrotado como justificado, cuando todos sus derrotadores se encuentran finalmente derrotados. Algunos contraejemplos al restablecimiento han sido propuestos en la literatura. Éstos sugieren que el restablecimiento no puede ser considerado como un principio general de la argumentación rebatible porque los argumentos restablecidos pueden sustentar conclusiones incorrectas. Algunos autores argumentan que el problema no se debe al restablecimiento, sino a la formalización de los ejemplos. La solución consiste en hacer al lenguaje lo suficientemente expresivo como para obtener los resultados correctos. También advierten que no se debe retocar la formalización sólo para obtener los resultados deseados frente a ejemplos concretos. Por lo tanto, este enfoque debe combinarse con la búsqueda de principios generales para la elección de una formalización correcta. Teniendo en cuenta que buscar principios generales de representación puede ser una empresa difícil, el objetivo de esta tesis consiste en la identificación de algún criterio que permita i. neutralizar los contraejemplos, ii. preservar el lenguaje formal original, tanto como sea posible, y iii. mantener el restablecimiento como un principio general. Para identificar ese criterio, se analizan los contraejemplos y se detectan posibles causas del problema. Como resultado, se ha encontrado que la preferencia por especificidad entre argumentos puede ser usada para obtener tal criterio. Tres enfoques basados en tal preferencia se proponen y evalúan. Dos de ellos proponen la introducción de relaciones alternativas de derrota entre argumentos. El tercero se basa en un criterio de filtrado de los argumentos no máximamente específicos.

Abstract

Reinstatement is a principle of argumentation systems that enables the justification of a defeated argument when all its defeaters are in turn ultimately defeated. Some counterexamples to reinstatement have been offered in the literature. Specifically, counterexamples suggest that reinstatement cannot be taken as a general principle of defeasible argumentation because the reinstated arguments may support incorrect conclusions. Some authors argued that the problems are not due to reinstatement but to the formalization of those examples. Then, the solution is to make the language expressive enough to obtain the correct results. They also warn that one should avoid tinkering with the formalization in concrete examples just to get a desired outcome. Therefore, this approach should be combined with the search of general principles for choosing the proper formalization. Taking into account that finding general principles of representation could be a hard enterprise, the goal of this thesis is to identify some criterion that allows *i.* neutralize the counterexamples, *ii.* preserve the original formal language as much as possible, and *iii*. maintain reinstatement as a general principle. To identify that criterion, counterexamples are analyzed and possible causes of the problem are detected. As a result it is found that the preference by specificity among arguments can be used to obtain that criterion. Three approaches based on specificity are proposed and evaluated. Two of them introduce alternative defeat relations among arguments. The third one is based on filtering the non maximally specific arguments.

Índice

Tabla de Ilustraciones

Capítulo I: Introducción

1.1 Sistemas argumentativos y posibles contraejemplos

Los sistemas argumentativos son formalismos que permiten modelar información tentativa y potencialmente contradictoria mediante la construcción, comparación y evaluación de argumentos a favor o en contra de ciertas conclusiones. Diferencias específicas aparte entre los diversos sistemas argumentativos concretos, estos pueden ser caracterizados mediante un proceso consistente de varias etapas: construcción de argumentos, identificación del marco argumentativo y selección de los argumentos que constituirán las extensiones del sistema.

Los argumentos, en tanto entidades que sustentan conclusiones, se construyen a partir de una base de conocimiento previamente especificada en un lenguaje formal determinado y de las condiciones que un argumento debe satisfacer (*fase 1*).

Una vez establecido el conjunto de argumentos, puede suceder que dos o más de ellos no puedan ser simultáneamente aceptados. Con vistas a estipular qué argumento prevalece se apela a diversas relaciones que entre ellos pueden darse. Una de las relaciones más importantes recibe el nombre de *derrota* (Dung, 1995).

Lo dicho hasta aquí permite advertir que los argumentos pueden entenderse como entidades que sustentan conclusiones y que además mantiene relaciones (como la de *derrota*) con otras entidades del mismo tipo. El conjunto de argumentos y las

relaciones de derrota que entre ellos se dan originan lo que Dung (1995) ha denominado *'marco argumentativo'* (*fase 2*).

El principal problema para un sistema argumentativo consiste en determinar qué argumentos, de todos los construidos, pueden ser aceptados. Por ello, luego de que argumentos y relaciones se han establecido se procede a seleccionar aquellos argumentos que constituirán las extensiones del sistema (*fase 3*), i.e. el conjunto de argumentos que un agente estaría dispuesto a aceptar. Los argumentos elegidos serán aquellos que prevalezcan frente a sus rivales y que se supone constituyen buenas razones para las conclusiones que sustentan.

Esta última fase puede hacerse en base a la satisfacción de condiciones previamente especificadas, denominadas '*semánticas'*, que un conjunto de argumentos debe verificar, o mediante algún procedimiento de prueba denominado '*juegos argumentativos*'. En ambos casos diversas exigencias adicionales podrán pedirse a los argumentos para que estos califiquen como extensiones del sistema. Tales exigencias estarán regidas por criterios más tolerantes, usualmente bajo teorías crédulas o criterios estrictos bajo teorías escépticas. Sin embargo, hay acuerdo general en que el requisito mínimo que un argumento debe verificar es el de ser *aceptable* (Dung, 1995).

Un argumento será aceptable cuando para cada argumento que lo derrota, existe al menos uno que lo defiende de ese derrotador. En términos más o menos precisos es posible decir que cuando un argumento aceptable satisface ciertos requerimientos escépticos recibe la denominación de *argumento justificado.* Básicamente se dirá que está *justificado* cuando la *cadena* de defensores de tal argumento descansa en un/os

último/s argumento/s que no posee/n derrotador/es. Obviamente cualquier argumento que no cuenta con derrotadores es un argumento justificado.

Ya sea mediante las semánticas o mediante juegos, la selección de los argumentos es realizada en base a su interacción en el marco argumentativo al que pertenecen, es decir, a las relaciones de derrota que entre los argumentos se dan. La noción de aceptabilidad permite observar que un argumento aunque se encuentre derrotado por otro no necesariamente significa que tal argumento deba rechazarse. Puede suceder que un argumento *B,* que oficia como derrotador de otro *A,* esté a su vez derrotado por un argumento *C*, de manera que *A* puede verse *restablecido*. Por ello, para determinar el estado final de un argumento será necesario conocer todas las interacciones entre los argumentos, incluida la del restablecimiento. El restablecimiento puede ser ilustrado con el siguiente ejemplo:

Ejemplo 1.1.1

A: *Tweety vuela porque se sabe que es un ave.*

B: *Tweety no vuela porque según la observación realizada por Paul, Tweety es un pingüino.*

C: *La observación de que Tweety es un pingüino no es confiable porque Paul la hizo en condiciones inadecuadas para las técnicas que empleó.*

A partir de la consideración de los argumentos *A*, *B* y *C* es posible creer que Tweety vuela puesto que la única objeción para creer en ello, el argumento *B*, ha sido desacreditada por el argumento *C*.

El principio del restablecimiento goza de una amplia aceptación en los sistemas argumentativos. No obstante, se han sugerido una serie de ejemplos que parecen

ponerlo en duda. Ejemplos similares a los siguientes fueron propuestos originalmente por Horty (2001) para sustentar la crítica a tal principio:

Ejemplo 1.1.2

A: *Dado que Al es un ave y teniendo en cuenta que las aves por lo general vuelan es posible concluir que vuela.*

B: *Sin embargo, Al es una gallina, y dado que las gallinas no vuelan se puede concluir que Al no vuela.*

C: *Al es una gallina salvaje, una clase particular de gallinas que tienen la habilidad de volar, por lo que puede concluirse que Al vuela.*

Cuando el ejemplo 1.1.2 es modelado en sistemas argumentativos, por ejemplo en el sistema propuestos por Simari y Loui (1992) o por Prakken y Sartor (1996[b]), se obtiene que tanto el argumento *C* como *A* son argumentos justificados, es decir, argumentos que prevalecen frente a sus adversarios. Ahora bien, intuitivamente pareciera que las razones por las que alguien estaría dispuesto en aceptar que *Al vuela* no se deben al hecho de que *Al sea un ave* sino al hecho de que *Al es una gallina excepcional*, una gallina que vuela. La razón por la que los argumentos *C* y *A* están justificados en los sistemas apuntados se debe a que el argumento *C*, un derrotador de *B*, defiende a *A* de *B,* i.e. lo restablece.

Estando así las cosas, parecería que el restablecimiento permite que un argumento esté justificado, en este caso el argumento *A,* a pesar de que estar basado en razones incorrectas (*vuela porque es ave*), aunque vale la pena aclarar que sustenta una conclusión correcta (*Al vuela*).

Este ejemplo no parece demasiado problemático ya que en fin de cuentas el argumento sustenta una conclusión correcta. Sin embargo, el siguiente ejemplo permitirá detectar un impacto semántico más serio del restablecimiento y que, según Horty (2001), sugiere la idea de que no debe ser aceptado como un principio conveniente.

Ejemplo 1.1.3

A: *Dado que Beth es una empleada de Microsoft y teniendo en cuenta que tales empleados tienden a ser millonarios es posible concluir que Beth es millonaria.*

B: *Dado que Beth es una nueva empleada de Microsoft y teniendo en cuenta que tales empleados por lo general poseen menos de medio millón es posible concluir que Beth posee menos de medio millón.*

C: *Dado que Beth es una nueva empleada de Microsoft en el departamento X y teniendo en cuenta que tales empleados por lo general poseen al menos medio millón se puede concluir que Beth posee al menos medio millón.*

En este ejemplo, el argumento *B* derrota a *A.* Por otro lado, la aceptación de *C* lleva al rechazo de *B* de modo que, restablecimiento mediante, los argumentos *C* y *A* cuentan como argumentos justificados. A diferencia de lo que ocurre con el ejemplo 1.1.2, en este caso, el restablecimiento lleva a aceptar una conclusión incorrecta, *Beth es millonaria*. Si se tiene en cuenta toda la información disponible, no hay razones para creer que Beth sea millonaria. Las razones que llevarían a la aceptación de tal conclusión se deben a que es una empleada de Microsoft *estándar*, pero *no es el caso aquí.*

Con ejemplos similares, Horty (2001) sentencia la cuestión: el restablecimiento, además de aceptar ciertos argumentos problemáticos para conclusiones correctas (*el ejemplo de Al*), lleva a aceptar conclusiones que son simplemente incorrectas (*el ejemplo de Beth*).

El dictamen de Horty parece sumir a los sistemas argumentativos en un situación grave puesto que los ejemplos propuestos apuntan contra una noción central. Sin embargo, es importante recordar una sugerencia de Hansson en "*La formalización en la filosofía*", parafraseando, *aun si los contraejemplos divulgasen convincentemente una deficiencia en el modelo, esto no representa una razón suficiente para abandonarlo. Si los contraejemplos no pueden ser neutralizados sin una pérdida sustancial de simplicidad, entonces una respuesta apropiada es continuar usando el modelo, recordando sus debilidades* (Hansson, 2007: 46). En la presente tesis, por lo tanto, se pretenderá neutralizar los contraejemplos intentando que los sistemas argumentativos no tengan una pérdida sustancial. A continuación se explica en mayor detalle el objetivo.

1.2 Objetivo de la tesis

El restablecimiento es un principio involucrado en la mayoría de los sistemas argumentativos (Baroni y Giacomin, 2009). Los ejemplos 1.1.2 y 1.1.3 parecen ponerlo en duda. Frente a tales cuestionamientos diversas respuestas pueden encontrarse en la literatura (Horty, 2001; Loui y Stiefvater, 1992; Prakken, 2002, Prakken y Vreeswijk, 2002; Horty, 2012). Tales propuestas pueden alinearse en dos ideas principales. Por un lado, la sugerida por Horty (2001), el restablecimiento es un principio incorrecto. La otra, el restablecimiento no es la causa del problema, sino la representación de la información (Loui y Stiefvater, 1992; Prakken, 2002, Prakken y Vreeswijk, 2002; Horty, 2012).

La indicación de Horty viene acompañada de una sugerencia drástica, abandonar el restablecimiento. Por otro lado, la versión *representacional* sugiere optar por un lenguaje de representación que sea lo suficientemente expresivo a fin de evitar la aparición de resultados inadecuados.

En la presente tesis se parte del supuesto de que el restablecimiento en sí es un principio válido, siguiendo en este aspecto a (Loui y Stiefvater, 1992; Prakken, 2002; Prakken y Vreeswijk, 2002; Horty, 2012). Sin embargo, es claro que una solución representacional debe ser acompañada de un estudio de principios que rijan la elección de una representación correcta so pena que frente a ejemplos que son extrañamente modelados en los sistemas suponga una constante modificación del lenguaje. Obviamente, encontrar principios generales que permitan elegir la correcta representación parece una tarea difícil, por ello, el objetivo de esta tesis consiste en *identificar algún criterio que permita la obtención de los resultados correctos sin tener que recurrir a la modificación del lenguaje de representación del conocimiento*. Este abordaje, a diferencia del cambio o extensión de la representación, es más sencillo de implementar, dado que si es posible detectar tal criterio, sólo será necesario anexarlo sin necesidad de modificar otros aspectos de los sistemas. A su vez la estrategia, en caso de ser exitosa, permitirá neutralizar los contraejemplos sin significar una pérdida sustancial en los sistemas argumentativos.

Con vistas a identificar tal criterio se realizará un análisis de los ejemplos, detectando posibles causas del comportamiento anómalo. A partir de tal detección se procederá al diseño de la estrategia formal en base a las causas identificadas con la respectiva implementación de los ejemplos en un sistema argumentativo particular. A su vez, la propuesta será considerada a la luz de otros ejemplos canónicos (*benchmark problems*) planteados en la literatura a fin de valorar la generalidad de la misma y en

comparación con los resultados obtenidos por una estrategia representacional sugerida por Prakken (2002).

Por razones de simplicidad, las propuestas que se sugieran serán implementadas en un sistema específico denominado *DeLP*. La razón de su elección se debe al hecho de que *DeLP* es un sistema de representación del conocimiento y el razonamiento de tipo modular, de manera que las modificaciones en diversos aspectos del mismo no exige modificaciones en otros. Por otro lado, *DeLP* cuenta con un mecanismo implícito de comparación de argumentos basado en especificidad, i.e. no hace falta que las preferencias entre las reglas y/o argumentos sean representadas explícitamente por el usuario, y cuenta con la posibilidad adicional de definir también preferencias entre las reglas mediante criterios de cualquier tipo. Cabe aclarar, que la implementación en *DeLP* no significa que las propuestas sugeridas en la presente tesis no puedan ser implementadas en otros sistemas argumentativos puesto que en todos los casos se parte de una definición conceptual más general.

1.3 Resultados obtenidos

Para hacer frente al problema señalado anteriormente, *restablecimiento y resultados contraintuitivos*, se proponen diversas vías de escape.

La primera vía de escape parte del supuesto de que el problema expresado en los ejemplos 1.1.2 y 1.1.3 aparece cuando los argumentos *restablecedor* y *restablecido* se relacionan por especificidad (un principio de comparación ampliamente utilizado en la modelación de razonamiento default) y sus conclusiones se implican lógicamente. Teniendo en cuenta el rasgo señalado, se procedió a definir una relación de derrota alternativa, denominada S_1, entre argumentos consistentes.

La derrota S_1 puede ser expresada intuitivamente de la siguiente manera: Un argumento A derrota a otro B mediante S_1 cuando A es estrictamente más específico que B y la conclusión de B implica la conclusión de A. De manera que tanto en casos como el ejemplo 1.1.2, como en situaciones similares al ejemplo 1.1.3, mediante S_1, C derrota a A. Esta estrategia permite obtener el resultado esperado. En ambos ejemplos el único argumento justificado es el intuitivamente esperable. Adicionalmente, permite obtener otros resultados interesantes, pero, como se verá, no es lo suficientemente general y además introduce derrotas innecesarias e intuitivamente incorrectas. Esto motivó el desarrollo de la siguiente vía de escape.

La segunda vía de escape parte del supuesto de que el problema es debido al hecho de que los argumentos justificados contraintuitvamente por restablecimiento no verifican una propiedad, la de ser máximamente específicos. El objetivo consiste en introducir una relación que permita derrotar a aquellos argumentos que no son máximamente específicos. Tal derrota, llamada '*derrota por socavamiento*' se basa en la intuición subyacente al principio de máxima especificidad propuesto por Hempel (1965).

En términos generales, un argumento A será considerado máximamente específico con respecto a la conclusión que sustenta cuando no exista evidencia socavadora en contra de A. Se dirá que existe evidencia socavadora para A cuando la evidencia que permite *activarlo*, también *activa* al menos a un contraargumento B que también es más específico.

En el ejemplo 1.1.2 la evidencia "*Al es una gallina"* permite activar el argumento A, i.e. permite inferir la conclusión "*Al vuela"* a partir de las reglas "*por lo general las aves vuelan*" y "*todas las gallinas son aves*". Sin embargo, la misma evidencia, "*Al es un*

gallina", permite activar el argumento *B* que es un derrotador de *A*. De manera que el dato "*Al es una gallina*" constituye una evidencia que permite indicar que el argumento *A* no es maximente específico.

Por su parte, si se atiende a la evidencia total disponible en el ejemplo 1.1.3, claramente el argumento para "*Beth es millonaria",* no debería estar autorizado a inferir tal conclusión, al igual que el argumento para "*Beth tiene menos de medio millón"* dado que ambos argumentos son, en sentido de lo anteriormente señalado, no máximamente específicos i.e. existe evidencia socavadora para cada argumento.

Con vistas a implementar la anterior idea se introdujo la definición de una relación de derrota alternativa, denominada '*derrota por socavamiento'*. La propuesta permite modelar adecuadamente los ejemplos problemáticos y a su vez casos que ejemplifican otros problemas en los sistemas argumentativos. Sin embargo, un comportamiento deseable de S_1 no puede ser emulado en esta propuesta. Por ello se procedió a definir una tercera vía que integre ambas alternativas.

La tercera propuesta consiste en un refinamiento de la idea de máxima especificidad, la máxima x-especificidad, definida en la propuesta basada en la derrota por socavamiento. A partir de ella, se determina un criterio de preselección de argumentos aplicado antes de la construcción de un marco argumentativo. Aquellos argumentos que no satisfacen el requerimiento son simplemente descartados en la preselección. Posteriormente, los argumentos que hayan sido preseleccionados son comparados y luego, de entre ellos, se seleccionan a los argumentos justificados, i.e. el marco argumentativo es construido en base a los argumentos máximamente x-específicos. Los resultados obtenidos son prometedores porque permiten evitar los

resultados contraintuitivos de los ejemplos propuestos por Horty y a su vez permiten resolver una serie de inconvenientes adicionales sugeridos en (Horty, 2001, 2012).

1.4 Organización de la tesis

La tesis se encuentra estructurada en cuatro secciones principales. En la primera se realiza una presentación general del área. En la segunda se definen sistemas argumentativos específicos. Posteriormente se presenta la problemática señalada en la sección 1.1 frente sistemas particulares. Finalmente se proponen una serie de posibles vías de escape. En concreto la Tesis se organiza como sigue.

En el *capítulo II* se da una introducción a los sistemas para argumentación rebatible presentando un panorama histórico y las características generales que los diversos formalismos basados en argumentación poseen. Adicionalmente se definen los marcos argumentativos abstractos y procesos de justificación basados en juegos argumentativos.

En el *capítulo III* se presentan algunos sistemas argumentativos específicos a fin de ilustrar cómo los componentes generales son instanciados en sistemas concretos y con vistas a tomar algunos de ellos para destacar la problemática señalada la sección 1.1. El primer sistema considerado es el propuesto por Prakken y Sartor (1996[b]). Luego se exponen los elementos principales del sistema planteado por Bondarenko, Dung, Kowalski y Toni (1997). También se presenta el formalismo conocido como *MTDR* sugerido por Simari y Loui (1992) continuando con uno de sus refinamientos, *DeLP* (García & Simari, 2004). Finalmente se presentan los conceptos principales que estructuran parte del proyecto Oscar de Pollock (1995).

En el *capítulo IV* se analizan los diversos ejemplos que ponen en duda la generalidad del principio de restablecimiento frente a sistemas argumentativos concretos. Una

vez que los resultados muestran que los sistemas son incapaces de evitar la aparición de resultados no intuitivos se discute sobre cómo interpretar tal cuestión. Adicionalmente, se presenta una versión representacional que permite resolver la problemática destacada.

En el *Capítulo V* diversas vías de escape al problema, alternativas a una solución representacional, son presentadas. El capítulo se vertebra en tres partes donde se van exponiendo y evaluando cada una de las vías. En todos los casos se tienen en cuenta los ejemplos del capítulo IV. En el *Capítulo VI* finalmente se concluye.

Capítulo II: Sistemas Argumentativos

2.1 Introducción

Los sistemas argumentativos son una forma de modelar razonamientos rebatibles (*defeasible reasoning*) mediante la construcción, comparación y evaluación de argumentos a favor y en contra de ciertas conclusiones. En este enfoque se considera que un argumento es un *buen argumento* cuando esta correctamente construido y es defendido de todos sus contraargumentos.

La caracterización de los sistemas argumentativos puede hacerse mediante la respuesta a tres preguntas: ¿cómo se construyen los argumentos?, ¿cómo se comparan? y ¿cómo se evalúan? La respuesta a tales interrogantes permitirá identificar a los sistemas argumentativos como formalismos definidos en base a un proceso consistente de tres etapas: construcción de argumentos, identificación de las relaciones de derrotas que entre ellos pueden darse y selección del conjunto de argumentos que, finalizado el proceso de comparación y evaluación, satisfagan determinados requerimientos (Prakken, 2011).

Los sistemas argumentativos son estructurados a partir de la noción de argumento rebatible y de las relaciones de derrota que entre ellos puedan darse. Un argumento se dice rebatible cuando da buenas pero tentativas razones para su conclusión y puede ser derrotado por un argumento mejor (Pollock, 1987). El conjunto de argumentos y

las relaciones de derrota constituyen la estructura básica de cualquier sistema basado en argumentos.

Diversos enfoques para ambas nociones pueden encontrarse en la literatura. Por un lado, están aquellos enfoques denominados *abstractos* donde la estructura interna de los argumentos no se encuentra especificada y la relación de derrota es primitiva como en (Dung, 1995) o está definida a partir de otras nociones primitivas no especificadas como en (Amgoud & Cayrol, 1998). Por otro lado, están aquellos enfoques en que se especifican tanto la estructura interna de un argumento junto a la noción de prueba asociada a una lógica subyacente, como las condiciones específicas bajo las cuales un argumento derrota a otro (ej. Simari & Loui, 1992; Prakken & Sartor, 1997). Recientemente un tercer enfoque se ha desarrollado en el que se define un marco abstracto en el que los argumentos poseen cierta estructura interna aunque abstracta, lo que permite especificar también la relación de derrota (Martínez et al., 2006; Prakken, 2009; 2010; Baláž & Frtúz, 2012).

Teniendo en cuenta que los sistemas argumentativos son modelos formales para el razonamiento no monótono y con vistas a poner tales sistemas en contexto, en el presente capítulo, se expondrán los antecedentes del estudio del razonamiento rebatible desde abordajes filosóficos como formales y luego se realizará una aproximación general a los sistemas argumentativos.

El capítulo se organiza como sigue. En la sección 2.2 se presenta un panorama histórico conceptual del ámbito de estudio. En la sección 2.3 se propone una caracterización general de los sistemas argumentativos a partir de las nociones de argumento y derrota. En la sección 2.4 se define un modelo abstracto de argumentación rebatible y se establecen las condiciones de aceptabilidad de conjuntos

de argumentos. En la sección 2.5 se brinda una introducción de los juegos argumentativos en tanto teoría de prueba para los sistemas en cuestión. En la sección 2.6 se concluye.

2.2 Argumentación rebatible: un panorama

Aunque el estudio de la argumentación rebatible puede rastrearse hasta los trabajos de Aristóteles, en las últimas décadas su estudio ha recibido gran atención gracias al desarrollo de sistemas formales en el ámbito de la representación del conocimiento y el razonamiento (KR&R). Aportes desde el ámbito de la filosofía, el derecho y la Inteligencia artificial han contribuido en la empresa. A continuación se reseñarán algunas de las principales contribuciones hasta llegar a la constitución de los sistemas argumentativos propiamente dichos a finales del siglo XX.

Si bien, el razonamiento deductivo juega un rol central para Aristóteles en la articulación del conocimiento científico, el razonamiento dialéctico se ocupa de las opiniones generalmente admitidas. Obviamente, el razonamiento dialéctico es revisable y falla en la validez deductiva aunque no por ello deja de ser razonable. Aristóteles dio un amplio y variado número de ejemplos de tales razonamientos en los *Tópicos*. En la universidad medieval, el estilo dialéctico constituyó parte de la formación general en el *trívium* y se cristalizó en las prácticas de defensa de tesis y en los escritos de la época. Ya avanzando en el tiempo, es posible constatar que la tradición dialéctica permaneció ajena al desarrollo formal que evidenció la lógica a partir de Frege.

A pesar de lo señalado anteriormente, es posible rescatar parte del estudio del razonamiento rebatible en el terreno de la moral y el derecho apelando a los trabajos de Ross (1930) y Hart (1949) por ejemplo, como también en el ámbito epistemológico con los aportes ofrecidos por Chisholm (1966) y Pollock (1974). Otras contribuciones

al estudio de la rebatibilidad y la argumentación pueden encontrarse en los trabajos de Toulmin (1958) y Rescher (1977).

Ross en 1930 estudió la noción de obligaciones *prima facie* en el contexto del razonamiento moral. Según Ross un acto es obligatorio *prima facie* si éste tiene una característica que en virtud del principio moral subyacente, tiende a ser un deber adecuado. Cumplir una promesa es *prima facie* un deber porque hay un principio moral que lo estipula de ese modo: *uno debe hacer lo que ha prometido hacer.* Pero el acto puede tener otras características que impidan su cumplimiento y bloqueen la aplicación del principio. Por ejemplo, supóngase que Juan le ha prometido a Pedro ir a tomar un café con él y cuando es el momento de cumplir la promesa ocurre que la madre de Juan, la Señora McCarthy, repentinamente se enferma. Entonces, Juan tiene una obligación *prima facie* de hacer las compras de su casa, basado en el principio moral, *se debe ayudar a los padres cuando lo necesitan.*

Para poder determinar cuál es la obligación correcta que se debe cumplir se deben considerar todas las condiciones, i.e. comparar todas las obligaciones *prima facie* que están basadas en aspectos determinados y encontrar o decidir cuál es la de mayor incumbencia entre aquellas que se encuentran en conflicto. Si se califica a la cláusula *considerar todas las condiciones*, como *considerar todas las condiciones que el agente sepa*, entonces claramente el razonamiento involucrado es revisable. Si primero se sabe que Juan le *prometió* a Pedro ir a visitarlo, se puede concluir que Juan *debe* ir a visitar a Pedro. Pero si luego se sabe que en ese momento, la madre de Juan se enferma se concluye que Juan *debe* cuidar a su madre. De modo que frente a dos argumentos rivales, uno de ellos pude prevalecer sobre el otro basado en los principios morales correspondientes.

En la obra *"The Ascription of Responsability and Rights"* de Hart (Hart, 1949), puede encontrarse tal vez el primer uso de la expresión *revisabilidad* o *rebatibilidad* (*defeasibility*) explícitamente nombrada y una definición técnica de su uso. Según Hart, los conceptos legales son *rebatibles* en el sentido de que las condiciones para cuando un hecho o situación clasifica como una instancia de un concepto legal son sólo *presumibles* u *ordinariamente* suficientes. Si en un proceso legal, una de las partes da pruebas para que un hecho o una situación clasifique como una instancia de un concepto legal determinado, no significa que esto permanezca estable, de hecho, en el proceso legal el peso de la prueba pasa al oponente, quien tiene la misión de probar hechos adicionales, que, a pesar de las pruebas del proponente, eviten o impidan la conclusión que ha sido concedida. La discusión de Hart trató sobre el fenómeno de la rebatibilidad en términos legales, pero claramente, esto supuso un desafío a la lógica a fin de que permita explicar tales situaciones.

Toulmin en (Toulmin, 1958) introdujo un modelo conceptual de argumentación. En su bien conocido esquema para argumentos, cuatro partes son distinguidas, *conclusión* (claim), *justificación* (warrant), *dato* (datum) y *soporte* (backing). Los contraargumentos son argumentos que pueden atacar alguna de esas partes. Para Toulmin, un argumento es *válido* si este puede resistir las críticas en una disputa conducida adecuadamente. En diversos trabajos como (Prakken, 2005c; Fuentes & Santibáñez, 2014) se destacan los aportes de Toulmin en el ámbito de los sistemas argumentativos.

Roderick Chisholm (1966) desarrolló una teoría del conocimiento basada en la noción de derrotabilidad (*defeasibility*). Los sentidos, según Chisholm, dan buenas pero rebatibles razones para creer en los hechos del mundo. Si un sujeto está frente a algo que le parece rojo (*experiencia sensorial*) entonces tiene razones para *presumir* que

efectivamente está en la presencia de algo rojo. Sin embargo, tal presunción es revisable si, por ejemplo, descubre que el entorno es relevantemente anormal, por caso, todo el ambiente está iluminado por una luz roja.

John L. Pollock desarrolló las ideas de Chisholm en una teoría epistemológica (Pollock, 1974) basada en dos tipos de razones. Las razones rebatibles y las conclusivas. Las rebatibles, también llamadas razones *prima facie,* son definidas en términos de la posibilidad de contar con derrotadores. Los derrotadores son razones con la potencialidad de poner en duda otras razones. Según Pollock existen dos tipos de razones derrotadoras, las *rebatidoras* y las *socavadoras*. Las primeras son razones que atacan las conclusiones mediante el soporte de una conclusión complementaria, mientras que las razones socavadoras atacan la conexión existente entre una razón y una conclusión.

A partir de la posibilidad de sustentar creencias en razones *prima facie,* Pollock hace un estudio de la noción de justificación (*warrant*) de una creencia. Para que una creencia esté justificada (*warrant*) es necesario que haya un argumento que la sustente y que tal argumento emerja de un proceso iterativo justificatorio como no derrotado en el que el mismo argumento puede verse sucesivamente restablecido y derrotado. Posteriormente su trabajo fue refinado y adquirió una mayor formalización. En la actualidad es uno de los principales exponentes del área no sólo en filosofía sino también en el ámbito de la representación del conocimiento y el razonamiento con su *proyecto Oscar*.

Por otra parte, Nicholas Rescher (1977) desarrolló un sistema de disputa formal. La disputa supone tres participantes: el proponente, el oponente y el juez. Tres tipos de jugadas fundamentales realizarán el proponente y el oponente. Las primeras

denominadas jugadas propiamente dichas, las contrajugadas y las contrajugadas dialécticas.

Las jugadas propiamente dichas son afirmaciones de tres tipos: categóricas, cautas y provisorias. Por su parte, las contrajugadas son contrajugadas para afirmaciones categóricas, contrajugadas para afirmaciones o negaciones cautas y contrajugadas para afirmaciones o negaciones provisorias. Finalmente, las contrajugadas dialécticas son contrajugadas para una negación provisoria, contrajugadas para una contraafirmación provisoria, contrajugadas para una excepción débil y contrajugadas para una excepción fuerte.

Uno de los aportes más interesantes en Rescher, además de la definición del proceso dialéctico, es la apelación a afirmaciones provisorias. Estas afirmaciones son expresadas como *P/Q* y representan sentencias como *"usualmente P es el caso cuando Q es el caso".* Claramente hay un gran parentesco entre las afirmaciones provisorias de Rescher y las reglas modeladas en sistemas para razonamiento no monotónico como *"por lo general los A son B".*

El juego definido por Rescher consiste en un intercambio de jugadas entre proponente y oponente. Las reglas estipulan qué tipos de jugadas puede hacer cada uno. El juego finaliza cuando uno de los participantes acepta las razones del otro, o si no existe consenso, cuando el juez asigna una decisión en base a los argumentos esgrimidos por las partes.

Adicionalmente, es importante destacar que en filosofía se encuentran otros estudios que apelan a la noción de derrotabilidad en el contexto de teorías epistemológicas

como en (Annis, 1973; Barker, 1976; Bergman, 1997; Klein, 1976; Lehrer & Paxson, 1969; Plantinga, 1986; 1993[a]; 1993[b]; 1995; Swain, 1974).

Pasando ya al terreno de la representación del conocimiento y del razonamiento (KR&R), el estudio de la rebatibilidad ha estado emparentado al desarrollo de sistemas que pretenden modelar lo que ha venido en llamarse razonamiento no monotónico. Minsky (1975) fue, tal vez, uno de los primeros en señalar que la lógica clásica es incapaz de modelar razonamientos que apelan a reglas tentativas, como "*por lo general las aves vuelan"* puesto que el agregado de nueva información puede obligar a abandonar conclusiones anteriormente obtenidas. De modo que según Minsky, es deseable que la lógica sea no monotónica.

Con vistas a modelar información tentativa y potencialmente contradictoria, diversos sistemas fueron propuestos desde aquella época. Algunos de ellos conocidos como sistemas de propósito especial tales como *base de datos*, definidas a partir de la hipótesis de mundo cerrado (Reiter, 1978; 1984); *programas lógicos* que emplean negación por falla (Clark, 1978); *sistemas de mantenimiento de verdad* (Doyle, 1979); o *redes de herencia* (Touretzky, 1986). Otros conocidos como sistemas de propósito general tales como la *Lógica No Monotónica* (McDermott & Doyle, 1980; McDermott, 1982); *Lógica Default* (Reiter, 1980; Reiter & Criscuolo, 1981; Brewka, 1994; Baader & Hollunder, 1993, Delgrande & Schaub, 1994); *Circunscripción* (McCharty, 1980); *Lógica Autoepistémica* (Moore, 1984; Konolige, 1987), *Implicación Preferencial* (Shoham, 1988). Como para citar algunos ejemplos paradigmáticos. En este contexto surgen los *sistemas basados en argumento*.

A continuación se presentan una breve reseña de algunos de los trabajos que contribuyeron directamente en el desarrollo de los sistemas argumentativos.

Uno de los primeros abordajes que brindó una fuerte influencia en el desarrollo de los sistemas argumentativos fue la propuesta de Doyle (1979). Doyle definió un sistema de mantenimiento de verdad (TMS) que puede entenderse como un método para la representación de creencias conjuntamente con una justificación para tales creencias y dotado con la capacidad de manejar la incorporación de nueva información. Para hacer frente a este último requerimiento, Doyle le dio al sistema la capacidad de retractar creencias mediante el empleo de razones *derrotadoras*. En la misma línea que Doyle, otros enfoques han sido propuestos (de Kleer, 1986; Elkan, 1990; McAllester, 1990; Forbus & de Kleer, 1993).

En TMS se utilizan dos estructuras de datos: nodos y justificaciones. A partir de tales estructuras se almacenan argumentos para las creencias potenciales de un agente. Los nodos representan creencias, mientras que las justificaciones codifican razones que sustentan los nodos. Un nodo *n* será creído en el sistema si existe un argumento válido para tal nodo. Se dirá que un argumento es válido si está basado en un conjunto de nodos que pertenecen a las creencias de TMS. Claramente podría existir argumentos que no estén basados en ningún nodo, en ese caso, se dice que se está en presencia de suposiciones.

Para crear y mantener la estructura conformada por nodos y justificaciones se definen diversas acciones en un TMS. Estas son: crear un nodo, al cual se le asocia una determinada sentencia representando una creencia potencial; agregar una justificación a un nodo determinado; eliminar una justificación y marcar un nodo como contradictorio a fin de señalar la inconsistencia de cualquier conjunto de creencias que respalden a un argumento para ese nodo.

Finalmente, cabe destacar que en la propuesta de Doyle se utiliza además un tipo particular de justificaciones, las no monótonas. Este tipo de justificaciones permite realizar inferencias tentativas que pueden ser eliminadas al obtener información adicional.

Otro trabajo que tuvo un fuerte impacto en el desarrollo de los sistemas argumentativos (en particular en Loui, 1987; Simari & Loui, 1992) fue el planteado por Poole (1985). Poole desarrolló un sistema para modelar razonamiento default basado en la noción de explicación y un comparador de teorías para manejar la aparición de múltiples extensiones.

Por otro lado, Donald Nute, propuso una lógica revisable empleando una noción de derrota basada en la fuerza de los antecedentes de las reglas condicionales involucradas en los argumentos. Nute (1988[a]; 1988[b]; 1988[c]) refina su formalismo mediante la introducción de criterios que permiten resolver conflictos entre argumentos mediante un criterio de especificidad de la regla principal de un argumento (*Top-rule*).

Ronald Loui (1987) extrapoló los resultados obtenidos en epistemología (Chisholm, 1966; Pollock, 1974) al ámbito de la representación del conocimiento en su trabajo *Defeat Among Arguments.* La propuesta se basa en un lenguaje de primer orden *L* al que se le adiciona reglas rebatibles de la forma $\Phi \succ\!\!\sim \Psi$ donde Φ y Ψ son fórmulas de *L*. Estas reglas definen una relación metalingüística que expresa lo siguiente: *ante la falta de información que indique lo contrario,* Φ *es una razón para sustentar* Ψ. La base de conocimientos del sistema está constituida por un conjunto EK de hechos y un conjunto de reglas derrotables R.

Un argumento se define como un grafo de prueba en el cual la información *fluye* desde las premisas (*fuentes*) hasta la conclusión (*desagüe*). El grafo que define un argumento es acíclico, dirigido y posee un único *desagote*. Los nodos del grafo está etiquetados por fórmulas de L y no existen dos nodos con la misma etiqueta. Adicionalmente, las fuentes del grafo son elementos de EK, cada arco desde un nodo P hacia otro Q debe ser tal que Q se pueda inferir de EK $\cup$ Q o existe una regla $r \in$ R de la forma $P \succ\!\!- Q$.

Entre los argumentos es posible que se den diversas relaciones como la contraargumentación y la derrota. Un derrotador de un argumento A es un argumento B tal que las conclusiones de ambos argumentos son contradictorias y B es preferido a A de acuerdo a un criterio de comparación de argumentos previamente especificado. Por otro lado, un argumento A contraargumenta a B si existe alguna etiqueta n en el grafo de B tal que la conclusión de A y n son contradictorias entre sí.

Las inferencias en el sistema de Loui son caracterizadas a partir del conjunto de argumentos que se denominan justificados o garantizados, i.e. aquellos argumentos que prevalecen frente a sus adversarios. Un argumento A justifica la conclusión h que sustenta cuando no exista un argumento B que lo derrote y para cada contraargumento C de A existe un argumento que lo derrota.

El sistema de Loui sentó las bases fundamentales de un sistema argumentativo mediante la definición de las nociones de derrota, justificación e inferencia.

En 1989, Fangzhen Lin y Yoav Shoham propusieron un sistema argumentativo con el objetivo de modelar razonamiento de sentido común a partir de una perspectiva argumentativa.

Un argumento es definido como un árbol de reglas de inferencia siguiendo las siguientes condiciones:

i. Sea A un hecho base. El árbol formado por A como único nodo es un argumento que posee a A como raíz.

ii. Sean $p_1,\ldots,p_n$ argumentos cuyas raíces son $A_1,\ldots,A_n$ respectivamente y sea $A_1,\ldots,A_n \rightarrow B$ una regla de inferencia tal que B no es un nodo en ninguno de los árboles para $p_1,\ldots,p_n$ entonces el árbol p con raíz B y $p_1,\ldots,p_n$ como subárboles inmediatos constituyen un argumento. La regla $A_1,\ldots,A_n \rightarrow B$ es la etiqueta de todos los arcos que conectan a B con sus hijos.

iii. Sean $p_1,\ldots,p_n$ argumentos cuyas raíces son $A_1,\ldots,A_n$ respectivamente y sea $A_1,\ldots,A_n \Rightarrow B$ una regla de inferencia tal que B no es un nodo en ninguno de los árboles para $p_1,\ldots,p_n$ entonces el árbol p con raíz B y $p_1,\ldots,p_n$ como subárboles inmediatos constituyen un argumento. La regla $A_1,\ldots,A_n \Rightarrow B$ es la etiqueta de todos los arcos que conectan a B con sus hijos.

Si φ es la raíz del árbol se dice que p es un argumento para φ o que φ está basada en p. Con vistas a evitar un número infinito de argumentos redundantes de la forma A → A, A→A→A, A→A→A→A, etc. φ sólo puede aparecer como raíz del árbol.

$A_1,\ldots,A_n \rightarrow B$ denotan *reglas monótonas* donde $n > 0$ y $A_1,\ldots,A_n$, B son *fbfs*. en un lenguaje L. Por otro lado, $A_1,\ldots,A_n \Rightarrow B$ denotan *reglas no-monótonas* y al igual que las anteriores, $n > 0$ y $A_1,\ldots,A_n$, B son *fbfs*. Finalmente, toda *fbf*. A en el lenguaje L es una regla de inferencia denominada *hecho base*. Un hecho base representa información explicita, como "*Tweety es ave",* una regla monótona expresa conocimiento deductivo o irrefutable como por ejemplo "*Todos los pingüinos son aves".* Las reglas no monótonas modelan información tentativa como "*Por lo general las aves vuelan".*

Lin y Shoham definen el conjunto de inferencias del sistema apelando a la noción de estructura de argumento. Si R es un conjunto de reglas de inferencia, un conjunto de T es una *estructura de argumento* si verifica las siguientes condiciones:

i. Si p es un hecho base en R entonces $p \in T$

ii. T es cerrado: para cada $p \in T$ tal que p' es un subconjunto de p se cumple que $p' \in T$

iii. T es monótonamente cerrado: si p se puede obtener a partir de $p_1,\ldots,p_n$ mediante una regla monótona y $p_1,\ldots,p_n \in T$ entonces $p \in T$

iv. T es consistente: para cada fórmula φ se cumple que T no contiene argumentos para φ y ¬φ simultáneamente.

La noción de estructura de argumento es complementada con la exigencia de que tal estructura sea completa con respecto a una fórmula φ y lo será si contiene un argumento para φ o para ¬φ.

Por otra parte, en 1988, Konolige define un sistema como solución al problema conocido como *Yale Shooting Problem* (*YSP*) originalmente propuesto en (Hanks & McDermott, 1987). El sistema *ARGH* (*Argumentation with Hypotheses*) se adelantó en varios aspectos al desarrollo de los sistemas argumentativos que serian definidos en la siguiente década.

El problema *YSP* es en un caso de razonamiento sobre eventos. El principal problema consiste en la aparición de conflictos entre los hechos que tiende a persistir y los cambios acaecidos por ciertos eventos. Konolige emplea argumentación para permitir varios tipos de argumentos basados en la consideración de la persistencia y el cambio y dirimir situaciones entre argumentos conflictivos mediante principios de derrota. Uno de tales principios dice: *un cambio causado por un evento derrota el argumento*

basado en la persistencia. Sin embargo, tal principio de derrota es también derrotable, dado que la persistencia de un evento determinado puede ser más fuerte que el cambio de un evento particular. De hecho, una de las principales observaciones de Konolige consiste en que un principio de prioridades para definir las derrotas que sea general e independiente del domino sería muy débil y que la información sobre la semántica del dominio es más importante a la hora de decidir entre argumentos rivales.

Todos los sistemas y enfoques presentados hasta aquí pretenden dar cuenta de los razonamientos de sentido común. Los sistemas argumentativos fueron desarrollados en ese contexto, y tal como se ha señalado, son una forma de modelar razonamiento no monotónico mediante la construcción y comparación de argumentos a favor y en contra de cierta conclusión. En estos sistemas, tal como lo señalan Prakken y Vreeswijk (2000), la noción básica es la de argumento rebatible. La no monotonicidad es explicada mediante la interacción entre los argumentos en conflicto, nueva información puede llevar a la construcción de nuevos argumentos que derrotan los inicialmente propuestos.

Otro aspecto importante a destacar en el enfoque basado en argumentos consiste en que estos sistemas tienen un amplio rango de aplicación para la modelación de la derrotabilidad, no sólo para razonamientos defaults. Según Prakken y Vreeswijk (2000), tales sistemas pueden ser aplicados a cualquier forma de razonamiento con información contradictoria aunque la fuente de las contradicciones sea debida o no a la presencia de excepciones. Por ejemplo, la contradicción puede aparecer porque los argumentos rivales parten de diferentes fuentes de información, o estar causada por desacuerdos con respecto a creencias o principios éticos, morales o políticos. También, señalan Prakken y Vreeswijk (2000) varios sistemas permiten la

construcción y derrota de argumentos que han sido tradicionalmente llamados como ampliativos, tales como razonamientos inductivos, analógicos, abductivos o estadísticos. Claramente estas formas de razonamiento caen fuera del ámbito de otros sistemas para razonamiento no monotónico.

Desde el surgimiento de los sistemas argumentativos a la actualidad diversas propuestas han sido realizadas ya sea mediante la construcción de sistemas para representar razonamientos (Loui, 1987; Pollock, 1987, 1995; Bench-Capon et al., 1991, 1993; Simari & Loui, 1992; Vreeswijk, 1993ª, 1993ᵇ, 1993ᶜ, 1995, 1997; Loui et al., 1993; Sartor, 1994; Bench-Capon & Leng, 1994; Krause et al., 1995; Bodanza & Simari, 1995; Delrieux, 1995; Cayrol, 1995; Gordon, 1995; Loui & Norman, 1995; Kowalski & Toni, 1996; Starmans, 1996; Freeman & Farley, 1996; Verheij, 1996; Prakken y Sartor, 1996ª, 1996ᵇ; Prakken, 1993, 1995, 1997; Gordon & Karacapilidis, 1997; Amgoud & Cayrol, 1997; Augusto, 1998; Dung & Son, 2001; Besnard & Hunter, 2001; García & Simari, 2004) o la propuestas de modelos abstractos como por ejemplo (Dung, 1995; Bondarenko et al., 1997; Amgoud y Cayrol,1998, 2002; Bochman, 2003; Bench-Capon, 2003; Amgoud et al., 2004, 2008; Modgil, 2006; Martínez et al., 2006; Prakken, 2009; 2010; Baláž & Frtúz, 2012) o de aplicaciones y combinaciones con otros enfoques o sistemas como por ejemplo (Fox et al., 1993; Das et al., 1996; Fox & Parson, 1997; Parson et al., 1998; Clark, 1990; Falappa et al., 2002, 2004; Bench-Capon, 2003; Gómez & Chesñevar, 2004; Chesñevar et al., 2006; Možina et al., 2007).

También un área de fuerte desarrollo fue la que puede ser denominada *"Basada en Dung"* siguiendo a Wu (2012). En ésta área se ha buscado el desarrollo de semánticas para marcos argumentativos abstractos. Además de las semánticas clásicas *completa*, *estable*, *preferida* y *básica*, que serán estudiadas más adelante, varias alternativas han sido propuestas tales como *CF2* (Baroni et al., 2005), *prudente* (Coste-Marquis et al.,

2005), *semi-estables* (Caminada, 2006b), *stage* (Verheij, 1996b; Bench-Capon et al., 2005), *semántica ideal* (Dung et al., 2007) y *eager* (Caminada, 2007b) entre otras.

Con vistas a brindar un acercamiento conceptual de los sistemas argumentativos propiamente dichos se procederá a caracterizar las dos nociones fundamentales de tales sistemas: *argumento y derrota*.

2.3 Argumento y derrota

Los sistemas argumentativos gravitan en torno a la noción de argumento en tanto *prueba* de la conclusión que sustentan. Los sistemas propuestos establecen diversas maneras de cómo definir un argumento pero es claro que un argumento está constituido por un conjunto de premisas (*que ofician de razones para una conclusión*), una conclusión y una relación de justificación entre las razones y las conclusiones.

Pollock (1987) sostiene que existen dos tipos de argumentos, los estrictos y los rebatibles. Los argumentos estrictos son definidos a partir de *razones conclusivas*. Las razones conclusivas son razones que implican lógicamente una conclusión, de modo que todos los razonamientos deductivos forman parte de este tipo de argumentos, por ejemplo (P & Q) es una razón conclusiva para P. Por otro lado, los argumentos rebatibles son construidos a partir de razones *prima facie*. Estas brindan un sustento tentativo para una conclusión dado que información adicional puede invalidar ese sustento. *Tweety vuela porque es ave* es un ejemplo típico de argumento rebatible, y es rebatible porque la información *Tweety es pingüino* invalida el argumento debido a que los pingüinos son aves que no vuelan.

Además de ejemplos similares al de Tweety, basado en el empleo de reglas default, es posible identificar una gran variedad de razonamientos rebatibles. En general todos los argumentos que apelan a información brindada por la percepción, la memoria, el

uso de inferencias estadísticas o argumentos construidos a partir de testimonios (Pollock, 1987), permiten la construcción de argumentos que dan buenas razones para la conclusión que sustentan pero éstas no quedan establecidas de manera firme y pueden ser revisadas en presencia de argumentos mejores.

La rebatibilidad de un argumento está determinada por la posibilidad de la identificación de objeciones o ataques, i.e. contraargumentos. La presencia de objeciones contra ciertos argumentos no implica que tales argumentos dejen, inmediatamente, de ser buenas razones para las conclusiones que sustentan. Por ello en los sistemas argumentativos se distinguen las nociones de '*conflicto*', '*derrota*' y '*estado final*'. El conflicto permite detectar qué argumentos no pueden ser conjuntamente aceptados, la derrota dirime el conflicto y establece cuál de tales argumentos es mejor, y el estado final consiste en determinar, luego de comparar las diversas interacciones con otros argumentos, si el argumento en cuestión ha prevalecido frente a todos sus rivales.

Una vez que los argumentos son construidos puede suceder que dos o más de ellos no puedan ser conjuntamente aceptados debido a *inconsistencia* o algún criterio preestablecido que así lo declare. Por ejemplo, supóngase que a partir de un conjunto de información inicial pueden construirse los siguientes argumentos:

Ejemplo 2.3.1

A: *A partir del dato de que Tweety es ave y dado que por lo general las aves vuelan es posible concluir que Tweety vuela.*

B: *A partir del dato de que Tweety es pingüino y dado que por lo general los pingüinos no vuelan es posible concluir que Tweety no vuela.*

En el ejemplo es fácil advertir que *A* y *B* no pueden ser conjuntamente aceptados de un modo racional puesto que se aceptaría al mismo tiempo que Tweety vuela y que no vuela. Este tipo de relación entre argumentos es conocida en el marco de los sistemas argumentativos como relación de *conflicto*, aunque también puede ser llamada *ataque* o *desacuerdo.*

El conflicto entre un par de argumentos puede entenderse como una relación que señala la imposibilidad de aceptarlos conjuntamente. Tres tipos de conflictos pueden ser identificados en la literatura: *i. conflicto conclusión-conclusión* o *conflicto por rebatimiento* (Pollock, 1987, 1995, Prakken & Sartor, 1996[a], 1996[b]); *ii. conflicto conclusión-premisa o ataque a premisas* (Prakken & Sartor, 1995, 1996[a], 1996[b]; Vreeswijk, 1993; Prakken, 2010) y *iii. conflicto conclusión-inferencia* o *conflicto por socavamiento* (Pollock, 1987, 1995).

El primer tipo de ataque, *conflicto conclusión-conclusión*, se da cuando las conclusiones de dos argumentos son contradictorias. Por ejemplo:

Ejemplo 2.3.2

A: *Nixon es pacifista porque Nixon es cuáquero y se sabe que por lo general los cuáqueros son pacifistas.*

B: *Nixon no es pacifista porque Nixon es republicano y se sabe que por lo general los republicanos no son pacifistas.*

Este tipo de conflicto es definido y empleado en la mayoría de los sistemas propuestos (Pollock, 1987; Loui, 1987; Simari & Loui, 1992; Prakken & Sartor, 1995, 1996[a], 1996[b]).

El segundo tipo de conflicto, *conflicto conclusión-premisa* o *ataque a premisas,* se da cuando la conclusión de un argumento niega una premisa de otro argumento. Por ejemplo:

Ejemplo 2.3.3

A: *Juan nació en el país X y por lo general los nacidos en el país X hablan el idioma P. Si Juan habla el idioma P entonces cuando viaje al país H (que también hablan el idioma P) Juan no tendrá problemas de comunicación.*

B: *Juan nació en la aldea F del país X y por lo general los nacidos en la aldea F no hablan el idioma P. Por lo tanto Juan no habla el idioma P.*

La conclusión del argumento *B, "Juan no habla el idioma P",* niega un paso en las premisas del argumento *A*: "*Juan habla el idioma P*".

Dependiendo de los sistemas es posible identificar dos tipos de ataque a premisas. Uno de ellos se da cuando se niega un hecho sustentado por un sub-argumento del argumento atacado (como en el ejemplo 2.3.3) y otro se da cuando se niega una hipótesis a la que apela un argumento. Por ejemplo:

Ejemplo 2.3.4

Supóngase que a partir de cierta base de conocimientos se puede construir un argumento para:

A: *Tweety es Ave (P) y no es probable que Tweety sea pingüino (¬R) por lo tanto, Tweety vuela (T).*

y construirse otro para:

B: *Los medios usados para la observación de Tweety fueron precisos y fiables (Q), por lo tanto se llega a la conclusión de que Tweety es pingüino (R).*

En este caso una premisa del argumento *A,* esto es, $\neg R$ está en conflicto con la conclusión *R* del otro argumento dándose un caso de conflicto *conclusión-premisa* (Prakken & Sartor, 1995). Para equipar a un sistema de este tipo de derrota se requiere que el lenguaje de representación se capaz de expresar formalmente sentencias como "*no es probable que P(x)".*

El conflicto por socavamiento o *conflicto conclusión-inferencia* se da cuando la conclusión de un argumento niega la conexión entre premisas y conclusión del otro o establece la falta de sustento que las premisas ofrecen a la conclusión (o cuando niega un paso inferencial en un subargumento). Este tipo de conflicto se denomina *socavamiento* (*undercutting*) y fue propuesto inicialmente en (Pollock, 1987). El siguiente es el ejemplo canónico de ataque por socavamiento.

Ejemplo 2.3.5

Supóngase que a partir de cierta base de conocimientos se puede construir un argumento para:

A: *La mesa que esta ante mí, parece color rojo (P), por lo tanto es posible concluir que la mesa es roja (T)*

y a su vez puede construirse el siguiente argumento:

B: *La mesa que está ante mí parece color roja (P), porque se encuentra iluminada por una luz roja (Q), esto me hace dudar acerca de la conexión entre las premisas y conclusión de A.*

Nótese que el argumento *B* no niega la conclusión del argumento *A* ni tampoco alguna de las premisas, sino que es un argumento que señala que las premisas del argumento

en cuestión no *alcanzan,* dada la información disponible, para sustentar adecuadamente la conclusión.

Esquemáticamente, los conflictos pueden representarse como en la figura 2.3.1, donde las flechas punteadas representan la relación de inferencia entre premisas y conclusión, las flechas sólidas representan relaciones de conflicto y "¬(*A*≫*B*)" representa la negación de la inferencia de *A* a *B*:

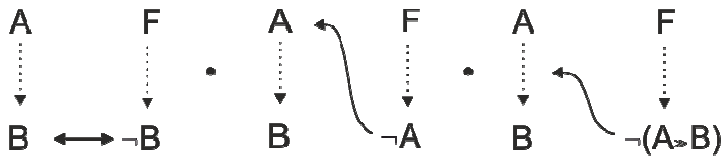


Fig. 2.3. 1: Conflictos: rebatimiento, ataque a premisas, socavamiento

Aunque la única relación de conflicto simétrica es la de rebatimiento es posible identificar conflictos mutuos de ataque a premisas, esto se puede ilustrar en la figura 2.3.2.

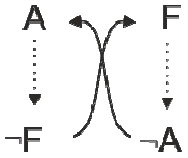


Fig. 2.3. 2: Ataque a premisas mutuo

Pollock (2001) también ha señalado la posibilidad de socavamiento mutuo:

Ejemplo 2.3.6

A: *Juan dice que Pedro no es digno de confianza (Q). Por lo tanto, es posible concluir que Pedro no es digno de confianza (R).*

B: *Pedro dice que Juan no es digno de confianza (S). Por lo tanto, es posible concluir que Juan no es digno de confianza (T).*

Es claro que *R* es una razón para creer que *S* no es una buena razón para *T.* Al mismo tiempo, *T* es una razón para creer que *Q* no es una buena razón para *R*.

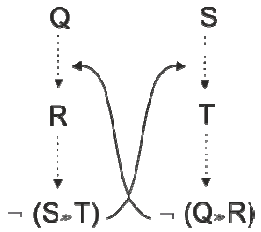


Fig. 2.3. 3: Socavamiento mutuo

La noción de conflicto no dice nada acerca del éxito de los argumentos rivales, simplemente señala la imposibilidad de su aceptación conjunta. Para determinar cuál de ellos prevalece es necesario algún mecanismo que permita decidir cuál (o ninguno) debe aceptarse. Esta necesidad lleva a la otra noción básica en sistemas argumentativos, la noción de derrota. La misión de esta consiste en determinar el éxito de un conflicto entre pares de argumentos.

En general, la relación de derrota es definida a partir de la relación de conflicto y de un orden previamente establecido entre los argumentos. El orden es realizado mediante una relación de preferencia. Dependiendo del tipo de información modelada es posible identificar diversos tipos de preferencia para resolver los conflictos, por ejemplo, *prioridades de leyes* en un sistema legal, *deseos* o *valores* en razonamiento práctico, *confiabilidad* en razonamiento epistémico, o *especificidad* en razonamiento default. Sin embargo, es importante destacar, como lo hacen Modgil y Prakken (2013), que existen derrotas dependientes de la relación de preferencia y derrotas que son independientes.

Al igual que las relaciones de conflicto es posible identificar tres tipos de derrotas: derrota por rebatimiento (*rebutting*), derrota por socavamiento (*undercutting*) y derrota de premisas (*undermining*).

La derrota por rebatimiento es dependiente de la noción de preferencia y puede ser definida como sigue:

Definición 2.3.1 **[Prakken & Sartor, 1996[b]]**

(Derrota por rebatimiento)

Sean *A* y *B* dos argumentos. Si la conclusión de *A* y la conclusión de *B* implican contradicción y *B* no es preferido a *A* se dice que *A rebate* a *B*.

□

La noción general de rebatimiento permite dos casos, una versión simétrica y una asimétrica.

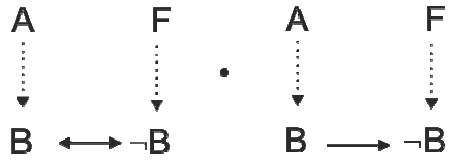


Fig. 2.3. 4: Derrota por rebatimiento: simétrico y asimétrico

El ejemplo 2.3.1 usualmente ha sido entendido como un caso de rebatimiento asimétrico, aunque esto es discutible puesto que puede ser expresado como un caso de socavamiento (Pollock, 2001; Walton, 2011), mientras que el ejemplo 2.3.2 es un caso de rebatimiento simétrico.

Como anteriormente se ha ejemplificado (en los ejemplos 2.3.3 y 2.3.4) es posible identificar dos casos de derrota a premisas, uno dependiente de la relación de preferencia mientras otro, independiente. Obviamente, esta diferencia estará dada por

la naturaleza de la premisa que es atacada. Tal como lo señala Modgil y Prakken (2013), las preferencias son necesarias para resolver estos conflictos excepto cuando la premisa establece alguna hipótesis que apela a la ausencia de evidencia de lo contrario, por ejemplo, negación por falla en programación lógica (Clark, 1978) como en el ejemplo 2.3.4.

El siguiente ejemplo ilustra que ciertos ataques a premisas deben apelar a un criterio de preferencia, puesto que de lo contrario atenta contra la intuición.

Ejemplo 2.3.7

A: *Juan, especialista en violines, dice que el violín en cuestión es caro porque es un Stradivarius. Luego, se puede concluir que el violín es caro.*

B: *Luis, el hijo de Pedro, de tres años de edad dice que el violín no es un Stradivarius. Luego, se puede concluir que el violín no es un Stradivarius.*

Si la derrota a premisas no involucrara preferencias, entonces, el argumento *B* derrotaría al argumento *A* pero no parece tener sentido aquí. Pero si se establece un criterio de confiabilidad (o autoridad), la sentencia "*no es un Stradivarius*" no derrota la premisa: "*Juan dice: El violín es un Stradivarius*". Porque el hijo de Pedro no es confiable en el tema en cuestión.

Cuando la derrota a premisas es dependiente de la noción de preferencia puede ser definida como un *rebatimiento* (indirecto) entre un subargumento propio de *A* y el argumento *B* de la siguiente manera, donde se especifica una de las versiones de la noción de derrota propuesta en Prakken y Sartor (1996[b]).

Definición 2.3.2 **[Prakken & Sartor, 1996[b]]**

(Derrota por rebatimiento indirecto)

Sean *A* y *B* dos argumentos. Si la conclusión de *B* y la conclusión de *A'* (donde *A'* es un subargumento propio de *A*) implican contradicción y *A'* no es preferido a *B*, *B* derrota por *rebatimiento indirecto* a *A*.

□

En el ejemplo 2.3.7 el argumento *B* no es capaz de derrotar el subargumento *Juan, especialista en violines dice que el violín es un Stradivarius,* por lo tanto es razonable concluir que *el violín es caro (por ser un Stradivarius).*

Por otro lado, se puede advertir que el ejemplo 2.3.4 no es un caso de derrota a premisas dependiente de la noción de preferencia de modo que es conveniente distinguirla de la anterior. Para ello se la denominará como *derrota a hipótesis*.

Definición 2.3.3 **[Prakken & Sartor, 1996 [b]]**

(Derrota a hipótesis)

Sean *A* y *B* dos argumentos. Si la conclusión de *A* niega una hipótesis en *B*, *A* derrota a *B*.

□

Ahora bien, si se tiene en cuenta el socavamiento debido a Pollock (1987) es claro que es una derrota independiente de un criterio preferencia y puede ser definido como:

Definición 2.3.4 **[Pollock, 1987]**

(Derrota por socavamiento)

Sean *A* y *B* dos argumentos. Si la conclusión de *A* niega la relación de inferencia en *B*, *A derrota por socavamiento* a *B*.

□

Los argumentos rebatibles, en tanto entidades que sustentan conclusiones pueden ser derrotados por argumentos mejores. Un contraargumento puede ser mejor que su rival debido a que tal contraargumento es preferido, o porque pone en duda el sustento que brinda, o porque niega una hipótesis aceptada como plausible. Con lo visto hasta aquí ya se cuenta con las herramientas necesarias para saber cuándo un argumento está derrotado. Sin embargo, el hecho de que un argumento se encuentra derrotado no significa que el argumento en cuestión no pueda prevalecer frente a sus adversarios. Para estudiar el estado final de un argumento se hace indispensable estudiar la condición de argumento aceptable que se hará a continuación.

2.4 Argumentos aceptables

Lo dicho hasta aquí permite señalar que la relación binaria de derrota entre argumentos no determina qué argumentos, finalizado el proceso de comparación, pueden ser considerados como argumentos que un agente *a* estaría dispuesto a aceptar, ni cuáles argumentos serán descartados. Puede suceder que un argumento B derrote a otro argumento *A* pero si un argumento *C* derrota a *B, A* puede restablecerse. El siguiente ejemplo, debido a Wu (2012), ilustra la idea. En el ejemplo se supone una conversación entre Paul y Olga.

Ejemplo 2.4.1

Paul: Mi auto es seguro porque tiene airbag.

Olga: Es verdad que tu auto tiene airbag pero yo no pienso que esto haga seguro a tu auto porque los airbags no son confiables. En el diario es posible encontrar varias noticias que reportan casos de airbags que no han funcionado.

Paul: Yo también he leído los diarios, pero un reciente estudio científico ha mostrado que los autos con airbags son más seguros que aquellos que no tienen y claramente los estudios científicos son más confiables que los informes periodísticos.

La conversación podría ser resumida mediante la presentación de tres argumentos:

A: *El auto de Paul es seguro porque tiene airbag.*

B: *El auto de Paul no es seguro porque los airbags no son confiables.*

C: *Estudios científicos muestran que los autos con airbags son más seguros que aquellos que no lo tienen.*

En esta conversación, el argumento *A* es derrotado por *B* y *B* es a su vez derrotado por *C*. El argumento A, entonces, puede ser restablecido ya que el contraargumento que llevaría al rechazo de *A*, ha sido derrotado. En suma, un argumento prevalece frente a sus rivales (*un argumento ganador)* aunque esté derrotado si todos sus derrotadores estén a su vez derrotados. Tal principio es denominado restablecimiento y se encuentra ilustrado en la figura 2.4.1.

Lo anterior permite advertir que para conocer el estado final de los argumentos es necesario hacerlo teniendo en cuenta la interacción de todos los argumentos.

La definición del estado final de un argumento usualmente divide a los argumentos en dos conjuntos: los argumentos ganadores y los perdedores (e.g. Dung, 1995). Algunas veces una tercera categoría es definida, la de los argumentos no decididos (e.g. Wu & Caminada, 2010). Por otro lado y dependiendo de si la teoría de base es crédula o escéptica los argumentos pueden ser considerados como defendibles o justificados

respectivamente. Una amplia discusión al respecto puede encontrarse en (Prakken & Vreeswijk, 2000)

Que un argumento sea ganador o perdedor puede ser definido de una manera declarativa o procedimental. La forma declarativa, usualmente obtenida a partir de una definición de punto fijo, *declara* cierto conjunto de argumentos como aceptables sin definir un procedimiento para testear si un argumento es miembro o no de tal conjunto. La manera procedimental especifica un procedimiento para determinar cuándo un argumento pertenece al conjunto de los argumentos aceptados.

En 1995, Dung propuso una teoría abstracta para argumentación rebatible, denominada marcos argumentativos, que permite concentrarse en el estado final de los argumentos. En general, los principales sistemas argumentativos y modelos formales para el razonamiento no monotónico pueden ser expresados en esta teoría.

La propuesta de Dung gravita en torno a la noción de *argumento aceptable* que es definida a partir de un conjunto de argumentos y una relación binaria entre ellos. Un subconjunto del conjunto de argumentos es seleccionado, el conjunto de los argumentos que un agente estaría dispuesto en aceptar, mediante un criterio predefinido denominado *semántica de extensiones* (*extension semantics*) ya sea desde un criterio crédulo o escéptico. Tal conjunto recibe el nombre de '*extensión*' de un marco argumentativo.

A continuación se brinda la definición de marco argumentativo propuesta por Dung (1995).

Definición 2.4.1 [**Dung, 1995**]
(marco argumentativo)

Un *marco argumentativo,* MA, es un par

$$\text{MA} = \langle AR, derrota \rangle$$

donde *AR* es un conjunto de entidades abstractas denominadas argumentos y *derrota* $\subseteq AR \times AR$ es una relación entre los miembros de *AR*.

Un argumento *A derrota* a un argumento *B* si y sólo si (*A, B*) $\in$ *derrota.* Un argumento *A derrota estrictamente* a *B* si *A* derrota a *B* y *B* no derrota a *A.*

Un conjunto de argumentos S derrota a un argumento *A* si y sólo si algún argumento en *S* derrota a *A*.

□

Ejemplo 2.4.2: El ejemplo 2.4.1 constituye un marco argumentativo donde AR = {*A, B, C*} y *derrota* = {(*B, A*); (*C, B*)}.

Una vez establecido el marco argumentativo se procede a determinar qué conjunto (o conjuntos) de argumentos satisfacen ciertas condiciones establecidas por las semánticas. Al menos dos condiciones elementales debe satisfacer un conjunto para calificar como extensión de un marco argumentativo: *defender a cada argumento que pertenece a él* y *ser consistente*, i.e. no hay pares de argumentos, en tal conjunto, que verifican la relación de derrota. Tales propiedades son capturadas en la siguiente definición.

Definición 2.4.2 **[Dung, 1995]**
(conjunto libre de conflicto, aceptabilidad y admisibilidad)
Sea S $\subseteq AR$

- S es *libre de conflicto* si y sólo si no existe un par de argumentos A, B en S tal que A derrote a B.
- Un argumento $A \in AR$ es *aceptable* con respecto a un conjunto S de argumentos si y sólo si para cada argumento $B \in AR$: si B derrota a A, entonces B es derrotado por S.
- S es un *conjunto admisible* si y sólo si S es libre de conflicto y cada argumento en S es aceptable con respecto a S

□

Ejemplo 2.4.3: El marco argumentativo MA = $\langle\{A, B, C\}\ \{(B, A), (C, B)\}\rangle$ cuenta con los siguientes conjuntos que son libre de conflicto: $\{A, C\}$, $\{B\}$, $\{A\}$, $\{C\}$, $\{\ \}$, mientras que $\{A, B, C\}$, $\{B, C\}$ o $\{A, B\}$ no lo son. Claramente de aquellos conjuntos que son libre de conflictos, $\{A, C\}$, $\{C\}$ y $\{\ \}$ son admisibles.

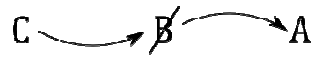


Fig. 2.4. 1: restablecimiento

Además de la defensa conjunta (admisibilidad) y la coherencia (conjunto libre de conflicto), las semánticas exigen otras condiciones a un conjunto de argumentos. Por ejemplo, la semántica *preferida* exige que sea maximal con respecto a esas propiedades. La semántica *estable* requiere que un conjunto sea coherente y derrote a cada argumento que no pertenezca a él. La semántica *completa* demanda, además de la coherencia y la autodefensa del conjunto en cuestión, que incluya cada argumento defendido por tal conjunto. Finalmente, la semántica *básica* (*grounded*) requiere que el conjunto sea construido a partir de argumentos aceptables con respecto al conjunto vacío. A continuación serán presentadas las diversas extensiones según fueron definidas originalmente en Dung (1995).

Definición 2.4.3 **[Dung, 1995]**

(extensión preferida)

Sea S $\subseteq$ *AR*. Se dice que S es una *extensión preferida* si y sólo si S es un conjunto admisible maximal (c.r. a $\subseteq$) de MA.

□

La extensión preferida es una semántica crédula y cómo tal puede haber más de una.

Ejemplo 2.4.4 (ejemplo 2.4.1 revisitado) No es difícil ver que el MA del ejemplo 2.4.1 cuenta con una única extensión preferida: E= {*A, C*}.

Ejemplo 2.4.5 (diamante de Nixon) El ejemplo clásico conocido como diamante de Nixon puede ser representado como un MA tal que MA = ⟨{*A,B*} {(*A,B*), (*B,A*)}⟩ donde *A* representa el argumento *Nixon es pacifista porque es cuáquero* y *B* representa el argumento *Nixon no es pacifista porque es republicano.* Este marco argumentativo, a diferencia del ejemplo 2.4.4, cuenta con dos extensiones: una en la que Nixon es pacifista, otra en la que no lo es: E_1= {*A*}, E_2= {*B*}.

Fig. 2.4. 2: Diamante de Nixon

Dung (1995) demuestra que las extensiones preferidas siempre existen en un marco argumentativo.

Lema 2.4.1 **[Dung, 1995]**

(lema fundamental)

Sea S *un conjunto admisible de argumentos, y A y A' argumentos que son aceptables con respecto a S. Entonces i.* S'= S $\cup$ {*A*} *es admisible, y ii. A' es aceptable con respecto a* S'.

□

Teorema 2.4.1 **[Dung, 1995]**

Sea MA un marco argumentativo. i. El conjunto de todos los conjuntos admisibles de MA forma un orden parcial completo con respecto al conjunto inclusión. ii. Para cada conjunto admisible S de MA, existe una extensión preferida E de MA tal que S $\subseteq$ E.

□

El teorema 2.4.1 conjuntamente con el hecho de que el conjunto vacio es siempre admisible implica el siguiente corolario:

Corolario 2.4.1 **[Dung, 1995]**

Cualquier marco argumentativo posee al menos una extensión preferida.

□

La *semántica estable* es menos tolerante que la preferida pero a diferencia de la anterior, no siempre existe.

Definición 2.4.4 **[Dung, 1995]**

(extensión estable)

Un conjunto de argumentos S, libre de conflicto, se dice que es una *extensión estable* si y sólo si S derrota a cada argumento que no pertenece a S.

□

Como se observa en la definición, la semántica estable, es una semántica agresiva pues derrota a todo argumento que no pertenece a la extensión.

El siguiente lema permite caracterizar a las extensiones estables como el conjunto constituido por todos aquellos argumentos no derrotados por S.

Lema 2.4.2 **[Dung, 1995]**

S es una *extensión estable* si y solo si S = {$A \mid A$ *no es derrotado por* S}.

□

Ejemplo 2.4.6 (ejemplo 2.4.1 revisitado) El MA del ejemplo 2.4.1, E= {A, C} es una extensión estable.

Ejemplo 2.4.7 (ejemplo 2.4.5 revisitado) El MA del caso de Nixon cuenta con dos extensiones estables, E_1= {A} y E_2= {B}.

El siguiente ejemplo permite mostrar que las extensiones estables no siempre existen, de hecho es fácil de advertir que el conjunto vacio no puede verificar la propiedad de derrotar todo aquello que no pertenezca al conjunto.

Ejemplo 2.4.8 Si MA=⟨{A, B, C} {(A, B), (B, C), (C, A)}⟩ es un marco argumentativo, ilustrado en la figura 2.4.3, tal marco no posee una extensión estable. Por ejemplo, el conjunto {C} es consistente pero no derrota a todo aquello que no pertenece a {C} por caso, no derrota a A.

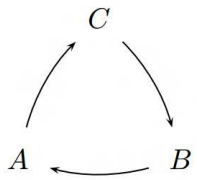


Fig. 2.4. 3: ciclo impar

Dung (1995) demostró que la semántica estable se corresponde con las extensiones de la lógica default (Reiter, 1980), las expansiones estables de la lógica autoepistémica (Moore, 1984) y los modelos estables de programación lógica (Gelfond & Lifschitz, 1991).

Las relaciones entre las extensiones estables y las preferidas, es capturada por el siguiente lema.

Lema 2.4.3 **[Dung, 1995]**

Cualquier extensión estable es una extensión preferida pero no viceversa.

□

La semántica escéptica de un marco argumentativo es capturada por la semántica básica. La misma es definida a partir de la función característica dada a continuación.

Definición 2.4.5 **[Dung, 1995]**

(función característica)

La función característica, denotada como F_{MA}, de un marco argumentativo MA = ⟨*AR*, *derrota*⟩ es definida como sigue:

- F_{MA}: $2^{AR} \rightarrow 2^{AR}$,
- $F_{MA}(S)$= {A | A *es aceptable c.r. a* S}.

□

La función característica permite definir el conjunto admisible de la siguiente manera.

Lema 2.4.4 **[Dung, 1995]**

Un conjunto de argumentos libre de conflictos S *es admisible si y sólo si* $S \subseteq F_{MA}(S)$

□

Si un argumento A es aceptable con respecto a S entonces A será aceptable con respecto a cualquier superconjunto de S.

Lema 2.4.5 **[Dung, 1995]**

F_{AF} es monotónica con respecto a $\subseteq$.

□

Definición 2.4.6 **[Dung, 1995]**

(extensión básica)

La *extensión básica* de un marco argumentativo MA, denotada como GE_{MA} es el menor punto fijo de F_{MA}.

□

La extensión básica se construye en base al siguiente proceso. En primer lugar se incorpora al conjunto los argumentos que no poseen derrotadores, i.e. F_{MA} se aplica al conjunto vacío. Luego los argumentos restablecidos por algún elemento del conjunto son agregados. Dado que la función es monótona, es posible asegurar la existencia de un menor punto fijo que puede ser obtenido mediante la aplicación reiterada de F_{MA}.

Ejemplo 2.4.9 (Ejemplo 2.4.1 reconsiderado) La extensión básica para el ejemplo 2.4.1 es $\{C, A\}$ puesto que 1. $F_{MA}(\emptyset)=\{C\}$; 2. $F_{MA}(\{C\})=\{C,A\}$; 3. $F_{MA}(\{C,A\})=\{C,A\}$.

Ejemplo 2.4.10 (diamante de Nixon) La extensión básica para el ejemplo de Nixon es el conjunto vacío dado que 1. $F_{MA}(\emptyset)=\{\ \}$; 2. $F_{MA}(\{\ \})=\{\ \}$.

De modo que ante una situación conflictiva, el razonador escéptico, no obtiene ninguna de las conclusiones alternativas.

La siguiente definición permite vincular las diversas semánticas.

Definición 2.4.7 **[Dung, 1995]**

(extensión completa)

Sea S $\subseteq AR$. S es una *extensión completa* si y sólo si S es admisible maximal (c.r. a $\subseteq$) y contiene a todos los argumentos defendidos por S.

□

Lema 2.4.6 **[Dung, 1995]**

Un conjunto de argumentos libre de conflictos E es una extensión completa si y sólo si $E=F_{MA}(E)$

□

Ciertas relaciones entre extensiones están expresadas en el siguiente teorema.

Teorema 2.4.2 **[Dung, 1995]**

- *Cada extensión preferida es una extensión completa pero no viceversa.*
- *La extensión básica es la menor extensión completa c.r. a* $\subseteq$
- *La extensión completa forma un semi-lattice completo c.r. a* $\subseteq$

□

Ejemplo 2.4.11: En el ejemplo 2.4.3 {*A, C*} es un conjunto admisible que es una extensión preferida, estable, completa y básica.

Las semánticas establecen diversos criterios de exigencia al conjunto de argumentos, aunque hay condiciones de coincidencia e inclusión que hacen que todos los criterios tengan algunos puntos en común. Su uso dependerá del tipo de información que se

modele o el tipo de extensión que se quiera obtener. Para poder definir estas propiedades es necesario tener en cuenta lo siguiente.

Definición 2.4.8 **[Dung, 1995]**

(marco argumentativo bien fundado)

Un marco argumentativo se dice *bien fundado* si y sólo si no existe una secuencia infinita A_0, A_1,..., A_n,... tal que para cada i, A_{i+1} derrota a A_i.

□

El siguiente teorema muestra que los marcos argumentativos bien fundados tienen exactamente una única extensión completa que es básica, preferida y estable. La demostración puede consultarse en (Dung, 1995).

Teorema 2.4.3 **[Dung, 1995]**

Cualquier marco argumentativo bien fundado tiene exactamente una extensión completa que es básica, preferida y estable.

□

El ciclo impar expresado en la figura 2.4.3 permite ilustrar el hecho de que una extensión sancionada por la semántica preferida, el conjunto vacío, no coincide con la semántica estable. Ello es así porque el marco argumentativo del ejemplo 2.4.8 no es bien fundado.

Otro punto de coincidencia entre las semánticas está dado por el hecho de que si la extensión preferida es también estable, se está en la presencia de un marco argumentativo coherente según la siguiente definición.

Definición 2.4.9 **[Dung, 1995]**

(marco argumentativo coherente)

Un marco argumentativo se dice *coherente* si cada extensión preferida es estable.

□

Por otro lado si la extensión básica coincide con la intersección de todas las extensiones preferidas (lo cual no siempre es el caso) se dice que tal marco es relativamente básico.

Definición 2.4.10 **[Dung, 1995]**

(marco argumentativo relativamente básicos)

Un marco argumentativo se dice *relativamente básico* si su extensión básica coincide con la intersección de todas las extensiones preferidas.

□

El siguiente ejemplo permite ilustrar el hecho de que no siempre la extensión básica coincide con la intersección de las preferidas.

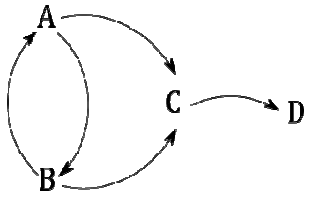


Fig. 2.4. 4: Argumento flotante

Ejemplo 2.4.12: Si MA=⟨{*A, B, C, D*} {(*A, B*), (*B, A*), (*B, C*), (*A, C*), (C, D)}⟩ es un marco argumentativo, ilustrado en la figura 2.4.4. Existen dos extensiones preferidas, E={*A, D*} y F={*B, D*}, obviamente, la intersección de ambos conjuntos es E ∩ F={*D*} mientras que la extensión básica es { }. De modo que este marco argumentativo no es relativamente básico.

En un marco argumentativo puede suceder que un argumento *C* al mismo tiempo que defiende a un argumento *A* de *B* pueda derrotar a *A*. En ese caso se dice que *C* es un argumento controvertido. Esta situación lleva a considerar marcos argumentativos no controversiales o limitadamente controversiales.

Ejemplo 2.4.13: En el ejemplo 2.4.12, ilustrado en la figura 2.4.4, *B* es un argumento controvertido con respecto a *C*, puesto que derrota a un derrotador de *C,* en este caso a *A,* y al mismo tiempo derrota a *C*.

Definición 2.4.11 **[Dung, 1995]**

(marco argumentativo no controversial)

Un marco argumentativo se dice *no controversial* si ninguno de sus argumentos es controversial.

□

Definición 2.4.12 **[Dung, 1995]**

(marco argumentativo limitadamente controversial)

Un marco argumentativo se dice *limitadamente controversial* si no existe una secuencia infinita de argumentos $A_0,\ldots,A_n,\ldots$ tal que A_{i+1} es controversial con respecto a A_i.

□

Atendiendo a las definiciones precedentes es posible inferir el siguiente teorema como también los lemas 2.4.7 y 2.4.8 y el corolario 2.4.2.

Teorema 2.4.4 **[Dung, 1995]**

- *Cualquier marco argumentativo limitadamente controversial es coherente.*

- *Cualquier marco argumentativo no controversial es coherente y relativamente básico.*

□

Lema 2.4.7 **[Dung, 1995]**

Sea MA *un marco argumentativo limitadamente controversial. Entonces existe al menos una extensión completa no vacía E de* MA.

□

Lema 2.4.8 **[Dung, 1995]**

Sea MA *un marco argumentativo no controversial y sea A un argumento tal que A no es derrotado por una extensión básica de* MA *y* $A \notin GE$*, entonces*

i. Existe una extensión completa E tal que $A \in E$*, y*

ii. Existe una extensión completa F tal que F derrota a A.

□

Corolario 2.4.2 **[Dung, 1995]**

Cualquier marco argumentativo limitadamente controversial posee al menos una extensión estable.

□

A partir de las definiciones anteriores es posible modelar cualquier conjunto de argumentos en base a la noción de derrota en tanto interacción entre argumentos. Para ello primero se define un marco argumentativo donde argumentos y relaciones entre ellos son considerados. Luego, mediante la aplicación de las condiciones señaladas por las semánticas se seleccionan aquellos conjuntos de argumentos que la verifican constituyendo las extensiones del sistema. Un argumento que pertenece a

una extensión podrá considerarse, ya sea desde una teoría crédula o escéptica, como un argumento que ha prevalecido frente a sus rivales.

Ahora bien, además de una semántica basada en extensiones es posible definir el estado final de un conjunto de argumentos mediante una semántica basada en estados de asignación o en etiquetas (Verheij, 1996; Jakobovits & Vermeir, 1999; Caminada & Gabbay, 2009; Caminada, 2006, 2007; Wu & Caminada, 2010). Baroni, Caminada y Giacomin (2011) desarrollan un panorama completo sobre las semánticas en los sistemas argumentativos.

A continuación se presenta un abordaje procedimental de la obtención de las extensiones de un sistema argumentativo a partir de la propuesta conocida como juegos argumentativos.

2.5 Juegos argumentativos

Las extensiones pueden verse como la semántica de un sistema argumentativo, mientras que una teoría de prueba puede hacerse mediante la propuesta de juegos argumentativos. Los juegos argumentativos, a diferencia de las semánticas que se focalizan en las propiedades que un conjunto de argumentos debe satisfacer, son una teoría de prueba o la dimensión procedimental de la argumentación donde se pretende *establecer el estado de un argumento individual.* Diversas propuestas han sido realizadas (Vreeswijk, 1993b; Simari et al., 1994; Gordon, 1994, 1995; Prakken & Sartor, 1996[b]; Loui, 1998; Vreeswijk & Prakken, 2000; Amgoud et al., 2000; Prakken, 2005; Caminada & Wu, 2008; Modgil & Caminada, 2009; Prakken & Sartor, 2009; Prakken, 2011; Bodanza et al., 2012). La idea de este enfoque puede ser explicada en términos de un juego de diálogo entre dos jugadores, un *proponente* y un *oponente*.

Un *juego argumentativo* es un diálogo en tanto serie de movidas alternadas entre el proponente (**P**) y el oponente (**O**). El proponente comienza con un argumento, argumento que será testeado, cada movida siguiente consistirá en un argumento que derrota a la última movida de la contraparte. El argumento inicial podrá tener cierto status si **P** cuenta con una estrategia ganadora, i.e. si el proponente es capaz de lograr que el oponente se quede sin movimientos. Las reglas exactas del juego dependerán de la semántica que se intente capturar pero los roles de los jugadores pueden caracterizarse de manera general tal como se indica a continuación.

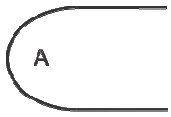
El rol de **P** es *constructivo*. Su función, mostrar que un argumento determinado, por caso *A,* pertenece a un extensión determinada. Para satisfacer el rol constructivo de **P**, **P** cumple una función de contra-contra-argumentador. Si existen argumentos que derrotan a *A*, buscará argumentos que derroten los derrotadores de *A* (aceptabilidad). Si encuentra tales argumentos, estos deben ser consistentes con los argumentos previamente jugados (libre de conflicto) dado que cualquier semántica es al menos libre de conflicto. Como se puede observar, el rol de **P** es eminentemente defensivo. Sólo construye argumentos necesarios para defender a *A*. Además, es importante para **P** mantener tal conjunto lo más pequeño que sea posible dado que mientras más grande sea, más difícil será de defender.

Por su parte, **O** asume un rol *crítico*. Su función consiste en proponer argumentos que derroten los argumentos propuestos por **P**.

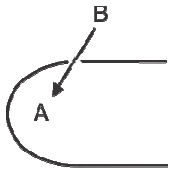
Si **P** tiene éxito en su tarea, entonces **O** falla, y viceversa. Por ello, un juego argumentativo es un juego de suma cero. A su vez, es deseable que las jugadas entre **P** y **O** sean finitas, de modo que también es un juego finito.

Antes de brindar una definición formal se procederá a presentar una serie de ejemplos que permitirán discutir algunas nociones estratégicas en los juegos tal como ha sido sugerido por Vreeswijk y Prakken (2000).

Ejemplo 2.5.1: Sea MA = ⟨{*A, B*} {(*B, A*)}⟩ un marco argumentativo. Supóngase que hay interés por conocer el status de *A* en base a la semántica preferida. Dado que las extensiones preferidas son conjuntos admisibles maximales, es suficiente para **P** mostrar que *A* pertenece a un conjunto admisible. La primera acción de **P** simplemente consiste en colocar *A* y el objetivo, construir un conjunto consistente de argumentos que defiendan a *A* de todos los posibles derrotadores:

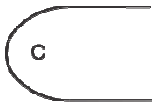


Si *A* no puede ser criticado, i.e. no existen derrotadores de *A*, S={*A*} es admisible y **P** tiene éxito en mostrar que *A* es preferido. Sin embargo *B* derrota a *A*.

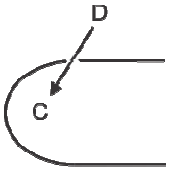


Ante esta situación **P** debe defender a *A* de *B*, pero dado que no existe un argumento para tal fin, **P** falla en su función de construir un conjunto admisible en torno a *A*. De modo que *A* no es admisible y en consecuencia, *A* no es preferido.

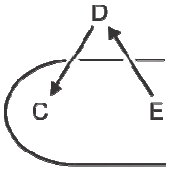
Ejemplo 2.5.2: Sea MA = ⟨{*C, D, E, F, G*} {(*D, C*), (*E, D*), (*F, C*), (*G, F*)}⟩ un marco argumentativo. Suponga que **P** quiere demostrar que *C* es preferido. La primera acción de **P** es jugar *C*:



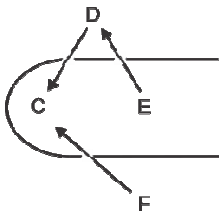
O derrota *C* con *D*:



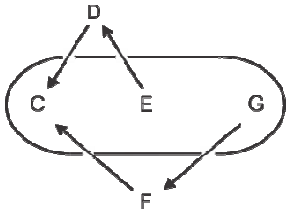
P defiende esta derrota de *C* con *E*:



La derrota de **O** contra *C* mediante *D* ha fallado. **O** vuelve a atacar *C* pero ahora con *F:*

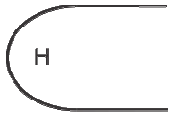


Ahora, **P** defiende a *C* de *F* con *G*. Dado que **O** no cuenta con otro argumento capaz de impedir que *C* sea considerado admisible, **P** puede cerrar el conjunto S:

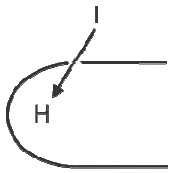


Ejemplo 2.5.3: Sea MA=⟨{*H, I, J, K*} {(*I, H*), (*J, I*), (*K, J*), (*J, K*)}⟩ un marco argumentativo.

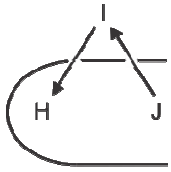
P quiere mostrar que *H* es admisible, para ello propone *H*:



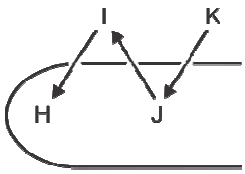
O derrota *H* con *I*



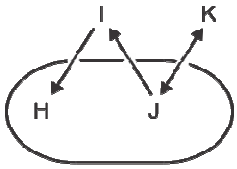
P defiende *H* de *I* con *J*



O ataca el defensor de *H* con *K*:

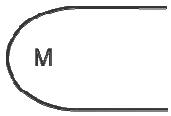


P defiende a *J* de *K* con *J* mismo, de modo que *J* se autodefiende. **O** es incapaz de proponer otro argumento, de modo que **P** gana y puede cerrar el conjunto S:

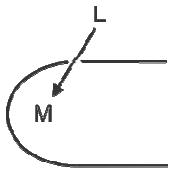


Este ejemplo, muestra que tiene sentido que **P** pueda repetir sus argumentos y **O** no, al menos desde el punto de vista de la semántica crédula y en la misma línea de disputa.

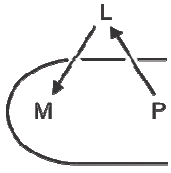
Ejemplo 2.5.4: Sea MA=⟨{*M, L, P, K, H*} {(*L, M*), (*P, L*), (*H, P*), (*K, L*), (*M, K*)}⟩ un marco argumentativo. Supóngase que **P** quiere probar que *M* es admisible. La primera acción es poner *M*:



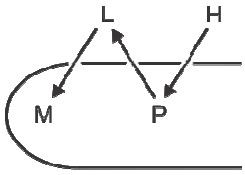
O ataca *M* con *L*



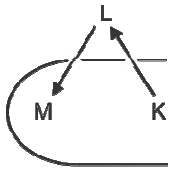
P defiende a *M* con *P*



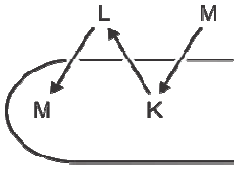
O ataca *P* con *H*



P vuelve la jugada e intenta defender *M* con *K* de *L* puesto que no puede defender a *P* de *H*

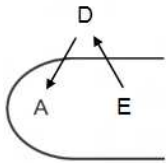


Sin embargo, **O** ataca *K* con *M* revelando inconsistencia en el conjunto defensor de *M*



P es incapaz de cerrar S. De modo que *M* no pertenece a un conjunto admisible. Note que **P** no puede derrotar *M* con *L*, dado que **P** debe construir un conjunto admisible para *M* y si lo hace, el conjunto no será libre de conflicto. Este ejemplo muestra que **P** no puede repetir las jugadas de **O** pero **O** puede repetir las jugadas de **P** dado que en tales repeticiones se puede revelar un conflicto en la posición de **P**.

Ejemplo 2.5.5: Sea MA=⟨{*A, B, C, D, E*} {(*C, A*), (*D, A*), (*E, D*), (*B, C*), (*B, E*), (*E, B*)}⟩ un marco argumentativo. Supóngase que **P** quiere mostrar la admisibilidad de *A.* **O** ataca *A* con *D*, **P** defiende a *A* de *D* con *E*.



Si **O** ataca *E* con *B*, **P** puede defender *E* repitiendo *E* mismo. Sin embargo, **O** puede *volver a atacar A*, pero ahora con *C*, luego **P** sólo puede defender *A* con *B*, entonces **O** puede repetir la jugada de **P** jugando *E* y revelando que la posición de **P** no es libre de conflicto.

Teniendo en cuenta los ejemplos considerados es posible preguntarse sobre la estrategia de repetición tanto para **P** como para **O**: ¿Tiene sentido para **P** (**O**) repetir sus propios argumentos?, ¿tiene sentido para **P** (**O**) repetir los argumentos de **O** (**P**)?

Las respuestas son dadas en los siguientes puntos:

- *Si* **P** *repite sus propios argumentos,* **O** puede fallar en encontrar o producir un argumento contra el argumento repetido. De modo que tiene sentido que **P** repita sus propios argumentos.
- *Si* **O** *repite sus propios argumentos*, **P** siempre contará con una estrategia defensiva para los argumentos repetidos por **O**. De modo que la repetición para **O** no tiene sentido. Aunque, claro está, que **O** puede r*epetir sus propios argumentos cuando la línea de argumentación sea distinta*.
- *Si* **O** *repite los argumentos de* **P**, **O** puede poner en evidencia que el conjunto de argumentos de **P** no es libre de conflicto y en consecuencia tampoco admisible. De modo que la repetición de los argumentos de **P** por **O** tiene sentido.
- Si **P** *repite los argumentos de* **O**, **P** al jugarlos introduce un conflicto en la colección de sus propios argumentos, de modo que no tiene sentido para **P** jugar los argumentos de **O**.

A continuación se presenta un proceso de justificación basado en juegos en el que se definen protocolos con vistas a establecer diversos criterios para determinar la justificación de un argumento. En concreto, será definido un modelo de diálogo general y protocolos específicos que capturarán diversas semánticas. Antes de continuar se introducirán los rasgos comunes o generales de los juegos, algunos ya enunciados anteriormente:

- En el juego hay dos partes: *proponente,* notado como **P**, cuya tarea consistirá en defender un argumento, y *oponente,* notado como **O**, cuyo objetivo será derrotar el argumento propuesto por **P**.
- El juego es un juego de suma cero, i.e. sólo un jugador gana.
- El juego es un juego finito, i.e. el número de argumentos jugados por **P** y **O** es finito.

Teniendo en cuenta los rasgos anteriores se procede a definir un juego argumentativo de la siguiente manera.

Definición 2.5.1 **[Bodanza et al., 2012]**

(juego argumentativo)

Un *juego argumentativo* en un marco argumentativo MA=⟨*AR*, *derrota*⟩ es un juego extensivo de suma cero en el que:

1. Existen dos jugadores, *i* y –*i,* quienes juegan los roles de **P** y **O** respectivamente.
2. Una *historia* en el juego es cualquier secuencia $A_0, A_1, A_2,\ldots,A_{2k},A_{2k+1},\ldots$ de elecciones de argumentos en *AR* realizada por los jugadores en el juego. A_{2k} corresponde a **P** y A_{2k+1} a **O**, para $k = 0,1,\ldots$
3. En cualquier historia, A_0 es el argumento que **P** intenta defender.

4. En una historia las *elecciones* del jugador i en el nivel $k > 0$ son $C_i(k) = \{A \in AR: \exists B \in C_{-i}(k-1), (A,B) \in derrota\}$.
5. Una historia de longitud finita K, $A_0,\ldots,A_K$ es *terminal* si A_K corresponde al jugador j ($j = i$ o $j = -i$) y $C_{-j}(k+1) = \varnothing$.
6. Los *pagos* son determinados en las historias terminales: el pago para **P** en $A_0,\ldots,A_K$ es 1 (**P** gana) si K es par (i.e. **O** no puede responder al último argumento de **P**) sino −1 (**P** pierde). A su vez, el pago en $A_0,\ldots,A_K$ para **O** es 1 si K es impar caso contrario es -1.

□

Un juego en el que **P** intenta defender un argumento A pude entenderse como un árbol con raíz en el que A es la raíz. Cada nodo no terminal en el nivel l consiste de una historia $A_0,\ldots,A_l$ y sus hijos son todas las historias en $A_0,\ldots,A_l, A_{l+1}$. Los nodos terminales son las historias terminales.

Definición 2.5.2 **[Bodanza et al., 2012]**

(estrategia)

Una *estrategia* para un jugador i es una función que asigna un elemento $A_{l+1} \in C_i(l)$, a cada historia no terminal $A_0,\ldots,A_l$ donde A_l corresponde al jugador $-i$. Una estrategia del jugador i se denomina *estrategia ganadora* para i si para cada estrategia elegida por $-i$, la historia terminal produce un pago 1 para el jugador i.

□

Si **P** tiene una estrategia ganadora significa que su argumento inicial puede ser defendido en contra de cualquier posible derrota. De lo contrario, si **O** tiene una estrategia ganadora significa que el argumento inicialmente propuesto por **P** no puede ser defendido.

Nótese que la estrategia ganadora para **P** u **O** no puede garantizarse si el árbol es infinito. Aún siendo finito, un marco argumentativo que no esté libre de ciclos en las relaciones de derrota puede producir un árbol infinito. Ahora bien, los juegos argumentativos de acuerdo a la definición presentada no son necesariamente finitos. Dadas las reglas del juego, es claro que una posible fuente de no finitud del juego es la existencia de ciclos en la relación de derrota. Existen básicamente dos formas de evitar la infinitud en el marco con un número finitos de argumentos:

a. Restringir el juego a marcos argumentativos en el que la relación de derrota es acíclica.

b. Agregar reglas prohibiendo ya sea a ambos o a algunos de los jugadores repetir los argumentos de alguna manera específica.

La primera alternativa lleva a evitar casos interesantes. Así que es conveniente seguir la última atendiendo también a lo discutido previamente sobre la repetitibilidad como estrategia en un juego.

La semántica básica (*grounded*) propuesta por Dung (1995) sanciona una única extensión y consiste en el menor punto fijo de la función característica F(S) tal como se ha visto anteriormente. Para capturar tal semántica son necesarias las reglas 1 a 6 en adición con el protocolo 1 a fin de asegurar la finitud del juego tal como es propuesto en (Bodanza et al., 2012).

Protocolo 1: Rules 1 a 6 + 7. Donde 7 es la siguiente regla:

7. **P** no tiene permitido jugar un argumento que ha sido previamente jugado por cualquier jugador.

Ejemplo 2.5.6 Sea MA = ⟨{*A, B, C*} {(*C, B*), (*B, C*), (*B, A*)}⟩ un marco argumentativo. Supóngase que **P** quiere probar que *A* es un argumento básico (*grounded*), entonces el juego consistirá en la siguiente secuencias de jugadas $A_{(P)} \leftarrow B_{(O)} \leftarrow C_{(P)} \leftarrow B_{(O)}$. Claramente **P** no cuenta con una estrategia ganadora para defender a *A*.

Ejemplo 2.5.7 Sea MA = ⟨{*A, B, C*} {(*C, B*), (*B, A*), (*A, B*)}⟩ un marco argumentativo. Supóngase que **P** quiere probar que *A* es básico (*grounded*). La secuencia de jugadas $A_{(P)} \leftarrow B_{(O)} \leftarrow C_{(P)}$ muestra que **P** tiene éxito, pues cuenta con una estrategia ganadora para defender a *A* de *B* mediante *C*.

El siguiente protocolo captura la admisibilidad, también debida a (Bodanza et al., 2012).

Protocolo 2: Rules 1 a 6 + 7- 8. Donde 7 a 8 son las siguientes:

7. Ningún jugador tiene permitido avanzar un argumento que fue previamente jugado por **O**.

8. Ningún jugador tiene permitido mover si el último argumento en la secuencia fue previamente jugado por **P.** (i.e., si la siguiente jugada corresponde a **O** y **P** ha repetido su jugada, **O** pierde; si la siguiente movida corresponde a **P** y **O** ha repetido un argumento jugado previamente por **P**, **P** pierde).

Ejemplo 2.5.9 Sea MA=⟨{*A, B*} {(*A, B*), (*B, A*)}⟩ un marco argumentativo. Supóngase que **P** quiere probar que *A* es admisible. El siguiente juego $A_{(P)} \leftarrow B_{(O)} \leftarrow A_{(P)}$ prueba que *A* es admisible.

Ejemplo 2.5.10 Sea MA=⟨{*A, B, C*} {(*B, A*), (*C, B*), (*A, C*)}⟩ un marco argumentativo. Supóngase que **P** quiere probar que *A* es admisible. El siguiente juego $A_{(P)} \leftarrow B_{(O)} \leftarrow C_{(P)} \leftarrow A_{(O)}$ prueba que *A* no es admisible.

2.6 Conclusión

En el presente capítulo se han caracterizado a los sistemas argumentativos en función de la respuesta a tres preguntas: ¿Cómo se construyen los argumentos?, ¿cómo estos pueden ser derrotados? y ¿cómo pueden ser defendidos en contra de los contraargumentos que los derrotan?

Las respuestas dadas a tales preguntas permitieron comprender que, independientemente del lenguaje formal utilizado, un argumento es construido a partir de un conjunto de información inicial. Una vez que tales argumentos son construidos, puede suceder que dos o más de ellos no puedan ser aceptados conjuntamente, cuando tal situación se da, se dice que los argumentos están en conflicto. Tres tipos de conflicto se han señalado como posibles: conflicto por rebatimiento, conflicto por socavamiento y ataque a premisas.

El conflicto únicamente señala la mutua incompatibilidad de los argumentos. La derrota es la relación que se define en función de la consideración de qué argumento es mejor. Se ha distinguido las derrotas que son dependientes de las independientes del criterio de preferencia. A su vez se ha brindado una tipología de las derrotas.

También se ha indicado que si un argumento es derrotado, tal situación no significa que ese argumento no prevalezca frente a sus adversarios. Por ello es necesario conocer el estado de todos los argumentos, dado que puede suceder que un argumento se vea restablecido.

Dos enfoques generales para conocer el estado final de un argumento se han presentado. Un enfoque semántico, basado en extensiones, que establece las condiciones que un conjunto de argumentos debe satisfacer para considerarlos como extensiones de la teoría. Otro procedimental, basado en un juego argumentativo o disputa, que permite probar si un argumento determinado es aceptable según determinado conjunto de reglas. Tales reglas caracterizan alguna semántica determinada.

Capítulo III: Algunos Sistemas Argumentativos

3.1 Introducción

En el capítulo anterior se presentaron las ideas y conceptos generales subyacentes a los sistemas basados en argumentos respondiendo a tres preguntas: ¿cómo se construyen los argumentos?, ¿cómo pueden ser derrotados? y ¿cómo pueden ser defendidos en contra de los contraargumentos que los derrotan? Las respuestas a tales preguntas permitieron reconocer a los sistemas argumentativos como formalismos que modelan información tentativa y potencialmente contradictoria mediante la construcción y comparación de argumentos rebatibles a favor o en contra de ciertas conclusiones.

Una vez construidos tales argumentos es posible que algunos de ellos no puedan ser aceptados conjuntamente, lo que exigirá algún mecanismo de selección o preferencia en caso de que lo haya. Aunque la comparación mediante alguna preferencia puede señalar qué argumento es mejor, esto no alcanza para determinar si tal argumento puede considerarse, finalizado el proceso de comparación, como un argumento justificado, i.e. un argumento que prevalece frente a todos sus rivales. Esto es así porque es necesario considerar otras formas de interacción entre argumentos, en particular, cómo se dan las derrotas entre ellos. De manera que un mecanismo que permita determinar cuándo un argumento puede ser considerado justificado es necesario. Tales nociones fueron caracterizadas de manera general en el capítulo anterior.

En el presente capítulo se presentarán algunos sistemas argumentativos concretos que instancian tales rasgos en una manera particular, i.e. definen a los argumentos de una manera determinada, establecen cómo se comparan los argumentos y estipulan un criterio específico a fin de dictaminar si un argumento en el sistema puede ser considerado como un *buen* argumento para la conclusión que sustenta. En el capítulo anterior puede verse una referencia más amplia a los diversos trabajos que se han desarrollado en el área de la argumentación rebatible.

Los sistemas que se presentarán a continuación fueron seleccionados en base a que cada uno representa algún aspecto característico y han recibido varios refinamientos con vistas a minimizar resultados inadecuados. Estos son el sistema *PS* debido a Prakken y Sartor (1996[b]), *BDKT* propuesto por Bondarenko, Dung, Kowalski y Toni (1997), *MTDR* planteado por Simari y Loui (1992), *DeLP*, un refinamiento del anterior sugerido por García y Simari (2004) *y Oscar* debido a Pollock (1995).

PS es un sistema capaz de modelar razonamientos rebatibles y fue propuesto originalmente para el contexto legal. Luego, el sistema *BDKT* propuesto en (Bondarenko et al., 1997) constituye un abordaje abstracto que permite diversas instanciaciones como casos del mismo sistema tales como la lógica default, programación lógica extendida, lógica modal no monotónica, lógica autoepistémica y Theorist. El sistema *MTDR* planteado por Simari y Loui (1992) es un sistema que, aunque inicialmente modela razonamiento default, goza de una particular característica de modularidad. A su vez, este sistema ha recibido varios refinamientos y se han establecido diversas relaciones con otros formalismos tales como sistemas multi-agentes, revisión de creencias, e implementaciones como en recomendación, toma de decisiones, entre otras tantas aplicaciones. Esta propuesta fue luego definida en el contexto de programación lógica, denominada *DeLP*. Finalmente la formulación

de Pollock, el proyecto *Oscar,* es tal vez el representante paradigmático de integración entre trabajos en teoría del conocimiento y computación (Pollock, 1995).

El capítulo se organiza como sigue. En la sección 3.2 se presentará el sistema *PS*. La sección 3.3 define el marco abstracto para razonamiento default propuesto en (Bondarenko et al., 1997). En la sección 3.4 se define el sistema *MTDR* formulado por Simari y Loui en (1992 y en la sección 3.5 un sistema de Programación lógica rebatible basado en *MTDR*, denominado *DeLP* (García y Simari, 1994). En la sección 3.6 se expone la teoría del razonamiento definida por Pollock en (1995). Finalmente se concluye haciéndose un repaso a los sistemas considerados.

3.2 El sistema *PS*

Prakken y Sartor en (Prakken y Sartor, 1995; 1996ª; 1996ᵇ y 1997) proponen un sistema argumentativo cuyos orígenes se remontan al trabajo de Prakken (1993). A continuación se presenta la versión expuesta en (Prakken y Sartor, 1996ᵇ), de ahí que sea denominado como el sistema *PS*. El sistema cuenta con dos versiones, una básica y una extendida. La diferencia entre ambas reside en que las prioridades son establecidas previamente en el sistema básico mientras que en el extendido, éstas son derivadas de las premisas.

El sistema es definido en un lenguaje formal subyacente dotado de dos tipos de negaciones. Una negación clásica y otra débil. Una formula atómica de primer orden es un literal positivo (*positive literal*); un literal positivo precedido por '¬' es un literal negativo (*negative literal*). Un literal positivo o negativo será un literal débil (*weak literal*) si se encuentra precedido por '∼', donde '∼' es una negación débil. De lo contrario, se estará en presencia de un literal fuerte (*strong literal*). Para cualquier

fórmula atómica *P(x), P(x) y ¬P(x)* son *complementos* una de otra. En el metalenguaje $\bar{L}$ denota el complemento de *L*.

Si *L* es un literal, ∼*L* puede entenderse como *"no existe evidencia que L sea el caso"* mientras que ¬*L* expresa la idea de que *L*, definitivamente, no es el caso. La negación débil '∼' es empleada para representar suposiciones (*assumptions*). A partir de los literales es posible construir reglas. Una *regla* es una expresión de la forma

$$r\colon L_0 \wedge \ldots \wedge L_j \wedge \sim L_k \wedge \ldots \wedge \sim L_m \Rightarrow L_n$$

donde *r* es el nombre de la regla y cada L_i, $0 \leq i \leq m$, es un literal fuerte. Las conjunciones a la izquierda de la flecha es el *antecedente* de la regla, mientras que a la derecha se encuentra el *consecuente* de la regla. Una regla con variables libres denota al conjunto de todas sus instancias.

La base de conocimiento o teoría está constituida por dos conjuntos disjuntos. El conjunto S de reglas estrictas y el conjunto D de reglas rebatibles. Las reglas en S no contienen literales débiles. Para simplificar, se empleará el conectivo '⇒' para las reglas estrictas y '→' para las rebatibles. La base de conocimiento se complementa con una relación predefinida de preferencia entre reglas, representada a través de un orden estricto parcial. De modo que dada la teoría T = {S, D}, una *teoría ordenada* es una par (T, <), si *r*<*r'* entonces la regla *r'* es preferida a *r*.

A partir de la base de conocimiento se construyen argumentos. Los argumentos se caracterizan mediante el encadenamiento sucesivo de reglas sugiriendo la idea de que una proposición cuenta con una *prueba* en el lenguaje del sistema. Formalmente

Definición 3.2.1 **[Prakken y Sartor, 1996[b]]**
(Argumento)

Un *argumento A* es una secuencia finita [r_n,..., r_m] de instancias básicas de reglas tal que:

1. Para cada *i*, $n \leq i \leq m$, se verifica que para cada literal *L* en el antecedente de r_i, existe un $j<i$, tal que *L* es el consecuente de r_j, y
2. Ningún r_i tiene como consecuente al consecuente de algún r_j ($j<i$).

Un argumento *A* se dice basado en una teoría (T,<) si y sólo si todas las reglas de *A* son instancias básicas de las reglas en T. *A'* será un *sub-argumento* (*propio*) de *A* si y sólo si *A'* es una sub-secuencia (propia) de *A*. Un literal *L* es una *conclusión* de *A* si y sólo si *L* es una consecuencia de alguna regla en *A*. *L* es una *suposición* (*assumption)* de *A* si y sólo si $\sim\overline{L}$ ocurre en alguna regla en *A*.

□

Para cualquier teoría ordenada Γ se denotará el conjunto de todos los argumentos en base a Γ como '$Args_\Gamma$'.

Definición 3.2.2 **[Prakken y Sartor, 1996[b]]**

(ataque)

Un argumento A_1 *ataca*, o es un *contraargumento de* un argumento A_2 si y sólo si la conclusión de A_1 es el complemento de alguna conclusión o suposición de A_2. Si un argumento ataca a otro, se dice que ambos están en conflicto.

□

La definición anterior permite la obtención de argumentos que se atacan a sí mismos, de ahí la necesidad de definir la noción de argumento coherente.

Definición 3.2.3 **[Prakken y Sartor, 1996[b]]**

(argumento coherente):

Un argumento *A* se dice *coherente* si y sólo si *A* no se ataca a sí mismo.

□

La idea de argumento coherente puede ser generalizada a conjuntos de argumentos que no se ataquen entre ellos. Esto es capturado en la siguiente definición.

Definición 3.2.4 **[Prakken y Sartor, 1996[b]]**

(conjunto libre de conflictos)

Un conjunto *Args* de argumentos se dice *libre de conflictos* si y sólo si ningún argumento en *Args* ataca otro argumento en *Args*.

□

Una vez determinado cuando dos argumentos están en conflicto es menester dirimir la cuestión y determinar si alguno es mejor que otro mediante algún mecanismo de comparación. Claramente, que un argumento sea mejor que otro no asegurará que éste será un argumento ganador, la derrota sólo señalará qué argumento es mejor entre dos argumentos individuales (y sus subargumentos).

Definición 3.2.5 **[Prakken y Sartor, 1996[b]]**

(rebatimiento y socavamiento)

Dados dos argumentos A_1 y A_2, se dice que

- A_1 *rebate* a A_2 si y sólo si existe un par de reglas r_1y r_2 ($r_1 \in A_1$; $r_2 \in A_2$) tal que *i).* r_1y r_2 tienen consecuentes complementarios y *ii).* $r_1 \prec r_2$.
- A_1 *socava (undercuts)* a A_2 si y sólo si alguna conclusión de A_1 es el complemento de alguna suposición de A_2.

□

La derrota combina la definición 3.2.5 con la idea de que un argumento incoherente es derrotado por el conjunto vacío.

Definición 3.2.6 **[Prakken y Sartor, 1996[b]]**

(derrota)

Dados dos argumentos A_1 y A_2, se dice que A_1 *derrota a* A_2 si y sólo si:

- A_1 es vacío y A_2 es incoherente, o
- A_1 socava a A_2, o
- A_1 rebate a A_2 y A_2 no socava a A_1.

Se dirá que A_1 *estrictamente derrota* A_2 si y sólo si A_1 derrota a A_2 y A_2 no derrota a A_1.

□

Comparar pares de argumentos no es suficiente para establecer qué argumento puede considerarse ganador en una disputa. Será necesaria una definición que determine el estado de un argumento en la base de todas las formas de interacción entre los argumentos. En particular la definición debe permitir el restablecimiento de ciertos argumentos cuando sus derrotadores sean a su vez derrotados. El estado de un argumento, finalizado el proceso de comparación, puede ser de tres tipos: *argumentos ganadores, perdedores* y *no decididos*.

La definición de argumento ganador, i.e., *argumento justificado* captura la siguiente idea intuitiva. El conjunto de argumentos justificados (*JustArgs*) se construye paso a paso, reuniendo primero en un conjunto $JustArgs_1$ todos los argumentos no derrotados. Luego, se añaden todos los argumentos que resultan justificados indirectamente a partir de $JustArgs_1$ obteniendo un conjunto $JustArgs_2$. Este proceso se repite hasta obtener un punto fijo $JustArgs_n$, el conjunto de todos los argumentos justificados.

Definición 3.2.7 **[Prakken y Sartor, 1996[b]]**

(argumento aceptable)

Un argumento *A* se dice *aceptable* con respecto a un conjunto *Args* de argumentos si y sólo si, cada argumento que derrota a *A*, es estrictamente derrotado por algún argumento contenido en *Args*.

□

Definición 3.2.8 **[Prakken y Sartor, 1996[b]]**

(conjunto $JustArgs_\Gamma$)

Sea Γ una teoría ordenada. Entonces se define la siguiente secuencia de subconjuntos de $Args_\Gamma$:

- $F_\Gamma^0 = \varnothing$
- $F_\Gamma^{i+1} = \{A \in Args_\Gamma \mid A \text{ es aceptable c.r. a } F_\Gamma^i\}$

Entonces el conjunto $JustArgs_\Gamma$ en base de Γ es $\bigcup_{i=0}^{\infty}(F_\Gamma^i)$.

□

Los argumentos restantes, junto con los justificados, pueden clasificarse en base a la siguiente definición:

Definición 3.2.9 **[Prakken y Sartor, 1996[b]]**

(estado final de un argumento)

Para cualquier teoría ordenada Γ usada como base para obtener *A*, se dirá que:

- *A* está *justificado* si y sólo si $A \in JustArgs_\Gamma$
- *A* está *denegado* si y sólo si *A* es atacado por $JustArgs_\Gamma$
- *A* es *defendible* si y sólo si *A* no está justificado ni denegado.

Para cualquier *L* se dirá, respectivamente, que *L* es una conclusión justificada o defendible o denegada si y sólo si *L* es una conclusión de un argumento justificado, o

no está justificada y es una conclusión de un argumento defendible, o ni es justificada ni defendible y es una conclusión de un argumento denegado.
Las siguientes son propiedades del conjunto de argumentos justificados. La primera establece que el conjunto de argumentos justificados es único, libre de conflictos y que cualquier sub-argumento de un argumento justificado es a su vez un argumento justificado.

Proposición 3.2.1: Sea $JustArgs_{\Gamma} = \bigcup_{i=0}^{\infty}(F_{\Gamma}^{i})$. Entonces para todo *i*, $F_{\Gamma}^{i} \subseteq (F_{\Gamma}^{i+1})$

□

Proposición 3.2.2: Para toda teoría ordenada Γ, $JustArgs_{\Gamma}$ es libre de conflicto.

□

Proposición 3.2.3: Si un argumento está justificado sobre la base de Γ, entonces todos sus subargumentos están justificados sobre la base de Γ.

□

Hasta aquí, se ha supuesto que existe un orden preestablecido entre las reglas de manera tal que conforman una teoría ordenada. Prakken y Sartor, propusieron adicionalmente, un mecanismo donde el orden es derivado de las premisas. Para ello incorporan una relación binaria que permite establecer un orden entre las premisas en el lenguaje objeto notado mediante '$\prec$'. El nuevo modelo requiere la modificación de la definición de argumentos que rebaten o socavan a otros a fin de que el ordenamiento no sea circular.

Definición 3.2.10 **[Prakken y Sartor, 1996[b]]**
(rebatimiento y socavamiento II)

Sean A_1 y A_2 dos argumentos. Entonces:

- A_1 *rebate* a A_2 si y sólo si:
  - O bien para algún par de reglas r_1 y r_2 ($r_1 \in A_1$; $r_2 \in A_2$) tal que *i*). r_1 y r_2 tienen consecuentes complementarios y *ii*). $r_1 \not\prec r_2$;
  - O bien para alguna secuencia de reglas $r_1,\ldots,r_n$ en A_1 y alguna regla r_m en A_2 *i*). El consecuente de r_m es $x \prec y$, y los consecuentes de $r_1,\ldots,r_n$ definen una cadena $y \prec z,\ldots,z' \prec x$; y *ii*). Para todo r_i ($1 \leq i \leq n$): $r_i \not\prec r_m$.
- A_1 *socava* a A_2 si y sólo si
  - alguna conclusión de A_1 es el complemento de alguna suposición de A_2; o
  - A_2 contiene alguna suposición $x \prec y$ y A_1 tiene una cadena de conclusiones $y \prec z,\ldots,z' \prec x$.

□

Las extensiones del sistema propuesto por Prakken y Sartor se caracterizan por una definición de punto fijo. La idea de que el conjunto de argumentos justificados se construye paso a paso se conserva y tiene su versión en el sistema extendido. Para caracterizar los argumentos aceptables será necesaria la siguiente noción:

Definición 3.2.11 **[Prakken y Sartor, 1996[b]]**

Sea *Args* un conjunto de argumento. Entonces

$$<_{Args} = \{r<r' \mid r \prec r' \text{ es una conclusión de algún } A \in Args\}$$

Se dirá que para cualquier conjunto *Args* de argumentos y cualquier conjunto T de reglas, que *A Args-derrota* (estrictamente) a *B* sobre la base de T si y sólo si *A* derrota (estrictamente) a *B* sobre la base de (T, $<_{Args}$). Se empleará una noción análoga para *Args-rebate.*

□

A partir de la noción de *Args-derrota* se entenderá que un argumento es aceptable cuando se verifique la propiedad expresada en la siguiente definición.

Definición 3.2.12 **[Prakken y Sartor, 1996[b]]**

(argumento aceptable II)

Un argumento se dice *aceptable* con respecto a un conjunto *Args* de argumentos si y sólo si todos los argumentos que *Args-derrotan* a *A* son estrictamente *Args-derrotados* por algún argumento en *Args.*

□

La aceptabilidad depende de la prioridad de las conclusiones de los argumentos en *Args*. Con esta modificación, la definición de argumentos justificados resulta similar a la definida anteriormente, excepto en que ahora Γ es un conjunto de reglas.

Definición 3.2.13 **[Prakken y Sartor, 1996[b]]**

(Conjunto *JustArgs*$_\Gamma$)

Sea Γ una teoría ordenada. Entonces se define la siguiente secuencia de subconjuntos de $Args_\Gamma$:

- $G_\Gamma^0 = \varnothing$
- $G_\Gamma^{i+1} = \{A \in Args_\Gamma \mid A \text{ es aceptable c.r. a } G_\Gamma^i\}$

Entonces el conjunto $JustArgs_\Gamma$ en base de Γ es $\bigcup_{i=0}^{\infty}(G_\Gamma^i)$.

□

Las propiedades del sistema extendido son análogas a las del sistema base.

Proposición 3.2.4: Sea $JustArgs_\Gamma = \bigcup_{i=0}^{\infty}(G_\Gamma^i)$.Entonces para todo *i*, $G_\Gamma^i \subseteq (G_\Gamma^{i+1})$.

□

Proposición 3.2.5: Para toda teoría ordenada Γ, $\cup_{i=0}^{\infty}(G_{\Gamma}^{i})$ es libre de conflicto.

□

3.3 El sistema *BDKT*

Bondarenko, Dung, Kowalski y Toni en (1997) propusieron un enfoque abstracto para modelar razonamiento default basado en argumentación. Tal enfoque incluyen varios formalismos como casos especiales, ejemplos de ellos son lógica default, programación lógica extendida, lógica modal no monotónica, lógica autoepistémica y el sistema Theorist.

Los argumentos son vistos como conjuntos de suposiciones que pueden ser agregadas a una teoría permitiendo derivar conclusiones que no pueden obtenerse considerando únicamente a la teoría.

Definición 3.3.1 **[Bondarenko et al., 1997]**

(marco basado en suposiciones)

Dado un sistema deductivo (L, R), un *marco basado en suposiciones* c.r. a (L, R) es una tupla $\langle T, Ab, {}^{-} \rangle$ donde

- $T, Ab \subseteq$ L
- $^{-}$ es un mapeo de Ab en L, donde $\overline{\alpha}$ denota lo contrario de α.

□

El sistema deductivo al que apela la definición 3.3.1 es entendido en (Bondarenko et al., 1997) como un par (L,R) donde L es un lenguaje formal consistente de sentencias contables y R es un conjunto de reglas de inferencia de la forma $\frac{\alpha 1,\ldots,\alpha n}{\alpha}$ donde α, α_1,…, $\alpha_n \in$ L y $n \geq 0$. Los axiomas lógicos, α, pueden ser representados como reglas de inferencia con n=0. Cualquier conjunto de sentencias T⊆L será llamado una teoría. La

teoría T expresa un conjunto de creencias dado y *Ab* es un conjunto de suposiciones que pueden ser usadas para extender T.

Una deducción a partir de T es una secuencia $\beta_1,\ldots, \beta_m$, donde $m>0$ tal que para cualquier $i = 1,\ldots, m$, $\beta_i \in T$ o existe una regla $\frac{\alpha 1,\ldots,\alpha n}{\beta}$ en R tal que $\alpha, \alpha_1,\ldots, \alpha_n \in \{\beta_1,\ldots, \beta_{i-1}\}$. $T \vdash \alpha$ significa que existe una deducción a partir de T cuyo último elemento es α. *Th*(T) es el conjunto $\{\alpha \in L: T \vdash \alpha\}$. En (Bondarenko et al., 1997: 69) se señala que el sistema deductivo (L,R) es *compacto* y *monotónico.*

La noción de derrota es definida en función de conjunto de suposiciones.

Definición 3.3.2 **[Bondarenko et al., 1997]**
(derrota)

Dado un marco basado en suposiciones $\langle T, Ab, \bar{} \rangle$,

- Un conjunto de suposiciones $\Delta \subseteq Ab$ *derrota una suposición* $\alpha \in Ab$ si y sólo si $T \cup \Delta \vdash \bar{\alpha}$
- Un conjunto de suposiciones $\Delta \subseteq Ab$ *derrota un conjunto de suposiciones* Δ' si y sólo si Δ derrota alguna suposición $\alpha \in \Delta'$.

□

Como consecuencia inmediata de la definición anterior se obtiene que dado un conjunto de suposiciones $\Delta \subseteq Ab$, si Δ es libre de conflicto entonces Δ no se derrota a sí misma.

Definición 3.3.3 **[Bondarenko et al., 1997]**
(conjunto libre de conflicto)

Dado un marco basado en suposiciones $\langle T, Ab, \overline{\ } \rangle$ y $\Delta \subseteq Ab$,

- Δ es *libre de conflicto* si y sólo si para cualquier $\alpha \in Ab$, $T \cup \Delta \nvdash \alpha, \overline{\alpha}$
- Δ es *libre de conflicto maximal* si y sólo si Δ es libre de conflicto y no existe un Δ' libre de conflicto tal que $\Delta' \supset \Delta$.

□

A partir de las nociones anteriores se define la semántica ingenua (*naive*) del sistema. La siguiente constituye una semántica menos tolerante:

Definición 3.3.4 **[Bondarenko et al., 1997]**

(semántica estable)

Un conjunto de suposiciones Δ es *estable* si y sólo si:

- Δ es cerrado, i.e., $\Delta = \{\alpha \in Ab : T \cup \Delta \vdash \alpha\}$
- Δ no se derrota a sí mismo
- Δ derrota cada suposición $\alpha \notin \Delta$

□

Buscando una semántica menos liberal que la ingenua y más tolerante que la estable, proponen la semántica admisible (*admissibility semantics*) propuesta inicialmente en (Dung, 1995)

Definición 3.3.5 **[Bondarenko et al., 1997]**

(semántica admisible)

Un conjunto de suposiciones cerrado $\Delta \subseteq Ab$ es *admisible* si y sólo si

- Δ no se ataca a sí mismo, y
- Para cada conjunto de suposiciones cerrado$\Delta' \subseteq Ab$, si Δ' derrota Δ entonces Δ derrota a Δ'.

Un conjunto de suposiciones que sea admisible maximal será considerado como el conjunto preferido.

Definición 3.3.6 **[Bondarenko et al., 1997]**

(semántica preferida)

Un conjunto de suposiciones $\Delta \subseteq Ab$ es *preferido* si y sólo si Δ es admisible maximal (c.r. a $\subseteq$).

□

Definición 3.3.7 **[Bondarenko et al., 1997]**

(defensa)

Un conjunto de suposiciones Δ *defiende* una suposición α si y sólo si para cada conjunto de suposiciones cerrado Δ', si Δ' derrota a α entonces Δ derrota a $\Delta' - \Delta$.

□

Definición 3.3.8 **[Bondarenko et al., 1997]**

Dado un marco basado en suposiciones $\langle T, Ab, \overline{\ } \rangle$ y un conjunto de suposiciones $\Delta \subseteq Ab$, $Def(\Delta) = \{\alpha | \Delta \text{ defiende a } \alpha\}$

□

El siguiente teorema se sigue de las definiciones anteriores:

Teorema 3.3.1 **[Bondarenko et al., 1997]**

Un conjunto de suposiciones Δ es *admisible si y sólo si* Δ es cerrado, y $\Delta \subseteq Def(\Delta)$.

□

Mientras que un conjunto de suposiciones cerrado es *admisible* siempre y cuando esté contenido en el conjunto de suposiciones que lo defiende, es *completo* cuando y sólo cuando es idéntico al conjunto que lo defiende.

Definición 3.3.9 **[Bondarenko et al., 1997]**

(conjunto completo)

Un conjunto de suposiciones Δ es *completo* si y sólo si Δ es cerrado, y $\Delta = Def(\Delta)$.

□

La semántica escéptica puede definirse como la intersección de los conjuntos de suposiciones sancionados por la extensión completa. De ahí la definición de conjunto bien formado.

Definición 3.3.10 **[Bondarenko et al., 1997]**

(conjunto bien formado)

Un conjunto de suposiciones Δ es *bien formado* si y sólo si Δ es la intersección de todos los conjuntos de suposiciones completos.

□

En (Kowalski y Toni, 1996) se generaliza el enfoque propuesto en (Bondarenko et al., 1997) a fin de poder definir programas que contengan reglas rebatibles de la forma

$$r: A \leftarrow B_1,\ldots,B_n$$

donde '*r*' funciona como una etiqueta de la regla $A \leftarrow B_1,\ldots,B_n$. $A \leftarrow B_1,\ldots,B_n$ es una regla sujeta a restricciones donde '$B_1,\ldots,B_n$' constituyen razones para *A*. Las reglas se encuentran ranqueadas mediante expresiones de la forma:

$$r < r' \Leftarrow$$

que señalan que *r'* tiene prioridad sobre *r*.

Kowalski y Toni no presentan un enfoque para programación lógica rebatible en sí. Sugieren más bien un esquema general para transformar cualquier programa lógico rebatible a un programa lógico ordinario sin contener reglas rebatibles aunque sí empleando los predicados adicionales: '*holds', 'defeated'* y '*conflict'*. El programa lógico ordinario al que el programa rebatible es transformado posee una precisa y adecuada manera de representar el significado del programa original. La idea es simple, cada regla rebatible de la forma:

$$r: A \leftarrow B_1,\ldots,B_n$$

es reemplazada por dos reglas estrictas dotadas de predicados adicionales de la siguiente manera:

$$A \Leftarrow holds(r),$$

$$holds(r) \Leftarrow B_1,\ldots,B_n,\ \sim defeated(r)$$

intuitivamente, el consecuente de la regla *r* puede ser aceptado si la regla *r* "*holds*" y la regla *r* "*holds*" si el antecedente de la regla *r* puede ser establecido y *r* no es derrotada.

La derrota es caracterizada a partir de la siguiente regla:

$$defeated(r) \Leftarrow r<r',\ conflict(r,r'),\ holds(r')$$

una regla *r* es derrotada si existe una regla *r'* de mayor prioridad en conflicto que puede mostrarse que *r' "holds*". Una regla está en conflicto con otra cuando la conclusión de ambas es complementaria:

$$conflict(r,r') \Leftarrow conclusion(r,A),\ conclusion(r',\neg A)$$

Obviamente, si *r* y *r'* están en conflicto:

$$conflict(r,r') \Leftarrow conflict(r',r)$$

Con vistas a ilustrar la idea, puede modelarse al ejemplo de Tweety de la siguiente manera:

Ejemplo 3.3.1:

$$Pt\Longleftarrow$$

$$Bt\Longleftarrow Pt$$

Donde "$Pt\Longleftarrow$" representa que Tweety es un pingüino y "$Bt\Longleftarrow Pt$" que Tweety es un ave si es un pingüino. Las reglas rebatibles;

$$r_1\text{: } Ft\longleftarrow Bt$$

$$r_2\text{: } \neg Ft\longleftarrow Pt$$

proveen razones para creer que Tweety vuela dado que es un ave (r_1) y que no vuela dado que es un pingüino (r_2). Obviamente es posible establecer un ranqueo entre las reglas de la siguiente manera:

$$r_1 < r_2 \Longleftarrow$$

Siguiendo a Kowalski y Toni (1996) este programa rebatible puede ser transformado en un programa lógico ordinario por remplazo de r_1 y r_2 de la siguiente manera:

$$r_1\text{: } Ft\Longleftarrow holds(r_1);\ holds(r_1) \Longleftarrow Bt,\ \sim defeated(r_1)$$

$$r_2\text{: } \neg Ft\Longleftarrow holds(r_2);\ holds(r_2) \Longleftarrow Pt,\ \sim defeated(r_2)$$

La relación de conflicto entre r_1 y r_2

$$conflcit(r_1,r_2) \Longleftarrow$$

Atendiendo al nuevo programa es posible preguntarse sobre cuál es la conclusión que el sistema puede sostener. Para ello es requerido saber cuál es el conjunto de suposiciones admisible. El primer paso será establecer qué conjuntos de suposiciones habrá en base a la información considerada, enfocando la atención a la conclusión $\sim defeated(r)$. Además del conjunto vacío existen tres conjunto de suposiciones:

$$\Delta_1 = \{\sim defeated(r_1)\}$$

$$\Delta_2 = \{\sim defeated(r_2)\}$$

$$\Delta_3=\{\sim defeated(r_1), \sim defeated(r_2)\}$$

El conjunto Δ_3 viola la cláusula de consistencia en la noción de admisibilidad ya que es un conjunto que se derrota a sí mismo en el sentido de que permite obtener las conclusiones Ft y $\neg Ft$. Por su parte, el conjunto Δ_1 es atacado por Δ_2 pero no se puede defender de Δ_2. Δ_2 por su parte es consistente y se defiende de todos los ataques que recibe, $\neg Ft$ es la conclusión justificada, como es esperable.

3.4 El sistema *MTDR*

Simari y Loui (1992) propusieron un sistema basado en argumentos para razonamiento derrotable combinando el criterio de preferencias propuesto en (Poole, 1985) y la justificación por niveles definida en (Pollock, 1987). Luego ha recibido sucesivas modificaciones con vistas a evitar algunos inconvenientes que el sistema padecía. El sistema recibe el nombre de *MTDR* porque estas son las siglas del trabajo original: *A mathematical treatment of defeasible reasoning.*

La propuesta consiste en construir un sistema formal capaz de modelar razonamiento rebatible. El lenguaje está compuesto por un lenguaje de primer orden *L* al que se le agrega una relación binaria metalingüística "$\succ\!\!-$". Las reglas de inferencia asociadas a *L* son modus ponens e instanciación. Los miembros de la relación "$\succ\!\!-$" se denominan reglas rebatibles y tienen la forma:

$$A_1,\ldots, A_n \succ\!\!- B \quad n\geq 1$$

donde $A_1,\ldots, A_n$ son literales de *L*. "$A_1,\ldots, A_n \succ\!\!- B$" intuitivamente expresa: "las razones para creer en $A_1,\ldots, A_n$ proveen razones para creer en B" o simplemente: "$A_1,\ldots, A_n$ son razones para creer B". Los literales del antecedente de la regla se encuentran en conjunción. Todo nombre de variable que aparece en ambos lados de la regla es la

misma variable. Una *instancia* de una regla rebatible se obtiene por sustitución uniforme de cada variable por una constante de *L*.

El conjunto *Sent*(*L*) de sentencias de *L* es el conjunto de fórmulas bien formadas cerrado en *L*. Este conjunto puede ser particionado en dos subconjuntos $Sent_N(L)$ y $Sent_C(L)$, correspondiendo a la información *necesaria* y *contingente*. La información *necesaria* simplemente está constituida por sentencias de variables libres o implicaciones en *L*, mientras que la información *contingente* está constituida por literales instanciados:

$$Sent(L) = Sent_C(L) \cup Sent_N(L)$$

Obviamente:

$$Sent_C(L) \cap Sent_N(L) = \varnothing$$

El conocimiento de un agente *a* es representado por un par (K, Δ), donde es K un subconjunto de *Sent*(*L*) y Δ es un conjunto finito de reglas rebatibles.

El par (K, Δ) será llamado una estructura lógica rebatible. K que representa la parte no rebatible del conocimiento de *a,* será denominado el contexto*.* Δ representa la información tentativa. K puede ser particionado en dos subconjuntos:

$$K_N = Sent_N(L) \cap K, \qquad K_C = Sent_C(L) \cap K$$

Claramente $K = K_N \cup K_C$. La única condición en K es la consistencia, i.e., $K \nvdash \perp$.

Dado un miembro *A* de *Sent*(*L*) y un conjunto $\Gamma=\{A_1, A_2,..., A_n\}$, donde cada A_i es miembro de K o una instancia de Δ, se establecerá una relación meta-metalingüística $\sim$, llamada consecuencia rebatible, entre Γ y *A* de la siguiente forma: Una *fbf A* será llamada una '*consecuencia rebatible*' de Γ si y sólo si existe una secuencia $B_1,..., B_m$ tal

que $A = B_m$ y para cada i, o bien B_i es un axioma de L, o B_i es en Γ, o B_i es una consecuencia directa de los miembros precedentes de la secuencia usando modus ponens o instanciación de una sentencia cuantificada universalmente. Se usará $\Gamma \vdash A$ como una abreviación de "A es una consecuencia rebatible de Γ".

La noción elemental del sistema *MTDR* es la de '*estructura de argumento*' y será presentada formalmente en la definición 3.4.1. Con vistas a facilitar las definiciones se introducirá el conjunto $\Delta^{\downarrow}$ que representa a todas las instancias de los miembros de Δ.

Definición 3.4.1 **[Simari y Loui, 1992]**

(argumento)

Dado una teoría rebatible (K,Δ), un subconjunto T de $\Delta^{\downarrow}$ será denominado un *argumento* para $h \in Sent_C(L)$ en el contexto K, denotado por $\langle T, h \rangle$ si y sólo si:

i. $K \cup T \vdash h$,

ii. $K \cup T \not\vdash \bot$, y

iii. $\nexists T' \subset T$, $K \cup T' \vdash h$.

□

La primera condición establece que el conjunto de reglas rebatibles permite derivar rebatiblemente la conclusión h en unión con K. La condición *ii*. exige que el argumento sea consistente. La condición *iii*. pide que el argumento sea minimal.

Ejemplo 3.4.1: Sean K = {P(a), Q(a)} y Δ={P(x)$\succ$R(x), Q(x)$\wedge$R(x)$\succ$H(x), M(x)$\succ$N(x)} el contexto y el conjunto de reglas rebatibles respectivamente. Por lo tanto, el subconjunto T de instancias de Δ, T={P(a)$\succ$R(a), Q(a)$\wedge$R(a)$\succ$ H(a)}, es una estructura de argumento para *H(a)*, i.e. $\langle T, H(a) \rangle$.

Definición 3.4.2 **[Simari y Loui, 1992]**

(subargumento)

Se dice que $\langle S, j\rangle$ es un *subargumento* de $\langle T, h\rangle$, notado como $\langle S, j\rangle \sqsubseteq \langle T, h\rangle$ si y sólo si $\langle T, h\rangle$ una estructura de argumento para h, y $\langle S, j\rangle$ una estructura de argumento para j tal que $S \subseteq T$. Si $S \subset T$ se dirá que S es un *subargumento propio* de T.

□

Ejemplo 3.4.2: Dado que $T = \langle \{P(a) \succ\!\!- R(a), Q(a) \wedge R(a) \succ\!\!- H(a)\}, H(a)\rangle$ es un argumento para H(a) y $S = \langle \{P(a) \succ\!\!- R(a)\}, R(a)\rangle$ es un argumento para R(a) y se verifica la propiedad de que $\{P(a) \succ\!\!- R(a)\} \subseteq \{P(a) \succ\!\!- R(a), Q(a) \wedge R(a) \succ\!\!- H(a)\}$ entonces S es un subargumento de T.

Diversos tipos de interacciones entre argumentos pueden ser definidos. Una de ellas es la de *subargumentación*. A continuación se definen otras.

Definición 3.4.3 **[Simari y Loui, 1992]**

(desacuerdo)

Dos estructuras de argumento, $\langle T_1, h_1\rangle$ y $\langle T_2, h_2\rangle$, están en *desacuerdo* si y sólo si: $K \cup \{h_1, h_2\} \vdash \bot$.

□

Ejemplo 3.4.3: Dados los siguientes argumentos

$$T: \langle \{A \succ\!\!- B\}, B\rangle \qquad S: \langle \{C \succ\!\!- \neg B\}, \neg B\rangle$$

Y atendiendo que K = {A, C}, se dice que T y S están en desacuerdo dado que $K \cup \{Con(T), Con(S)\} \vdash \bot$, donde Con(T) y Con(S) significa conclusión de T y S respectivamente.

Ejemplo 3.4.4: Dados los siguientes argumentos

$$T\text{: } \langle \{A \succ\!\!- B\}, B \rangle \qquad S\text{: } \langle \{C \succ\!\!- D\}, D \rangle$$

y teniendo en cuenta que K = {A, C, D→¬B}, T y S están en desacuerdo.

En el siguiente ejemplo puede observarse un caso en que no se da la relación de desacuerdo aunque intuitivamente ambos argumentos no pueden ser aceptados conjuntamente.

Ejemplo 3.4.5: Dados los siguientes argumentos

$$T\text{: } \langle \{A \succ\!\!- \neg B\}, \neg B \rangle \qquad S\text{: } \langle \{C \succ\!\!- B, B \succ\!\!- D\}, D \rangle$$

teniendo en cuenta que K = {A,C} son argumentos que no están en desacuerdo a pesar de que claramente son argumentos incompatibles dado que el argumento T sustenta una sentencia ¬B mientras que en S, se sustenta como un paso a B. Esta situación motivó la siguiente definición.

Definición 3.4.4 **[Simari y Loui, 1992]**

(contraargumentación)

Se dice que un argumento $\langle T_1, h_1 \rangle$ *contraargumenta* a $\langle T_2, h_2 \rangle$ si y sólo si existe un subargumento $\langle T, h \rangle$ de $\langle T_2, h_2 \rangle$ tal que $\langle T_1, h_1 \rangle$ y $\langle T, h \rangle$ están en desacuerdo, i.e. $K \cup \{h_1, h\} \vdash \bot$.

□

Atendiendo a la definición 3.4.4, es claro que en el ejemplo 3.4.5, T contraargumenta a S. En los ejemplos 3.4.3 y 3.4.4, T contraargumenta a S y viceversa.

Dos argumentos pueden compararse entre sí a través de la relación de especificidad, la cual establece un orden de preferencias entre ellos. Intuitivamente, un argumento

$\langle T_1,h_1\rangle$ es más específico que $\langle T_2,h_2\rangle$ cuando $\langle T_1,h_1\rangle$ está sustentado en mayor información, tal que cuando puede afirmarse $\langle T_2, h_2\rangle$ también puede afirmarse $\langle T_1,h_1\rangle$ pero existen caso en los que puede afirmarse $\langle T_1,h_1\rangle$ pero no $\langle T_2,h_2\rangle$, formalmente:

Definición 3.4.5 **[Simari y Loui, 1992]**

(especificidad)

Sean $\langle T_1, h_1\rangle$ y $\langle T_2, h_2\rangle$ dos argumentos. Se dice que $\langle T_1, h_1\rangle$ es *estrictamente más específico que* $\langle T_2, h_2\rangle$, notado como $\langle T_1, h_1\rangle \succ_{esp} \langle T_2, h_2\rangle$, si y sólo si:

1. $\forall e \in Sent_C(L)$ tal que $K_N \cup \{e\} \cup T_1 \vdash h_1$ y $K_N \cup \{e\} \nvdash h_1$, entonces $K_N \cup \{e\} \cup T_2 \vdash h_2$, y
2. $\exists e \in Sent_C(L)$ tal que:
   - $K_N \cup \{e\} \cup T_2 \vdash h_2$ *(activa T_2)*
   - $K_N \cup \{e\} \cup T_1 \nvdash h_1$ *(no activa T_1)*
   - $K_N \cup \{e\} \nvdash h_2$ *(condición de no trivialidad)*

□

La expresión '*activa*' empleada en la definición se usa con el siguiente significado: *conjuntamente con K_N el argumento T es suficiente para construir una derivación rebatible para h.*

La relación de derrota es un refinamiento de la noción de contraargumentación incorporando la especificidad como orden de preferencias entre argumentos:

Definición 3.4.6 **[Simari y Loui, 1992]**

(derrota)

Sean $\langle T_1, h_1\rangle$ y $\langle T_2, h_2\rangle$ dos argumentos. Se dice que $\langle T_1, h_1\rangle$ *derrota* a $\langle T_2, h_2\rangle$, notado como $\langle T_1, h_1\rangle \gg_{def} \langle T_2, h_2\rangle$, si y sólo si existe un subargumento $\langle T, h\rangle$ de $\langle T_2, h_2\rangle$ tal que

$\langle T_1, h_1 \rangle$ se encuentra en desacuerdo con $\langle T, h \rangle$ y $\langle T_1, h_1 \rangle$ es estrictamente más específico que $\langle T, h \rangle$.

□

Ejemplo 3.4.6 Dadas las siguientes estructuras de argumento

$$T: \langle \{A \wedge B \wedge E \succ\!\!- \neg C\}, \neg C \rangle \qquad S: \langle \{A \wedge B \succ\!\!- C, C \succ\!\!- D\}, D \rangle$$

La estructura $\langle \{A \wedge B \wedge E \succ\!\!- \neg C\}, \neg C \rangle$ contraargumenta a $\langle \{A \wedge B \succ\!\!- C, C \succ\!\!- D\}, D \rangle$ en C y $\langle \{A \wedge B \wedge E \succ\!\!- \neg C\}, \neg C \rangle$ es estrictamente más específico que $\langle \{A \wedge B \succ\!\!- C\}, C \rangle$ de modo que $T \gg_{def} S$.

Una vez que se han definido los mecanismos para detectar argumentos en conflicto y decidir entre pares de argumentos cuál es mejor, se hace necesario contar con un modo de determinar cuál o cuáles de los argumentos esgrimidos constituye un argumento justificado. Para determinar esto, Simari y Loui en (1992) proponen una justificación por niveles basado en las siguientes definiciones.

Definición 3.4.7 **[Simari y Loui, 1992]**

(argumentos activos por niveles)

1. Todos los argumentos en el nivel *0* están activos como argumentos de soporte (S-argumento) e interferencia (I-argumento).
2. Un argumento $\langle T_1, h_1 \rangle$ está activo en el nivel *n+1* como S-argumento si y sólo si no existe en el nivel *n* algún I-argumento $\langle T_2, h_2 \rangle$ tal que contraargumenta a $\langle T_1, h_1 \rangle$ en *h*.
3. Un argumento $\langle T_1, h_1 \rangle$ está activo en el nivel *n+1* como I-argumento si y sólo si no existe en el nivel *n* un I-argumento $\langle T_2, h_2 \rangle$ tal que $\langle T_2, h_2 \rangle$ derrote a $\langle T_1, h_1 \rangle$.

□

Definición 3.4.8 **[Simari y Loui, 1992]**

(justificación)

Una estructura de argumento $\langle T_1, h_1 \rangle$ constituye una *justificación* para h_1 si y sólo si existe un m a partir del cual para todo $n \geq m$, $\langle T_1, h_1 \rangle$ está activo en el nivel n como argumento de soporte.

□

Ejemplo 3.4.7: Sea KB = (K,Δ) una base de conocimiento, donde:

K= {pingüino_mágico(tweety);
pingüino_mágico(X) → pingüino(X),
pingüino(X) → ave(X)}

Δ = {ave(X) ≻ vuela(X);
pingüino_mágico(X) ≻ vuela(X);
pingüino(X) ≻¬vuela(X)}.

En base a KB, pueden formularse los siguientes argumentos:

T_1 ⟨{ave(tweety) ≻vuela(tweety)}, vuela(tweety)⟩

T_2 ⟨{pingüino(tweety) ≻¬vuela(tweety)}, ¬vuela(tweety)⟩

T_3 ⟨{pingüino_mágico(tweety) ≻ vuela(tweety)}, vuela(tweety)⟩

Considerando que T_2 resulta ser un contraargumento que es estrictamente más específico que T_1, y a su vez T_3 es un contraargumento más específico que T_2, los argumentos activos en cada nivel, como argumentos de soporte y de interferencia se encuentran ilustrados en la figura 3.4.1

Nivel	Soporte	Interferencia
0	T_1, T_2, T_3	T_1, T_2, T_3
1	-	T_3
2	T_1, T_3	T_1, T_3
3	T_1, T_3	T_1, T_3
…	*…*	*…*

Fig. 3.4. 1: Argumentos por niveles

Los argumentos T_1 y T_3 permanecen activos como argumentos de soporte en todo nivel $m \geq 2$. En consecuencia los argumentos T_1 y T_3 están justificados.

En (Simari et al., 1994) se presentó un refinamiento del sistema con vista a mejorar algunos aspectos y a evitar resultados inadecuados. Una de las incorporaciones fue la modificación de la noción de derrota.

Definición 3.4.8 **[Simari et al., 1994]**

(derrotador propio y de bloqueo)

$\langle T_1,h_1\rangle$ *derrota* a $\langle T_2,h_2\rangle$, notado como $\langle T_1,h_1\rangle \gg_{\text{def}} \langle T_2,h_2\rangle$, si y sólo si existe un subargumento $\langle T,h\rangle$ de $\langle T_2,h_2\rangle$ tal que $\langle T_1,h_1\rangle$ contraargumento $\langle T_2,h_2\rangle$ en h y se verifica que:

i. $\langle T_1,h_1\rangle$ es estrictamente más específico que $\langle T,h\rangle$, o bien
ii. $\langle T_1,h_1\rangle$ no se relaciona por especificidad con $\langle T,h\rangle$.

Si $\langle T_1,h_1\rangle$ derrota a $\langle T_2,h_2\rangle$ por *i* se dice que $\langle T_1,h_1\rangle$ es un *derrotador propio* de $\langle T_2,h_2\rangle$ pero si $\langle T_1,h_1\rangle$ derrota a $\langle T_2,h_2\rangle$ por *ii* se dice que $\langle T_1,h_1\rangle$ es un *derrotador por bloqueo* de $\langle T_2,h_2\rangle$.

□

Otra modificación importante fue realizada en términos de la justificación optando por un mecanismo dialéctico considerablemente más simple.

Definición 3.4.9 **[Simari et al., 1994]**

(árbol dialéctico)

Sea $\langle T, h\rangle$ una estructura de argumento. Un *árbol dialéctico para* $\langle T, h\rangle$, denotado como $T_{\langle T,h\rangle}$, se construye de la siguiente manera:

- Si $\langle T, h\rangle$ no posee derrotadores propios o por bloqueo, el árbol asociado se compone de un único nodo conteniendo a $\langle T,h\rangle$.
- Si $\langle T, h\rangle$ posee como derrotadores propios o por bloqueo a los argumentos $\langle T_1, h_1\rangle,\ldots,\langle T_n, h_n\rangle$ entonces el árbol asociado se obtiene colocando a los nodos raíces de los árboles dialécticos $T_{\langle T_1,h_1\rangle},\ldots,T_{\langle T_n,h_n\rangle}$ como sucesores de un nodo conteniendo a $\langle T, h\rangle$.

□

Ejemplo 3.4.8: El análisis dialéctico del ejemplo 3.4.7 puede observarse en la siguiente figura.

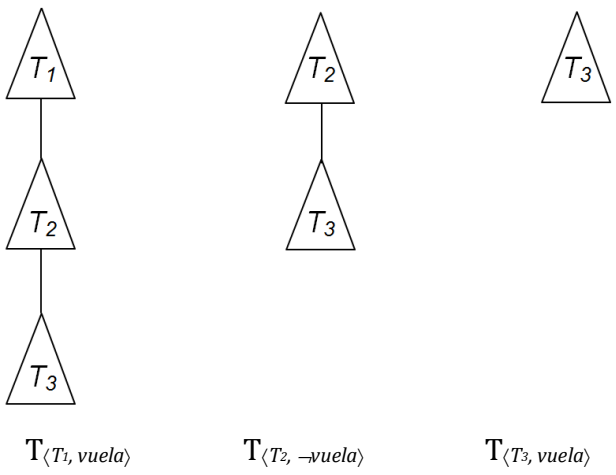


Fig. 3.4. 2: Análisis dialéctico del ejemplo 3.4.7

El análisis dialéctico de un argumento determinará si un argumento está justificado dependiendo del estado de sus derrotadores. La siguiente definición describe el procedimiento de marcado recursivo que permitirá conocer tal estado.

Definición 3.4.10 **[Simari et al., 1994]**

(marcado de un árbol dialéctico)

Sea $T_{\langle T,h\rangle}$ el árbol dialéctico asociado a $\langle T,h\rangle$. Los nodos en $T_{\langle T,h\rangle}$ pueden ser marcados como *derrotados* o *no derrotados* cuando se verifiquen las siguientes condiciones:

- Un nodo será marcado como *nodo derrotado* si y sólo si al menos un nodo sucesor está marcado como nodo no derrotado.
- Un nodo será marcado como *nodo no derrotado* si y sólo si todos sus sucesores están marcados como nodos derrotados.

□

La noción de argumento justificado puede hacerse ahora en términos del análisis dialéctico según la siguiente definición.

Definición 3.4.11 **[Simari et al., 1994]**

(justificación)

Sea $\langle T,h\rangle$ una estructura de argumento. $\langle T,h\rangle$ constituye una *justificación* para h si y sólo si es posible marcar al nodo raíz de su árbol dialéctico asociado $T_{\langle T,h\rangle}$ como nodo no derrotado.

□

Otras modificaciones han sido incorporadas al sistema propuesto originalmente. En particular en lo referido a evitar los ciclos. Dado a que las mismas se encuentran incorporadas en *DeLP* se omitirá su presentación.

3.5 Programación Lógica Rebatible (*DeLP*)

En (Gracía, 1997 y García, 2000) se propone un formalismo basado en (Simari y Loui, 1992 y Simari et al., 1994) denominado *DeLP* (*Defeasible Logic Programing*) donde se combina a la programación lógica con la argumentación rebatible a fin de modelar información tentativa y potencialmente contradictoria. A continuación se presentará el formalismo expuesto en (García y Simari, 2004).

En *DeLP* el lenguaje es definido a partir de tres conjuntos disyuntos: un conjunto de hechos, un conjunto de reglas estrictas y un conjunto de reglas rebatibles. Un literal '*L*' en *DeLP* será un átomo '*A*' o su negación '*~A*', donde '*~*' representa la negación fuerte (clásica). Los literales no tienen variables.

Un hecho es un literal, i.e. un átomo o su negación. Una regla estricta es un par ordenado denotado por '*head⟵body',* donde *'head'* es un literal y '*body'* es un conjunto finito no vació de literales. Una regla rebatible es un par ordenado *'head⟙body'* donde '*head'* es un literal y '*body'* es un conjunto finito no vacío de literales. Por la definición de literal, claramente, la negación fuerte puede ser empleada en *head*.

Las reglas estrictas representan información no rebatible como '*todos los pingüinos son aves'* mientras que las reglas rebatibles permiten representar información tentativa como *'por lo general las aves vuelan'.*

Una regla estricta con *body* vacío representa un hecho, mientras que una regla rebatible representará una presunción. En (Martínez, et al., 2012) se desarrolla un sistema de programación lógica que extiende a *DeLP* dotándolo de la capacidad de trabajar con presunciones denominado *PreDeLP*.

Definición 3.5.1 **[García y Simari, 2004]**

(programa lógico rebatible)

Un *programa lógico rebatible* $\mathcal{P}$, abreviado *DeLP*, es un conjunto posiblemente infinito de hechos, reglas estrictas y rebatibles. En un programa $\mathcal{P}$ se distinguirá el subconjunto Π de hechos y reglas estrictas y el subconjunto Δ de reglas rebatibles. Cuando sea requerido se denotará $\mathcal{P}$ como (Π,Δ).

□

A continuación se define la noción de derivación rebatible.

Definición 3.5.2 **[García y Simari, 2004]**

(Derivación rebatible)

Sea $\mathcal{P} = (\Pi, \Delta)$ un *delp* y L un literal. Una *derivación rebatible* para L a partir de $\mathcal{P}$, denotado como $\mathcal{P} \mid\sim L$, consiste de una secuencia finita $L_1, L_2,\ldots,L_n=L$ de literales, y para cada literal L_i en la secuencia se tiene que:

- L_i es un hecho en Π, o
- Existe una regla R_i en $\mathcal{P}$ (estricta o rebatible) con *head* L_i y *body* $B_1, B_2,\ldots,B_k$ y todos los literales de *body* es un elemento L_j de la secuencia aparecida antes de L_i (j<i).

□

Definición 3.5.3 **[García y Simari, 2004]**

(Derivación estricta)

Sea $\mathcal{P} = (\Pi,\Delta)$ un *delp* y L un literal con una derivación rebatible $L_1, L_2,\ldots,L_n=L$. Se dirá que L tiene una *derivación estricta* a partir de $\mathcal{P}$, denotado como $\mathcal{P} \vdash L$, si L es un hecho o todas las reglas usadas para obtener la secuencia $L_1, L_2,\ldots,L_n$ son reglas estrictas.

□

El símbolo ‘ ¯ ’ será usado para denotar el complemento de un literal *L* con respecto a la negación fuerte, i.e. $\bar{p}$ es ~p y $\overline{\sim p}$ es p. Dos literales son contradictorios si estos son complementarios.

Definición 3.5.4 **[García y Simari, 2004]**

(conjunto de reglas contradictorio)

Un conjunto de reglas es *contradictorio* si y sólo si existe una derivación rebatible para un par de literales complementarios a partir de tal conjunto.

□

DeLP, al igual que *MTDR,* se encuentra basado en la noción de estructura de argumento y en las relaciones de derrota entre ellos. A continuación se brindan las definiciones elementales del mismo.

Definición 3.5.5 **[García y Simari, 2004]**

(estructura de argumento)

Sea *h* un literal y $\mathcal{P}$ = (Π, Δ) un *delp*. Se dirá que ⟨*T, h*⟩ es una *estructura de argumento* si *T* es un conjunto de reglas rebatibles de Δ tal que:

i. Existe una derivación rebatible para h a partir de $\Pi \cup T$

ii. El conjunto $\Pi \cup T$ no es contradictorio

iii. *T* es minimal, no existe un subconjunto propio *T'* de *T* tal que *T'* satisfaga las condiciones i. y ii.

□

Definición 3.5.6 **[García y Simari, 2004]**

(estructura de subargumento)

Una estructura de argumento ⟨*T, h*⟩ es una *estructura de subargumento* de ⟨*S, k*⟩ si $T \subseteq S$.

Definición 3.5.7 **[García y Simari, 2004]**

(desacuerdo)

Sea $\mathcal{P}$ = (Π,Δ) un DeLP, se dirá que dos literales *h* y *h'* están en *desacuerdo* si y sólo si el conjunto $\Pi \cup \{h, h'\}$ es contradictorio.

□

Dos literales complementarios *p* y ~*p* están en desacuerdo para cualquier conjunto Π, dado que $\Pi \cup \{p, \sim p\}$ es contradictorio. Sin embargo, dos literales que no son complementarios podrían también estar en desacuerdo. Por ejemplo dado Π={(~*h* ⟵ *b*), (*h* ⟵ *a*)}, los literales *a* y *b* están en desacuerdo porque $\Pi \cup \{a, b\}$ es contradictorio. La noción de desacuerdo permitirá distinguir dos tipos de conflicto, directo e indirecto.

Definición 3.5.8 **[García y Simari, 2004]**

(contraargumento)

Se dirá que $\langle T_1, h_1 \rangle$ *contraargumenta* $\langle T_2, h_2 \rangle$ en el literal *h* si y sólo si existe un subargumento $\langle T, h \rangle$ de $\langle T_2, h_2 \rangle$ tal que h_1 y *h* están en desacuerdo.

Si $\langle T_1, h_1 \rangle$ contraargumenta a $\langle T_2, h_2 \rangle$ en el literal *h*, entonces *h* será llamado *punto de contraargumentación* y el subargumento $\langle T, h \rangle$ será llamado *subargumento de desacuerdo.*

□

Un contraargumento $\langle T_1, h_1 \rangle$ de $\langle T_2, h_2 \rangle$ puede atacar directamente la conclusión h_2 o un punto intermedio *h*. Esto permite distinguir entre ataque directo e indirecto. La figura 3.5.1 ilustra la idea.

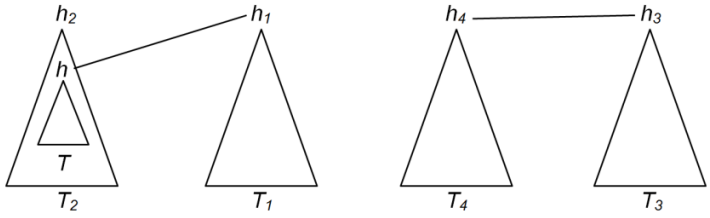


Fig. 3.5. 1: Ataque indirecto (izq.) y ataque directo (der.)

DeLP cuenta con dos maneras de comparar argumentos a fin de dirimir el desacuerdo. Uno basado en la noción de especificidad y el otro mediante prioridades entre reglas.

La noción de especificidad es capturada por la definición de especificidad generalizada que será expuesta a continuación. Intuitivamente, esta definición favorece dos aspectos en los argumentos: prefiere aquel que contenga mayor información o que use menos reglas, i.e. es más directo. De modo que un argumento será considerado mejor que otro si es más preciso y conciso.

Definición 3.5.9 **[García y Simari, 2004]**

(especificidad)

Sea $\mathcal{P}$ = (Π, Δ) un *delp* y sea Π_G el conjunto de todas las reglas estrictas en Π (sin incluir hechos). Sea $\mathcal{F}$ el conjunto de todos los literales para los que existe una derivación rebatible a partir de $\mathcal{P}$ ($\mathcal{F}$ será considerado un conjunto de hechos). Sea $\langle T_1, h_1\rangle$ *y* $\langle T_2, h_2\rangle$ dos estructuras de argumento obtenidas a partir de $\mathcal{P}$**.** Se dirá que $\langle T_1, h_1\rangle$ *es estrictamente más específico que* $\langle T_2, h_2\rangle$, notado como $\langle T_1, h_1\rangle \succ \langle T_2, h_2\rangle$, si se verifican las siguientes condiciones:

i. Para todo $H \subseteq \mathcal{F}$: Si $\Pi_G \cup H \cup T_1 \vdash h_1$ y $\Pi_G \cup H \not\vdash h_1$entonces $\Pi_G \cup H \cup T_2 \vdash h_2$, *y*

ii. Existe $H' \subseteq \mathcal{F}$ tal que $\Pi_G \cup H' \cup T_2 \vdash h_2$ y $\Pi_G \cup H' \not\vdash h_2$ y $\Pi_G \cup H' \cup T_1 \not\vdash h_1$

□

No es posible tener una derivación rebatible de un literal a partir de un conjunto de reglas sin hechos. Por lo tanto, del conjunto $\Pi_G \cup T_1$ no sería posible obtener una derivación rebatible para h_1, sin embargo, $\Pi_G \cup H \cup T_1$ sí porque H es un conjunto de literales. Por ello se dice que H activa a $\langle T_1,h_1\rangle$ o que H es un conjunto de activación para $\langle T_1, h_1\rangle$.

Definición 3.5.10 **[García y Simari, 2004]**

(equi-especificidad)

Dos argumentos $\langle T_1,h_1\rangle$ *y* $\langle T_2,h_2\rangle$ son *equi-específicos,* denotado como $\langle T_1,h_1\rangle \equiv \langle T_2,h_2\rangle$, si y sólo si $T_1 = T_2$, y el literal h_1 tiene una derivación estricta a partir de $\Pi \cup h_2$ y h_2 tiene una derivación estricta a partir de $\Pi \cup h_1$.

□

Dada la estructura modular de *DeLP*, es posible introducir mecanismos de comparación de argumentos que no requiera la noción de especificidad generalizada anteriormente. Puede ser implementada la comparación mediante el empleo de prioridades entre reglas de la siguiente manera:

Definición 3.5.11 **[García y Simari, 2004]**

(preferencia basada en reglas)

Sea $\mathcal{P}$ = (Π, Δ) un *delp* y > una relación de preferencia explícitamente definida entre las reglas rebatibles. Dadas dos estructuras de argumento $\langle T_1, h_1\rangle$ *y* $\langle T_2, h_2\rangle$, $\langle T_1, h_1\rangle$*será preferido a* $\langle T_2, h_2\rangle$ si:

i. Existe al menos una regla $r_a \in T_1$ y una regla $r_b \in T_2$ tal que $r_a > r_b$, y

ii. No existe un $r'_b \in T_2$ y $r'_a \in T_1$ tal que $r'_b > r'_a$

□

Atendiendo a la definición 3.5.9 y 3.5.11 es posible obtener un criterio de comparación más sofisticado combinándolos. Por ejemplo, considerando primero especificidad generalizada y si ningún argumento es preferido, usar prioridades existentes.

Dada una estructura de argumento $\langle T_1, h_1\rangle$ y un contraargumento $\langle T_2, h_2\rangle$ para $\langle T_1, h_1\rangle$ es posible comparar ambos argumentos y decidir cuál prevalece. Si el contraargumento $\langle T_2, h_2\rangle$ es mejor que $\langle T_1, h_1\rangle$ con respecto al criterio de comparación empleada entonces $\langle T_2, h_2\rangle$ será denominado un derrotador propio. Si ningún argumento es mejor, entonces se dirá que $\langle T_2, h_2\rangle$ es un derrotador por bloqueo. Si $\langle T_1, h_1\rangle$ es mejor que $\langle T_2, h_2\rangle$ entonces $\langle T_2, h_2\rangle$ no será considerado un derrotador para $\langle T_1, h_1\rangle$.

Como se señala, para definir cuándo un argumento derrota a otro, un criterio de comparación de argumentos es requerido. La derrota puede ser formulada independientemente del tipo de criterio empleada. Por ello, un criterio abstracto entre argumentos, notado como $\succ$, será empleada a continuación a fin de que con él se pueda estar refiriendo a especificidad generalizada o bien a otro criterio de comparación de argumentos.

Definición 3.5.12 **[García y Simari, 2004]**
(derrotador propio)

Dadas dos estructuras de argumento $\langle T_1, h_1\rangle$ *y* $\langle T_2, h_2\rangle$, $\langle T_1, h_1\rangle$ es un *derrotador propio* de $\langle T_2, h_2\rangle$ en el literal h si y sólo si existe un subargumento $\langle T, h\rangle$ de $\langle T_2, h_2\rangle$ tal que

i. $\langle T_1, h_1\rangle$ contraargumenta $\langle T_2, h_2\rangle$ en h, y

ii. $\langle T_1, h_1\rangle \succ \langle T, h\rangle$.

□

Definición 3.5.12 **[García y Simari, 2004]**

(derrotador por bloqueo)

Dadas dos estructuras de argumento $\langle T_1, h_1\rangle$ y $\langle T_2, h_2\rangle$, $\langle T_1, h_1\rangle$ es un *derrotador por bloqueo* de $\langle T_2, h_2\rangle$ en el literal h si y sólo si existe un subargumento $\langle T,h\rangle$ de $\langle T_2,h_2\rangle$ tal que $\langle T_1, h_1\rangle$ contraargumenta $\langle T_2, h_2\rangle$ en h, y $\langle T_1, h_1\rangle \not\succ \langle T, h\rangle$ y $\langle T, h\rangle \not\succ \langle T_1, h_1\rangle$.

□

Integrando ambas definiciones se obtiene la noción de derrota de la siguiente manera:

Definición 3.5.13 **[García y Simari, 2004]**

(derrota)

Dadas dos estructuras de argumento $\langle T_1, h_1\rangle$ *y* $\langle T_2, h_2\rangle$, $\langle T_1, h_1\rangle$ *derrota* a $\langle T_2, h_2\rangle$ si y sólo si

i. $\langle T_1, h_1\rangle$ es un derrotador propio de $\langle T_2, h_2\rangle$, o

ii. $\langle T_1, h_1\rangle$ es un derrotador por bloqueo de $\langle T_2, h_2\rangle$.

□

En orden a establecer si un argumento $\langle T_0,h_0\rangle$ es no derrotado, todos los derrotadores de $\langle T_0,h_0\rangle$ deben ser considerados. Supóngase que $\langle T_1,h_1\rangle$ es un derrotador de $\langle T_0,h_0\rangle$, pero $\langle T_1,h_1\rangle$ puede a la vez ser derrotado y así con otros. De esta manera, una secuencia de argumentos es creada, donde cada elemento de la secuencia derrota su predecesor. Esta idea puede ser formalizada así:

Definición 3.5.14 **[García y Simari, 2004]**

(Línea de argumentación)

Sea $\mathcal{P}$ = (Π,Δ) un *delp* y $\langle T_0, h_0\rangle$ una estructura de argumento obtenida a partir de $\mathcal{P}$. Una *línea de argumentación* para $\langle T_0, h_0\rangle$ es una secuencia de estructuras de

argumento a partir de $\mathcal{P}$, notado como $\Lambda=[\langle T_0, h_0\rangle, \langle T_1, h_1\rangle, \langle T_2, h_2\rangle, \langle T_3, h_3\rangle,\ldots]$, donde cada elemento de la secuencia $\langle T_i, h_i\rangle$, $i > 0$, es un derrotador de su predecesor $\langle T_{i-1}, h_{i-1}\rangle$.

□

Dado que una línea de argumentación puede resultar infinita se requiere la imposición de algunas restricciones.

Definición 3.5.15 **[García y Simari, 2004]**

(Argumentos de soporte e interferencia)

Sea $\Lambda=[\langle T_0, h_0\rangle, \langle T_1, h_1\rangle, \langle T_2, h_2\rangle, \langle T_3, h_3\rangle,\ldots]$ una línea de argumentación, se dirá que $\Lambda_S=[\langle T_0, h_0\rangle, \langle T_2, h_2\rangle, \langle T_4, h_4\rangle,\ldots]$ es *el conjunto de argumentos de soporte* y $\Lambda_I=[\langle T_1, h_1\rangle, \langle T_3, h_3\rangle, \langle T_5, h_5\rangle,\ldots]$ es *el conjunto de argumentos de interferencia.*

□

Dado un argumento $\langle T_0, h_0\rangle$, pueden existir varios derrotadores para $\langle T_0, h_0\rangle$ y cada uno de ellos generará una línea de argumentación diferente. En tal línea de argumentación, los derrotadores podrán tener más de un derrotador de manera que generarán más líneas de argumentación comenzando en $\langle T_0, h_0\rangle$. Por ello, se necesita un proceso que considere todas las posibles líneas. Adicionalmente se requerirán una serie de restricciones a fin de evitar algunos resultados indeseables.

Definición 3.5.16 **[García y Simari, 2004]**

(Argumentos concordantes)

Sea $\mathcal{P} = (\Pi, \Delta)$ un delp y $\langle T_1, h_1\rangle$ y $\langle T_2, h_2\rangle$ estructuras de argumento obtenidas a partir de $\mathcal{P}$. Se dirá que $\langle T_1, h_1\rangle$ y $\langle T_2, h_2\rangle$ son argumentos *concordantes* si y sólo si el conjunto

$\Pi \cup T_1 \cup T_2$ no es contradictorio. Más generalmente un conjunto de estructuras de argumentos $\{\langle T_i,h_i\rangle\}_{i=1}^{n}$ es concordante si y sólo si $\Pi \cup \bigcup_{i=1}^{n} T_i$ no es contradictorio.

Definición 3.5.17 **[García y Simari, 2004]**

(Línea de argumentación aceptable)

Sea $\Lambda=[\langle T_0, h_0\rangle,\ldots, \langle T_i, h_i\rangle,\ldots, \langle T_n, h_n\rangle]$ una línea de argumentación, se dirá que Λ es *aceptable* si y sólo si:

i. Λ es una secuencia finita.

ii. El conjunto Λ_S, de argumentos de soporte, es concordante y el conjunto Λ_I, de argumentos de interferencia, es concordante.

iii. Ningún argumento $\langle T_k, h_k\rangle$ en Λ es un subargumento de un argumento $\langle T_i, h_i\rangle$ aparecido previamente en Λ ($i < k$).

iv. Para todo *i,* tal que el argumento $\langle T_i, h_i\rangle$ es un derrotador por bloqueo para $\langle T_{i-1},h_{i-1}\rangle$, si $\langle T_{i+1},h_{i+1}\rangle$ existe, entonces $\langle T_{i+1},h_{i+1}\rangle$ es un derrotador propio para $\langle T_i, h_i\rangle$.

□

Una rasgo interesante en la definición consiste en que cambiar alguna de las propiedades puede modificar el comportamiento en el formalismo.

En *DeLP* un literal *h* estará justificado si existe una estructura de argumento $\langle T, h\rangle$ no derrotada. Con vistas a establecer si $\langle T, h\rangle$ es no derrotado, el conjunto de derrotadores para tal argumento debe ser considerado. Dado que para cualquier estructura de argumento que derrote a $\langle T, h\rangle$ puede a su vez ser derrotado y así con otros, más de una línea de argumentación puede aparecer.

El conjunto de las líneas de argumentación para un argumento cualquiera podrá tener la estructura de árbol. En la definición 3.5.18 se expresa formalmente la idea empleando la noción de '*árbol dialéctico'.*

Definición 3.5.18 **[García y Simari, 2004]**

(árbol dialéctico)

Sea $\mathcal{P}$ = (Π, Δ) un *delp* y $\langle T_0, h_0 \rangle$ una estructura de argumento obtenida a partir de $\mathcal{P}$. Un *árbol dialéctico* para $\langle T_0, h_0 \rangle$, notado $\mathcal{T}_{\langle T_0, h_0 \rangle}$, es definido como sigue:

i. La raíz del árbol es etiquetado con $\langle T_0, h_0 \rangle$.

ii. Sea N un nodo no raíz del árbol etiquetado $\langle T_n, h_n \rangle$, y Λ= **[**$\langle T_0, h_0 \rangle$,$\langle T_1, h_1 \rangle$,$\langle T_2, h_2 \rangle$,..., $\langle T_n, h_n \rangle$**]** la secuencia de etiqueta de los caminos de la raíz a N.

Sean $\langle S_1, j_1 \rangle$,$\langle S_2, j_2 \rangle$,..., $\langle S_k, j_k \rangle$ todos los derrotadores para $\langle T_n, h_n \rangle$.

Para cada derrotador $\langle S_i, j_i \rangle$ ($1 \leq i \leq k$), tal que, la línea de argumentación Λ'=**[**$\langle T_0, h_0 \rangle$,$\langle T_1, h_1 \rangle$,$\langle T_2, h_2 \rangle$,...,$\langle T_n, h_n \rangle$,$\langle S_i, j_i \rangle$**]** sea aceptable, existe un nodo hijo N_i de N etiquetado con $\langle S_i, j_i \rangle$.

Si no existe ningún derrotador para $\langle T_n, h_n \rangle$ o no existe un $\langle S_i, j_i \rangle$ tal que Λ' es aceptable entonces N es una hoja.

□

En un árbol dialéctico, cualquier nodo, excepto el nodo raíz, representa un nodo derrotador de su padre y las hojas corresponden a argumentos no derrotados. Cada camino desde la raíz a una hoja corresponde a una línea de argumentación diferente.

En orden a decidir si la raíz de un árbol dialéctico es derrotada, un proceso de marcado es definido en *DeLP*. Los nodos serán marcados recursivamente como "*D*" (derrotados) o "*U*" (no derrotados) como sigue:

Definición 3.5.19 **[García y Simari, 2004]**

(marcado de un árbol dialéctico)

Sea $\mathcal{T}_{\langle T,\ h\rangle}$ un árbol dialéctico para $\langle T,\ h\rangle$, el correspondiente *marcado del árbol dialéctico*, notado como $\mathcal{T}^*_{\langle T,\ h\rangle}$, será obtenido marcando cada nodo en $\mathcal{T}_{\langle T,\ h\rangle}$ de la siguiente manera:

i. Todas las hojas en $\mathcal{T}_{\langle T, h\rangle}$ serán marcadas como "U" en $\mathcal{T}^*_{\langle T, h\rangle}$.

ii. Sea $\langle S, j\rangle$ un nodo interno de $\mathcal{T}_{\langle T, h\rangle}$ entonces $\langle S, j\rangle$ será marcado como "*U*" en $\mathcal{T}^*_{\langle T, h\rangle}$ si y sólo si cualquier hijo de $\langle S, j\rangle$ es marcado como "*D*" en $\mathcal{T}^*_{\langle T,\ h\rangle}$. El nodo $\langle S, j\rangle$ será marcado como "*D*" en $\mathcal{T}^*_{\langle T,\ h\rangle}$ si y sólo si existe al menos un hijo marcado como "*U*" en $\mathcal{T}^*_{\langle T, h\rangle}$.

□

La noción de literal justificado será definida en base a la definición 3.5.19.

Definición 3.5.20 **[García y Simari, 2004]**

(literal justificado)

Sea $\langle T,\ h\rangle$ una estructura de argumento y $\mathcal{T}^*_{\langle T, h\rangle}$ su árbol dialéctico marcado asociado. El literal *h* es *justificado* si y sólo si la raíz de $\mathcal{T}^*_{\langle T, h\rangle}$ es marcada como "*U*". Se dirá que *T* es una justificación para h.

□

Tal como lo destaca García y Simari (2004) es interesante notar que las nociones de línea de argumentación aceptable y árbol dialéctico proveen una estructura flexible para definir diferentes protocolos de argumentación cuando se consideran diferentes estrategias para aceptar derrotadores durante un proceso argumentativo. Este aspecto constituye una ventaja sobre otros formalismos donde el cambio del protocolo significa un cambio de sistema.

En *DeLP* es posible consultar sobre el estado de justificación de un literal *h*. Dependiendo de si este está justificado o no, cuatro tipos de respuesta pueden darse. La respuesta a las preguntas se basa en un operador modal de creencia '*B*' que señalará *Bh* (cree en *h*) cuando *h* este justificado, ¬*Bh* (no cree en *h)* si *h* no está justificado.

Definición 3.5.21 **[García y Simari, 2004]**

(respuesta a preguntas)

Las respuestas de un intérprete *DeLP* pueden ser definidas en base a un operador modal *B.* En términos de *B*, existen cuatro respuestas a las preguntas sobre *h:*

- SÍ, si *Bh* (*h* está justificado)
- NO, si $B\bar{h}$(el complemento de *h* está justificado)
- NO DECIDIDO, si ¬*Bh* y ¬*B*~*h* (ni *h* ni ~*h* están justificados)
- DESCONOCIDO, si *h* no está en el lenguaje del programa.

□

Además de los componentes enunciados, *DeLP* cuenta con una *poda* para el proceso de justificación. También cuenta con extensiones que permiten incorporar negación por falla y presunciones. Las mismas pueden ser consultadas en (García y Simari, 2004) y en (Martínez et al., 2012).

3.6 El Proyecto *Oscar*

John Pollock propuso un marco para la argumentación rebatible en su proyecto *Oscar*. Aunque el proyecto es más amplio, se presentará la estructura del razonamiento rebatible expuesta en (Pollock, 1995).

Según Pollock el razonamiento procede mediante la construcción de argumentos donde las razones proporcionan links atómicos para esos argumentos. Algunas razones son conclusivas, i.e. implican lógicamente su conclusión, otras no conclusivas, denominadas razones *prima facie.* Las razones *prima facie* crean una presunción a favor de la conclusión que sustentan pero cabe la posibilidad de su retractación con la incorporación de nueva información en el sistema de razonamiento.

Una razón puede ser codificada como un par ordenado $\langle \Gamma, p \rangle$, donde Γ es un conjunto de premisas y p una conclusión. Cuando para una razón *prima facie* existe otra razón que pone en entredicho lo sustentado por aquella se dice que tal razón es un derrotador para la primera. Las razones *prima facie* pueden ser derrotadas de dos maneras, o bien por rebatimiento o bien por socavamiento. Los derrotadores por rebatimiento contradicen la conclusión de la razón *prima facie* mientras que la derrota por socavamiento ataca la conexión entre las premisas y la conclusión.

Definición 3.6.1 **[Pollock, 1995]**
(rebatimiento)

Sea $\langle \Gamma, p \rangle$ una razón prima facie. Una razón $\langle \Lambda, q \rangle$ es un *derrotador por rebatimiento (rebutting defeater)* para $\langle \Gamma, p \rangle$ si y sólo si $q = \ulcorner \neg p \urcorner$.

□

Definición 3.6.2 **[Pollock, 1995]**
(socavamiento)

Sea $\langle \Gamma, p \rangle$ una razón prima facie. Una razón $\langle \Lambda, q \rangle$ es un *derrotador por socavamiento* (*undercutting defeater*) de $\langle \Gamma, p \rangle$ si y sólo si $q = \ulcorner \sim (\Pi\Gamma \gg p) \urcorner$.

Donde ΠΓ es la conjunción de las premisas en Γ y ⌈P≫Q⌉ se interpreta como P no sería verdadero a menos que Q lo fuera. ⌈~ (P≫Q)⌉ es abreviado como ⌈ (P ⊗Q)⌉.

□

La derrota por socavamiento propuesta por Pollock es tal vez una de las propuestas más innovadoras en el conjunto de los sistemas argumentativos. La derrota por socavamiento se caracteriza principalmente por el hecho de brindar razones que ponen en duda o cuestionan la capacidad sustentadora de las razones con respecto a la conclusión. Como nota, es importante destacar que aunque en otros sistemas, como por ejemplo el de Prakken y Sartor (1996[b]) se emplea la expresión '*socavamiento'* esté no comparte el mismo significado.

Los razonamientos se retrotraen a un conjunto *inicial* de premisas (denominado "inicial") asumido como previamente especificado. A partir de *inicial*, el razonador obtiene inferencias (algunas conclusivas, otras rebatibles) mediante la aplicación de diversos patrones de inferencia. Si bien, la mayoría de los sistemas argumentativos se focalizan en argumentos elaborados a partir de una base de conocimiento preestablecida, la propuesta de Pollock es innovadora en el sentido de que prevé la obtención de dos tipos de razonamientos, lineales y suposicionales.

Los argumentos lineales se construyen mediante la derivación continuada de conclusiones a partir de creencias previas que constituyen las razones para las conclusiones. Dicho de otro modo, estos argumentos constituyen secuencias finitas de proposiciones cada una de las cuales es o bien un miembro de las premisas de entrada o inferible a partir de miembros previos de la secuencia de acuerdo con algún esquema de razonamiento. Por ejemplo, considérese los siguientes dos argumentos para P&Q a partir de la premisa Q&P:

i.	(P&Q)	*i.*	(P&Q)
ii.	Q de i.	*ii.*	P de i.
iii.	P de i.	*iii.*	Q de i.
iv.	(Q&P) de ii y iii.	*iv.*	(Q&P) de ii y iii.

Ambos ejemplos representan diversos órdenes en que la inferencia puede ocurrir pero las relaciones de dependencia son las mismas y pueden ser representadas más perspicuamente como un grafo de inferencia:

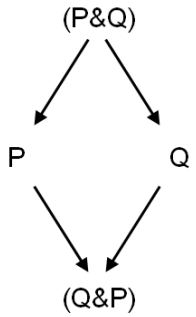


Fig. 3.6. 1: Grafo de inferencia

La primera manera de expresar un razonamiento, se denomina secuencial, en tanto que codifica las secuencias de inferencia. La representada en la figura 3.6.1 es una representación gráfica que codifica las relaciones de dependencia. Ambas estructuras son útiles para diferentes propósitos. A las estructuras secuenciales, Pollock las llama *argumentos* y a las estructuras gráficas, *grafos de inferencia*. Los argumentos y los grafos de inferencia son intercambiables. Pollock sostiene que el razonamiento puede entenderse como la producción de distintos argumentos que sustentan un conjunto de conclusiones y que todo este procedimiento puede representarse en un único grafo de inferencia. Tal grafo permite mostrar que inferencias se han realizado y como están sustentadas unas en otras y en consecuencia evaluar las creencias del razonador.

Los razonamientos suposicionales son un tipo de razonamientos caracterizado por suponer alguna proposición que no se puede inferir de las premisas, obtener conclusiones a partir de tal suposición y descargar, finalmente, tal proposición para obtener la conclusión deseada sin que dependa de la suposición inicial. La condicionalización, la reducción al absurdo y el razonamiento por casos son ejemplos de razonamiento suposicional.

El razonamiento suposicional puede ser codificado en un grafo de inferencia. Para esto cada nodo en el grafo se representa como un par ordenado $\langle X,p\rangle$ que consiste de una suposición X (*formada por un conjunto de proposiciones*) y una conclusión p (*una proposición*). En el razonamiento suposicional se avanza desde las proposiciones iniciales o suposiciones mediante reglas de inferencia. Estas pueden verse como reglas para agregar nodos a un grafo de inferencia. A continuación se presentan algunas de estas reglas.

Premisa: si p es una premisa y G un grafo de inferencia, entonces para cualquier suposición X es posible construir un nuevo grafo de inferencia agregando a G un nodo formado por $\langle X, p\rangle$ con ningún ancestro inmediato.

Suposición: si G es un grafo de inferencia y X es un conjunto finito de proposiciones tal que $p \in X$ entonces es posible construir un nuevo grafo de inferencia agregando a G un nodo formado por $\langle X, p\rangle$ con ningún ancestro inmediato.

Razón: si G es un grafo de inferencia formado por los nodos $\langle X, p_1\rangle$,..., $\langle X, p_n\rangle$ y $\langle [p_1,...,p_n], q\rangle$ es una razón (*conclusiva* o *prima facie*) entonces es posible construir un nuevo grafo de inferencia agregando a G un nodo formado por $\langle X, q\rangle$ con ancestros inmediatos $\langle X, p_1\rangle$,..., $\langle X, p_n\rangle$.

Condicionalización: si G es un grafo de inferencia que contiene al nodo ⟨X ∪{p}, q⟩ entonces es posible construir un nuevo grafo de inferencia agregando a G un nodo formado por ⟨X (p ⊃ q⟩ con ⟨X ∪{p}, q⟩ como ancestro inmediato.

Razonamiento por Casos: si G es un grafo de inferencia que contiene a los nodos ⟨X (p ∨ q⟩, ⟨X ∪{p}, r⟩ y ⟨X ∪{q}, r⟩ entonces es posible construir un nuevo grafo de inferencia agregando a G un nodo ⟨X, r⟩ con ⟨X (p ∨ q⟩, ⟨X ∪{p}, r⟩ y ⟨X ∪{q}, r⟩ como ancestros inmediatos.

Las relaciones de derrota anteriormente expuestas (definición 3.6.1 y 3.6.2) pueden ser definidas como una relación entre pares de nodos de un grafo de inferencia. Un nodo de un grafo de inferencia derrota a otro por tratarse de un derrotador para él. No obstante, frente a los casos de razonamiento suposicional no basta con esta idea. Si un nodo η se obtiene aplicando la razón *prima facie* ⟨ Γ, p⟩ y un nodo σ contiene un derrotador para ⟨ Γ, p⟩ esto no garantiza que σ derrota a η, pues η y σ pueden basar su conclusión en suposiciones distintas. De modo que un derrotador basado en una suposición *X* sólo puede derrotar a un nodo basado en una suposición *Y* tal que *X*⊆*Y*. Formalmente, las relaciones de derrota son definidas de la siguiente manera.

Definición 3.6.3 **[Pollock, 1995]**

(rebatimiento revisado)

Sean η y σ dos nodos en un grafo de inferencia. η *rebate* a σ si y sólo si se verifican las siguientes condiciones:

i. σ es un nodo que surge de una razón prima facie y sustenta una proposición *q* con fuerza ξ basado en una suposición *Y*;

ii. η sustenta ¬*q* con fuerza χ basado en una suposición *X*, donde *X*⊆*Y* ;

iii. χ ≥ ξ.

Definición 3.6.4 **[Pollock, 1995]**

(socavamiento revisado)

Sean η y σ dos nodos en un grafo de inferencia. Se dice que η *socava* a σ si y sólo si se verifican las siguientes condiciones:

i. σ es un nodo que surge de una razón prima facie y sustenta una proposición *q* con fuerza ξ basado en una suposición *Y* y $p_1,\ldots ,p_n$ son las proposiciones sustentadas por sus ancestros inmediatos;

ii. η sustenta a $(p_1 \,\&\, \ldots \,\&p_n)\otimes$q con fuerza χ basado en una suposición *X*, donde X ⊆ Y;

iii. χ ≥ ξ.

□

Una vez definida las relaciones que pueden darse entre los grafos de inferencia y establecido cómo obtener conclusiones a partir de los razonamientos resta determinar cómo evaluar los razonamientos que el sistema realiza. Las siguientes definiciones contribuirán a señalar las condiciones que permiten identificar cuándo una proposición está justificada. Formalmente:

Definición 3.6.5 **[Pollock, 1995]**

(justificación)

Una creencia está *justificada* en grado δ si y sólo si está basada en un nodo del grafo de inferencia que no está derrotado y posee una fuerza mayor o igual que δ.

□

Para aplicar la noción de justificación es necesario definir como se asigna el estado a los nodos de un grafo de inferencia de acuerdo a la relación de derrota dada entre ellos. Pollock señala dos situaciones en las que un nodo debe estar derrotado, cuando

σ es derrotado por un nodo sin derrotar o cuando σ se infiere a partir de un nodo derrotado.

La primera propuesta de Pollock para computar el estado de las derrotas fue la siguiente:

Definición 3.6.6 **[Pollock, 1995]**

(asignación de estados)

1. Un nodo η tal que ni η ni ninguno de sus ancestros está derrotado por algún nodo en el grafo de inferencia no está derrotado. Estos nodos se denominan *D-iniciales.*
2. Si los ancestros inmediatos de un nodo η no están derrotados y todos los nodos que derrotan a η están derrotados entonces η no está derrotado.
3. Si η posee un ancestro inmediato derrotado o existe un nodo sin derrotar que lo derrota entonces η es un nodo derrotado.

□

Esta definición funciona bien para casos sencillos pero es incapaz de resolver escenarios complejos como la derrota colectiva y la de argumentos autoderrotados. Para resolver tales inconvenientes Pollock propuso la siguiente definición:

Definición 3.6.7 **[Pollock, 1995]**

(estado de asignación revisado)

1. Los nodos D-iniciales *no están derrotados*.
2. Los nodos autoderrotados están *totalmente derrotados*.

3. Si η no es un nodo autoderrotado, sus ancestros inmediatos no están derrotados y todos los nodos que derrotan a η están totalmente derrotados entonces η *no está derrotado*.
4. Si η posee un ancestro inmediato totalmente derrotado o existe un nodo sin derrotar que lo derrota entonces η es un nodo *totalmente derrotado*.
5. En cualquier otro caso η está *provisionalmente derrotado*.

□

Es importante notar que los nodos provisionalmente derrotados conservan la capacidad de derrotar a sus adversarios, i.e. son capaces de propagar la derrota. Los argumentos autoderrotados son excluidos explícitamente mediante la cláusula 2 de la definición 3.6.7. Sin embargo, tal noción también presenta algunos problemas por lo que fue necesario proponer la siguiente noción de estados de asignación parcial.

Definición 3.6.8 **[Pollock, 1995]**

(estado de asignación parcial)

Una asignación σ de los estados derrotado y no derrotado a un subconjunto de nodos de un grafo de inferencia G es un *estado de asignación parcial* si y sólo si se verifican las siguientes condiciones:

1. σ asigna "no derrotado" a todos los nodos D-iniciales.
2. σ asigna "no derrotado" a un nodo α si y sólo si asigna "no derrotado" a todos sus ancestros inmediatos y "derrotado" a todos los nodos que derrotan a α.
3. σ asigna "derrotado" a un nodo α si y sólo si o α posee un ancestro inmediato al que se le asigna "derrotado" o existe un nodo β tal que β derrota a α y se le asigna el estado "no derrotado".

□

La noción de estado de asignación parcial soluciona los problemas detectados por Pollock en las otras definiciones (argumentos autoderrotados, derrota colectiva, argumentos flotantes). La siguiente definición complementa la idea:

Definición 3.6.9 **[Pollock, 1995]**

(estado de asignación)

Sea σ un estado de asignación parcial de un grafo G; σ es un *estado de asignación* si y sólo si σ no está propiamente contenido en ningún otro estado de asignación parcial de G.

□

Una conclusión puede encontrarse epistémicamente justificada o idealmente garantizada. Ambas nociones caracterizan una semántica práctica y una semántica idealizada. La primera computable en un tiempo finito, la otra se elabora en base a un conjunto de argumentos potencialmente infinito. La semántica idealizada permite establecer que una conclusión está idealmente garantizada en base al conjunto G, compuesto por todos los nodos que el razonador puede construir a partir de las premisas.

Definición 3.6.10 **[Pollock, 1995]**

(idealmente garantizado)

Un par $S = \langle X, \sigma \rangle$ formado por una razón y una conclusión está *idealmente garantizado* en grado δ relativo a un conjunto de premisas P si y sólo si existe un nodo α con fuerza conclusiva mayor o igual a δ tal que:

1. $\alpha \in G$,
2. α no está derrotado con respecto a G y
3. α sustenta a S.

La definición anterior permite identificar las creencias que poseería un razonador sin restricciones de recursos i.e. capaz de producir y analizar todos los argumentos relevantes para las creencias dadas. Ahora bien, una caracterización computacional requerirá poder seleccionar ciertas creencias en un momento determinado. Mientras que G es un conjunto compuesto por todos los nodos que el razonador *puede* construir a partir de las premisas, G_i es el conjunto de los primeros i-nodos. De manera que una creencia podrá estar justificada en base al *i-ésimo* paso como sigue:

Definición 3.6.11 **[Pollock, 1995]**
(epistémicamente justificado)

Un par $S = \langle X, \sigma \rangle$ formado por una razón y una conclusión está *justificado* en grado δ en el estadio *i* si y sólo si existe un nodo α con fuerza conclusiva mayor o igual a δ tal que:

1. $\alpha \in G_i$,
2. α no está derrotado con respecto a G_i y
3. α sustenta a S.

□

La justificación epistémica permite conocer el estado de una proposición en un momento determinado. Conforme el proceso continúe, el estado de las creencias podrán fluctuar entre justificado y no justificado. La siguiente definición modela el comportamiento real del sistema, mientras que la noción de idealmente garantizado representa el comportamiento esperado.

Definición 3.6.12 **[Pollock, 1995]**
(garantización)

Un par S = ⟨X, σ⟩ formado por una razón y una conclusión está *garantizado* en grado δ si y sólo si existe *i* tal que para todo *j* ≥ *i* se cumple que S está justificado en grado δ en el estadío *j*.

□

La definición anterior, en conjunto con las de *justificación* y *estados de asignación parcial,* permiten determinar si una razón y su conclusión pueden contar como creencias de un razonador. Obviamente, el sistema Oscar es más amplio, para una consideración completa del mismo, puede ser consultado en (Pollock, 1995).

3.7 Conclusión

El sistema propuesto por Prakken y Sartor en (1996[b]) modela información tentativa y potencialmente contradictoria mediante la introducción de una negación débil que permite caracterizar las reglas rebatibles. La noción de argumento es definida en base a un encadenamiento de reglas. Tales argumentos pueden estar en conflicto entre ellos. Los argumentos en conflictos pueden ser derrotados o bien por rebatimiento, cuando los argumentos poseen conclusiones complementarias pero uno no es preferido a otro o bien por socavamiento, i.e. cuando la conclusión del derrotador niega una suposición en el derrotado. Nótese que la definición de socavamiento propuesta en (Prakken & Sartor, 1996b) es diferente a la señalada por Pollock (1995). Por ello, en el capítulo anterior se definió el socavamiento de Prakken y Sartor como *ataque a hipótesis.*

Una vez que todas las interacciones posibles entre los argumentos son establecidas, se decide si un argumento puede considerarse justificado, defendible o denegado mediante una función de punto fijo. El conjunto de las consecuencias de tales argumentos constituirán conclusiones justificadas, defendibles o denegadas

dependiendo de si los argumentos que las sustentan son justificados, defendibles o denegados.

Las versiones básica y extendida del sistema se diferencian por la manera en cómo las preferencias son obtenidas. La versión básica expresa un sistema donde las prioridades entre las reglas son previamente establecidas, mientras que la versión extendida permite la obtención de argumentos que sancionan prioridades.

Bondarenko, Dung, Kowalski y Toni (1997) proponen un enfoque abstracto para modelar razonamiento default basado en argumentación. Tal enfoque incluye varios formalismos, como la lógica default, programación lógica extendida, lógica modal no monotónica, entre otros, como casos especiales del sistema.

En el sistema *BDKT* los argumentos son vistos como conjuntos de suposiciones. Las suposiciones son conjuntos de reglas defaults. Estas pueden ser agregadas a la teoría permitiendo derivar conclusiones que no pueden obtenerse considerando únicamente a la teoría. Los argumentos i.e. el conjunto de suposiciones, son entidades que sustentan extensiones de la teoría, y defiende o derrotan a alguna suposición o conjunto de suposiciones. A partir de las relaciones de derrota que entre ellas se dan es posible determinar si un conjunto de suposiciones constituye o no una extensión de la teoría.

La semántica más simple requiere que la extensión sea consistente y maximal. Por su parte una semántica no tan liberal denominada '*estable',* permite definir una extensión caracterizada porque derrota a toda aquella suposición o conjuntos de suposiciones que no pertenecen a ella. Un conjunto denominado '*extensión completa'* está caracterizado por que es admisible y contiene a todas las suposiciones defendidas por

ese conjunto. La versión escéptica sanciona como creencias del sistema al conjunto de suposiciones que puede ser derivado de todas las extensiones aceptables. Con tales semánticas, *BDKT* es capaz de capturar a otros sistemas como instancias del sistema.

Kowalski y Toni (1996) extienden la propuesta de *BDKT* con vista a modelar en un sistema de programación ordinario un sistema de programación rebatible mediante la traducción de las reglas rebatibles en reglas estrictas. Adicionalmente, en (Kowalski& Toni, 1996) se ordenan las reglas mediante un orden de prioridad establecido en el sistema. Las reglas permiten generar conjuntos de suposiciones que pueden estar en conflicto con otras y mediante las preferencias se define qué conjunto es mejor, i.e. qué conjunto derrota a cual. Una vez definida las relaciones de derrota es posible determinar el conjunto de suposiciones que extienden la teoría mediante una semántica basada en *BDKT*.

Simari y Loui en (1992) definen un sistema argumentativo basado en la preferencia por especificidad propuesta en (Poole, 1985) y la justificación por niveles debida a Pollock (1987). Los argumentos son entendidos como conjuntos de reglas rebatibles que permiten sancionar conclusiones rebatibles en unión con un conjunto de reglas estrictas y de hechos. Se exige de los argumentos que sean consistentes y minimales. Una vez construidos los argumentos estos pueden compararse mediante especificidad. En caso de que dos argumentos estén en conflicto es posible que uno sea derrotado por otro, apelando a la especificidad. El conjunto de argumentos justificados es definido en base a una justificación por niveles al estilo de (Pollock, 1987).

Posteriormente, en (Simari et al., 1994) se realizan diversas modificaciones con vistas a evitar una serie de inconvenientes que *MTDR* padecía. El cambio más importante

consiste en la modificación del mecanismo de justificación. El nuevo mecanismo se encuentra basado en principios dialécticos mediante la construcción de un árbol donde argumentos a favor y en contra de un argumento determinado se van sucediendo, conformando líneas argumentativas.

García y Simari (2004) proponen un sistema de programación lógica rebatible (*DeLP*) basado en *MTDR*. *DeLP* mantiene los componentes básicos propuestos en *MTDR* con los mejoramientos respectivos presentados en (Simari et al., 1994). Adicionalmente, *DeLP* cuenta con una versión alternativa de comparación de argumentos basada en reglas, tal vez al estilo de sistema básico de (Prakken & Sartor, 1996[b]). A su vez, el sistema ha sido enriquecido con vistas a modelar negación por falla y presunciones (Martínez et al., 2012).

Pollock en (1995) define una teoría del razonamiento rebatible partiendo de la distinción entre *razones conclusivas* y *prima facie*. Las razones *prima facie* constituyen buenas razones para sustentar su conclusión pero pueden verse derrotadas por razones mejores. Dos tipos de derrotas pueden darse entre los argumentos: derrota por rebatimiento y socavamiento.

Una vez construidos los argumentos y definida las derrotas es preciso determinar cuándo un argumento contará como justificado. En términos intuitivos, un argumento se encontrará justificado cuando ningún argumento finalmente lo derrote. Si bien la noción intuitiva es clara, la manera en cómo poder capturarla formalmente no fue tarea fácil. Pollock fue modificando a lo largo de su trabajo la noción de estados de asignación hasta proponer finalmente la de estado de asignación parcial. Con la noción de estado de asignación parcial, la definición de justificación y garantización, Pollock se vale para determinar cuándo un argumento contará como justificado.

A pesar de que los sistemas se caracterizan de manera particular es posible distinguir las nociones generales presentadas en el capítulo anterior: A partir de un conjunto de argumentos rebatibles es posible definir las relaciones de derrota entre ellos y, una vez establecidas todas las relaciones que pueden darse, se determina si un argumento está justificado. En tal idea subyace un principio fundamental en los sistemas argumentativos, el restablecimiento, en el capítulo siguiente se mostrarán algunos ejemplos que parecen ponerlo en duda.

Capítulo IV: Restablecimiento: algunos problemas

4.1 Introducción

Los sistemas argumentativos son formalismos que pretenden modelar razonamiento no monotónico mediante la construcción, comparación y evaluación de argumentos a favor o en contra de ciertas conclusiones. Los argumentos rebatibles pueden considerarse como entidades que sustentan conclusiones y que mantienen relaciones con otras entidades del mismo tipo. Diversas relaciones pueden darse entre ellos tales como conflicto, preferencia o derrota. Dependiendo de las relaciones de derrota y finalizado el proceso de comparación, un conjunto de argumentos conformará las *extensiones* del sistema. Diferentes criterios de selección para las extensiones han sido propuestos llamados *semánticas de argumentación.* Las conclusiones de aquellos argumentos que prevalecen frente a sus adversarios, bajo criterios determinados, serán el conjunto de las conclusiones aceptadas en el sistema.

Una noción involucrada en la obtención de las extensiones de un sistema argumentativo es la de *restablecimiento* y constituye el núcleo de la mayoría de los sistemas argumentativos, especialmente aquellos que pueden ser tratados como instancias de un marco argumentativo (Dung, 1995).

La intuición subyacente en el restablecimiento consiste en que un argumento prevalecerá frente a sus adversarios i.e. contará como parte de una extensión de un marco argumentativo, o tendrá una prueba en un juego argumentativo, cuando todos

sus posibles derrotadores estén a su vez derrotados. El siguiente ejemplo permite ilustrar la idea.

Ejemplo 4.1.1

Supóngase que el profesor X tiene razones para creer T: El profesor X ha visto a Tom Grabit robar un libro en la biblioteca por lo que puede concluir que Tom Grabit robó un libro de la biblioteca. Ahora supóngase que el profesor X cuenta con un derrotador D de T: la Señora Grabit dice que Tom está a miles de kilómetros de distancia y su hermano gemelo, que es cleptómano, estaba en la biblioteca el día en que el Profesor X supuestamente vio a Tom. Ahora bien, si luego el profesor X se entera, por su psiquiatra, que la Señora Grabit es una mentirosa compulsiva y desquiciada, y que el hermano gemelo de Tom es un invento de su mente, entonces ha adquirido un derrotador D para D.*

En el ejemplo 4.1.1 originalmente propuesto en (Lehrer y Paxson, 1969) es posible constatar que *D* hace que *T* sea considerado injustificado pero *D** restaura la justificación original de *T*. Aunque el ejemplo permite comprender la fuerza intuitiva de la noción del restablecimiento, otros ejemplos (debidos originalmente a Horty, 2001) parecen sugerir la idea de que el restablecimiento no debe considerarse como un principio general.

Ejemplo 4.1.2

A: *Dado que Al es un ave y teniendo en cuenta que las aves por lo general vuelan es posible concluir que vuela.*

B: *Sin embargo, Al es una gallina, y dado que las gallinas no vuelan se puede concluir que Al no vuela.*

C: *Ahora bien, dado que Al es una gallina salvaje, una clase particular de gallinas que tienen la habilidad de volar, se puede concluir que Al vuela.*

Atendiendo al ejemplo 4.1.2 y a cómo usualmente ha sido entendida la noción de derrota en sistemas argumentativos, el argumento *B* derrota al argumento *A* y *C* derrota al argumento *B,* de modo que *C* restablece al argumento *A*. Intuitivamente pareciera que las razones por las que alguien estaría dispuesto a aceptar que *Al vuela* no se deben al hecho de que *Al sea un ave* sino al hecho de que *Al es una gallina excepcional*, una gallina que vuela. El restablecimiento permite que un argumento basado en razones incorrectas (aunque sustente una conclusión correcta) cuente como un argumento justificado. Lo señalado aquí podrá ser confirmado luego, cuando este ejemplo sea implementado en los sistemas propuestos por Simari y Loui (1992) o por Prakken y Sartor (1996[b]) y se obtenga el mismo resultado.

El ejemplo 4.1.2 no parece demasiado problemático ya que en fin de cuentas el argumento sustenta una conclusión correcta. Sin embargo, el siguiente ejemplo permite señalar la posibilidad de que el restablecimiento tiene un impacto semántico más serio. El restablecimiento parece llevar a considerar como justificados a argumentos que sustentan conclusiones que son, simplemente, incorrectas.

Ejemplo 4.1.3

A: *Dado que Beth es una empleada de Microsoft y teniendo en cuenta que tales empleadas tienden a ser millonarios es posible concluir que Beth es millonaria.*

B: *Dado que Beth es una nueva empleada de Microsoft y teniendo en cuenta que tales empleadas por lo general poseen menos de medio millón es posible concluir que Beth posee menos de medio millón.*

C: *Dado que Beth es una nueva empleada de Microsoft en el departamento X y teniendo en cuenta que tales empleadas por lo general poseen al menos medio millón se puede concluir que Beth posee al menos medio millón.*

En este ejemplo, el argumento *B* da razones para rechazar al argumento *A,* por otro lado, la aceptación de *C* lleva al rechazo de *B,* de modo que, restablecimiento mediante, los argumentos *C* y *A* cuentan como argumentos justificados. A diferencia de lo que ocurre con el ejemplo 4.1.2, en este caso, el restablecimiento lleva a aceptar una conclusión incorrecta, *Beth es millonaria*. Atendiendo a la información total disponible, no hay razones para creer que Beth sea millonaria. La única razón que podría llevar a la aceptación de que es millonaria se debería a que es una empleada de Microsoft *estándar*, pero *no lo es.* Por otro lado y desde un punto de vista práctico no parece adecuado pensar que para defender el argumento *A* de *B* sea una buena idea emplear el argumento *C*. En realidad el argumento *C* parece sugerir un socavamiento a lo planteado por el argumento *A*.

Frente a tales ejemplos surgen varios interrogantes. El central puede expresarse con la siguiente pregunta: ¿Por qué, frente a ciertos casos, el restablecimiento lleva a la aceptación de argumentos incorrectos?

Una respuesta inmediata podría ser que tales resultados aparecen porque el restablecimiento no es un principio correcto y debería abandonarse. Esta idea ha sido sugerida en (Horty, 2001).

En contra de lo que sugiere Horty, Prakken (2002) señala que los contraejemplos no deben ser considerados como pruebas críticas sino más bien como generadores de investigaciones adicionales. Los ejemplos considerados entonces constituirán más bien casos que permiten poner en duda los mecanismos empleados con vistas a identificar, detectar o proponer estrategias de solución para minimizar el comportamiento anómalo.

Ahora bien, si se acepta *prima facie* que el restablecimiento es correcto, pueden surgir otras perspectivas al respecto en base a la pregunta enunciada. Unas de ellas pueden referirse a la validez general del restablecimiento, i.e. el restablecimiento es correcto pero existen condiciones que inhiben su aplicación. Otra que señale como responsables del comportamiento extraño del restablecimiento a las relaciones entre argumentos, como por ejemplo, la derrota, dado que podría pensarse que la relación de derrota tal y como es tratada en sistemas argumentativos es demasiado pobre como para capturar los casos problemáticos. Alternativamente podría pensarse que el lenguaje de representación empleado es inadecuado y esto hace que ciertos componentes implícitos, responsables del comportamiento extraño, no puedan ser considerados (Loui & Stiefvater, 1992; Prakken, 2002).

El objetivo del presente capítulo será abordar la problemática sugerida por los ejemplos propuestos 4.1.2 y 4.1.3 y discutir posibles maneras de interpretar los resultados. Para hacer frente a este objetivo, el capítulo se organiza como sigue: en la sección 4.2 se presentan los ejemplos problemáticos y su instanciación en algunos sistemas. En la sección 4.3 se reseñan las diversas respuestas que han sido dadas en la literatura y se ponderan las posibles maneras de interpretar los resultados. En la sección 4.4 se presenta un abordaje basado en la modificación del lenguaje de representación propuesto por Prakken (2002). Finalmente se concluye.

4.2 Restablecimiento contraintuitivo

El *restablecimiento* es un principio subyacente a los sistemas argumentativos y consiste en que un argumento podrá contar como aceptable a pesar de haber sido derrotado si todos sus posibles derrotadores son a su vez derrotados. Aunque este principio goza de un fuerte poder intuitivo, parecería que no es lo suficientemente general puesto que hay casos en los que determinados argumentos no deberían restablecerse, como en los ejemplos 4.1.2 y 4.1.3.

Con vistas a ilustrar el problema se instanciarán los ejemplos considerados en los sistemas *PS* (Prakken y Sartor, 1996[b]) y *KT* (Kowalski y Toni, 1996), como lo hizo Horty en (2001) al que se anexará un análisis en el sistema *MTDR* (Simari y Loui, 1992).

4.2.1 Restablecimiento contraintuitivo en *PS*

El sistema *PS* ha sido definido en la sección 3.2 del capítulo anterior. Tal como se ha señalado allí, la base de conocimiento o teoría T del sistema está constituida por un conjunto de reglas estrictas S y un conjunto de reglas rebatibles D. Como nota es menester recordar que para las reglas estrictas se empleará el conectivo '⇒' mientras que '→' será empleado para las rebatibles. T se complementa con una relación predefinida de preferencia entre reglas, representada a través de un orden estricto parcial. De modo que una *teoría ordenada* es una par (T, <) donde T = {S,D} y si $r<r'$, la regla r' es preferida a r.

Los argumentos son definidos como una secuencia finita $[r_n,..., r_m]$ de instancias básicas de reglas tal que para cada i, $n \leq i \leq m$, se verifica que para cada literal L en el antecedente de r_i, existe un $j<i$, tal que L es el consecuente de r_j, y ningún r_i tiene como

consecuente al consecuente de algún r_j ($j<i$). El conjunto de todos los argumentos en base a una teoría ordenada Γ será denotado como $Args_\Gamma$.

Puede suceder que en el conjunto de argumentos $Args_\Gamma$ existan argumentos que no puedan ser aceptados conjuntamente. Dados dos argumentos A_1 y A_2 en *PS* se dirá que A_1 *derrota* a A_2 cuando A_1 es vacío y A_2 es incoherente, o A_1 socava a A_2, o A_1 rebate a A_2 y A_2 no socava a A_1.

Un argumento A_1 *rebatirá* a otro A_2 cuando existe un par de reglas r_1y r_2 ($r_1 \in A_1$; $r_2 \in A_2$) tal que r_1y r_2 tienen consecuentes complementarios y r_2 no es preferido a r_1 notado como $r_1 \preccurlyeq r_2$. Por otro lado, un argumento A_1 *socavará* a A_2 cuando la conclusión de A_1 es el complemento de alguna suposición en A_2. En caso de que A_1 derrote a A_2 y A_2 no derrote a A_1, A_1 *estrictamente derrota a* A_2.

Comparar pares de argumentos no es suficiente para establecer qué argumento puede considerarse ganador en una disputa. El conjunto de argumentos ganadores o justificados (*JustArgs*) se construye paso a paso, colectando primero en un conjunto $JustArgs_1$ todos los argumentos no derrotados. Luego, se añaden todos los argumentos que resultan justificados indirectamente a partir de $JustArgs_1$ obteniendo un conjunto $JustArgs_2$. Este proceso se repite hasta obtener un punto fijo $JustArgs_n$. Tal punto fijo será el conjunto de todos los argumentos justificados. Si Γ es una teoría ordenada y $Args_\Gamma$ es el conjunto de todos los argumentos que pueden ser construidos en la teoría, el conjunto de argumentos justificados es $\bigcup_{i=0}^{\infty}(F_\Gamma^i)$ donde $F_\Gamma^0 = \varnothing$ y $F_\Gamma^{i+1} = \{A \in Args_\Gamma$: A es aceptable c.r. a $F_\Gamma^i\}$. Si A es un argumento tal que $A \in JustArgs_\Gamma$ entonces A está *justificado,* pero si A es atacado por $JustArgs_\Gamma$ A está denegado y si A no está justificado ni denegado, A es defendible.

El ejemplo 4.1.2 puede ser tratado en el sistema *PS* de la siguiente manera, téngase en cuenta que *WCa*, *Ca*, *Ba*, and *Fa* representan respectivamente las proposiciones *Al es una gallina salvaje*, *una gallina*, *un ave*, y *un volador*. De manera que el conjunto $Args_{\Gamma}$ estará constituido por:

Ejemplo 4.2.1 (Ejemplo 4.1.2 reconsiderado)

A: $\top \Rightarrow WCa$

B: $\top \Rightarrow WCa \Rightarrow Ca$

C: $\top \Rightarrow WCa \Rightarrow Ca \Rightarrow Ba$

D: $\top \Rightarrow WCa \Rightarrow Ca \Rightarrow Ba \rightarrow_{r1} Fa$

E: $\top \Rightarrow WCa \Rightarrow Ca \rightarrow_{r2} \neg Fa$

F: $\top \Rightarrow WCa \rightarrow_{r3} Fa$.

Dado que la regla r_2 es más específica que r_1 y r_3 es más específica que r_2, el orden de preferencia estará dado de la siguiente manera:

$r_2 > r_1$

$r_3 > r_2$

Atendiendo a las preferencias es posible determinar las relaciones de derrota entre los argumentos de la siguiente manera: *E* estrictamente derrota *D*, y *F* estrictamente derrota *E*. Dado que el único argumento que derrota a *D* es también derrotado, entonces, tal argumento es restablecido, y puede contarse como justificado de acuerdo con el sistema *PS*. El procedimiento iterativo confirma lo esperado: $F^1 = \{A, B, C, F\}$, que $F^2 = \{A, B, C, D, F\}$, y $F^3 = F^2$, así que este es el menor punto fijo del operador F_{Γ}.

La clasificación de *D* como justificado parece correcta, dado que la conclusión "*Al vuela*", se encuentra legítimamente soportado por *F*, pero *D* es problemático dado que sugiere la idea de que *Al vuela* porque *es un ave*. En el contexto, este argumento es

extraño, ya que en realidad la conclusión *Al vuela* pude afirmarse debido a que es Al una gallina salvaje y no por el hecho de ser un ave. Como se ve, el sistema *PS* permite obtener un resultado incorrecto frente al ejemplo 4.1.2.

Para representar el ejemplo 4.1.3 en *PS* considere la siguiente representación, donde *NMEXb*, *NMEb*, *MEb*, *1Mb*, *>1/2Mb*, *<1/2Mb* representarán respectivamente las siguientes proposiciones: *Beth es una nueva empleada de Microsoft en el departamento X, Beth es una nueva empleada de Microsoft, Beth es una empleada de Microsoft, Beth tiene un millón dólares, Beth tiene menos de medio millón de dólares y Beth tiene más de medio millón de dólares.* Los argumentos de $Args_{\Gamma}$ son (simplificando):

Ejemplo 4.2.2 (Ejemplo 4.1.3 reconsiderado)

A: $\top \Rightarrow NMEXb$,

B: $\top \Rightarrow NMEXb \Rightarrow NMEb \Rightarrow MEb \rightarrow_{r1} 1Mb$

C: $\top \Rightarrow NMEXb \Rightarrow NMEb \rightarrow_{r2} >1/2Mb$

D: $\top \Rightarrow NMEXb \rightarrow_{r3} <1/2Mb$

B es un argumento que sustenta la conclusión *Beth es millonaria* por el hecho de ser una empleada de Microsoft. *C* es un argumento que sustenta la conclusión *Beth posee menos de medio millón* por el hecho de ser una nueva empleada de Microsoft, mientras que *D* es un argumento que sustenta *Beth posee al menos medio millón* por ser una nueva empleada de Microsoft en el departamento X. El orden de preferencia está dado de la siguiente manera: $r_2 > r_1$, $r_3 > r_2$ ya que la regla r_2 es más específica que r_1 y r_3 es más específica que r_2.

C estrictamente derrota *B* y *D* derrota a *C*. Dado que *D* derrota a *C*, restablece a *B* de modo que B contará entre los argumentos justificados del sistema *PS*. En el procedimiento iterativo: F^1= {*A, D*}, y F^2= {*A, D, B*}, y $F^3 = F^2$.

La clasificación de *B* entre los argumentos justificados presenta una seria dificultad: la conclusión *1Mb* simplemente es un error. La única razón para creer que Beth tiene un millón de dólares se debe a que sea una empleada típica de Microsoft, pero esto es anulado por la consideración de que Beth es una nueva empleada de Microsoft. De modo que para el sistema *PS*; *Beth posee un millón de dólares*.

4.2.2 Restablecimiento contraintuitivo en *KT*

En la sección 3.3 del capítulo anterior se brindó una presentación general del sistema *BDKT* propuesto en (Bondarenko et al., 1997). Luego se señaló que Kowalski y Toni generalizaron tal enfoque mediante la transformación de cualquier programa lógico rebatible en un programa lógico ordinario.

En tal enfoque, cada regla rebatible de la forma *r*: $A \leftarrow B_1,\ldots,B_n$ es reemplazada por dos reglas estrictas dotadas de los predicados adicionales *holds y ~defeated* de la siguiente manera:

$$A \Leftarrow holds(r) \text{ y } holds(r) \Leftarrow B_1,\ldots,B_n,\ \sim defeated(r).$$

Intuitivamente, esto significa que el consecuente de la regla *r,* es decir *A,* puede ser establecido si la regla *r* "*holds*" y la regla *r* "*holds*" si los antecedentes de la regla *r* pueden ser establecidos y *r* no es derrotada (*~defeated(r)*).

Una regla *r* es derrotada si existe una regla *r'* de mayor prioridad en conflicto tal que *r'* "holds", formalmente:

$$defeated(r) \Leftarrow r<r', conflict(r,r'), holds(r').$$

Una regla está en conflicto con otra cuando la conclusión de ambas es complementaria, es decir,

$$conflict(r,r') \Leftarrow conclusion(r,A), conclusión(r', \neg A).$$

Obviamente, si *r* y *r'* están en conflicto

$$conflict(r,r') \Leftarrow conflict(r',r).$$

El conjunto de argumentos aceptables es definido en base a diferentes semánticas. Con vistas a presentar los ejemplos considerados en la sección anterior, baste con apelar solamente a la noción de conjunto preferido según la definición 3.3.6.

El sistema *KT* también exhibe un comportamiento similar pero no idéntico al sistema *PS* frente a los anteriores ejemplos. En particular porque diferencia de *PS*, en *KT* no se explicitan los argumentos construidos (las razones que justifican una conclusión), de modo que el ejemplo de Al no puede constituir un contraejemplo en este caso, dado que hay dos conjuntos de reglas para la misma conclusión, por ello se presentará sólo el ejemplo 4.1.3 que permitirá evidenciar el problema.

La representación de tal ejemplo incluye reglas estrictas que representan la siguiente información: *Beth es una nueva empleada de Microsoft en el departamento X*, *Beth es una nueva empleada de Microsoft dado que es una nueva empleada en el departamento X*, *Beth es una empleada de Microsoft dado que es una nueva empleada de Microsoft*, *Beth posee al menos medio millón dado que posee un millón*, *no es cierto que Beth posea menos de medio millón si posee un millón, etc*. Formalmente:

Ejemplo 4.2.3 (Ejemplo 4.1.3 reconsiderado)

$NMEXb \Leftarrow$

$NMEb \Leftarrow NMEXb$

$MEb \Leftarrow NMEb$

$<1/2Mb \Leftarrow 1Mb$

$\sim >1/2Mb \Leftarrow 1Mb$

$\sim >1/2Mb \Leftarrow <1/2Mb$

$\sim <1/2Mb \Leftarrow >1/2Mb$

$\sim 1Mb \Leftarrow >1/2Mb$

Las reglas rebatibles por su parte son:

$r_1 : 1Mb \leftarrow MEb$

$r_2 : >1/2Mb \leftarrow NMEb$

$r_3 : <1/2Mb \leftarrow NMEXb$

donde r_1 da razones para creer que *Beth tiene un millón de dólares si ella es una empleada de Microsoft*, r_2 expresa *Beth posee menos de medio millón si ella es una nueva empleada de Microsoft*, y r_3 representa la regla *Beth posee más de medio millón si ella es una empleada del departamento X*. La prioridad está dada de la siguiente manera:

$r_1 < r_2 \Leftarrow$

$r_2 < r_3 \Leftarrow$

ya que la regla r_2 es más específica que r_1 y r_3 es más específica que r_2.

Las reglas rebatibles serán reescritas en reglas estrictas de la siguiente forma:

$1Mb \Leftarrow holds(r_1)$

$holds(r_1) \Leftarrow MEb, \sim defeated(r_1)$

$>1/2Mb \Leftarrow holds(r_2)$

$holds(r_2) \Leftarrow NMEb, \sim defeated(r_2)$

$<1/2Mb \Leftarrow holds(r_3)$

$holds(r_3) \Leftarrow NMEXb, \sim defeated(r_3)$

El conflicto es expresado de la siguiente manera:

$conflict(r_1,r_2) \Leftarrow$

$conflict(r_2,r_3) \Leftarrow$

puesto que los consecuentes de la regla r_1 y la regla r_2 son contradictorios al igual que entre los consecuentes de r_2 y r_3.

Ahora bien, es fácil observar que el programa posee exactamente un conjunto de hipótesis preferido:

$\Delta_1 = \{\sim defeated(r_1), \sim defeated(r_3)\}$.

De modo que las *<1/2Mb* y *1Mb* son las conclusiones del sistema, i.e. obtiene la conclusión de que *Beth posee al menos medio millón* pero también, contrario a la intuición, que *Beth tiene un millón de dólares*.

4.2.3 Restablecimiento contraintuitivo en *MTDR*

Simari y Loui (1992) propusieron un sistema formal basado en argumentos para modelar información tentativa y potencialmente contradictoria mediante la vinculación de los abordajes realizados por Poole (1985) y Pollock (1987). La presentación general del mismo puede encontrarse en el capítulo precedente, en la sección 3.4.

Allí se dijo que *MTDR* es definido a partir de un lenguaje formal compuesto por un lenguaje de primer orden *L* dotado de una relación binaria metalingüística "$\succ\!\!-$". Los miembros de la relación "$\succ\!\!-$" son denominados reglas rebatibles y tienen la forma

$A_1,\ldots, A_n \succ\!\!- B$ ($n \geq 1$) donde $A_1,\ldots, A_n$ son literales de L. Los literales del antecedente de la regla se encuentran en conjunción mientras que el consecuente está constituido por un único literal. Mediante sustitución uniforme de cada variable de una regla rebatible se obtiene una *instancia.*

El conocimiento de un agente a es representado por un par (K, Δ), denominado estructura lógica rebatible. K representa el conocimiento no rebatible de a y $\mathrm{K} \subseteq Sent(L)$. $Sent(L)$ puede ser particionado en dos subconjuntos $Sent_N(L)$ y $Sent_C(L)$, correspondiendo a la información *necesaria* (constituido por sentencias de variables libres o implicaciones en L) y *contingente* (constituido por literales instanciados) respectivamente. K se encuentra formado por dos subconjuntos disyuntos $\mathrm{K_N}$ y $\mathrm{K_C}$. Δ, por su parte, representa la información tentativa y es un conjunto finito de reglas rebatibles.

Dado un miembro A de $Sent(L)$ y un conjunto $\Gamma = \{A_1, A_2,\ldots, A_n\}$, donde cada A_i es miembro de K o una instancia de Δ, A será llamada una *consecuencia rebatible* de Γ si y sólo si existe una secuencia $B_1,\ldots, B_m$ tal que $A = B_m$ y para cada i, o bien B_i es un axioma de L, o B_i es en Γ, o B_i es una consecuencia directa de los miembros precedentes de la secuencia usando modus ponens o instanciación de una sentencia cuantificada universalmente.

Los argumentos son estructuras constituidas por conjuntos minimales de reglas rebatibles instanciadas y consistentes que sustentan una conclusión. Estos son notados como $\langle T, h\rangle$ donde T es el conjunto de reglas rebatibles y h la conclusión de la estructura de argumento. Si $S \subseteq T$, entonces $\langle S, j\rangle$ es un subargumento de $\langle T, h\rangle$, y si $S \subset T$, entonces $\langle S, j\rangle$ es un subargumento propio de $\langle T, h\rangle$.

Dos o más argumentos pueden sustentar conclusiones contradictorias, de modo que no pueden ser conjuntamente aceptados. La definición general que permite capturar casos de argumentos mutuamente excluyentes es la de contraargumentación. Se dice que un argumento $\langle T_1, h_1 \rangle$ *contraargumenta* a $\langle T_2, h_2 \rangle$ si y sólo si existe un subargumento $\langle T, h \rangle$ de $\langle T_2, h_2 \rangle$ tal que $\langle T_1, h_1 \rangle$ y $\langle T, h \rangle$ están en desacuerdo y se dice que están en desacuerdo cuando las conclusiones h y h_1 en unión con K implican contradicción.

Con vistas a establecer si un contraargumento tiene éxito, en *MTDR* se define un comparador de argumentos denominado *especificidad* (que en desarrollos posteriores se ha visto modificado). Intuitivamente, un argumento $\langle T_1, h_1 \rangle$ es más específico que $\langle T_2, h_2 \rangle$ cuando todo aquello que activa $\langle T_1, h_1 \rangle$ activa $\langle T_2, h_2 \rangle$ pero algo que activa $\langle T_2, h_2 \rangle$ no activa $\langle T_1, h_1 \rangle$. En la definición 3.4.5 se la expresa formalmente.

Mediante las nociones de contraargumentación y especificidad se propone la relación de derrota. En *MTDR* se dice que una estructura de argumento $\langle T_1, h_1 \rangle$ *derrota* a otra $\langle T_2, h_2 \rangle$, notado como $\langle T_1, h_1 \rangle \gg_{\text{def}} \langle T_2, h_2 \rangle$, cuando existe un subargumento $\langle T, h \rangle$ de $\langle T_2, h_2 \rangle$ tal que $\langle T_1, h_1 \rangle$ se encuentra en desacuerdo con $\langle T, h \rangle$ y $\langle T_1, h_1 \rangle$ es estrictamente más específico que $\langle T, h \rangle$.

El proceso de justificación de argumentos originalmente propuesto en (Simari & Loui, 1992) se encuentra basado en la justificación por niveles de Pollock (1987). Posteriormente este se ha visto refinado (Simari et al., 1994; García & Simari, 2004). Aunque para la modelación de los ejemplos considerados basta tener en cuenta las definiciones 3.4.7 y 3.4.8.

A continuación se presenta la representación de los ejemplos 4.1.2 y 4.1.3 en *MTDR.* Los mismos resultados pueden ser obtenidas en los refinamientos del sistema realizados en (Simari et al., 1994) y en (García & Simari, 2004).

Ejemplo 4.2.4 (Ejemplo 4.1.2 reconsiderado)
WC(a), *C(a)*, *B(a)*, y *F(a)* representan respectivamente a las proposiciones: *Al es un pollo salvaje, un pollo, un ave, y un volador*. A partir de tal información es posible construir los siguientes argumentos que permiten modelar el ejemplo de Al.

$\langle A_1,h_1\rangle$ $\quad\quad$ $\langle\{B(a) \succ\!\!- F(a)\}, F(a)\rangle$

$\langle A_2,h_2\rangle$ $\quad\quad$ $\langle\{C(a) \succ\!\!- \neg F(a)\}, \neg F(a)\rangle$

$\langle A_3,h_3\rangle$ $\quad\quad$ $\langle\{WC(a) \succ\!\!- F(a)\}, F(a)\rangle$

Dadas las relaciones de especificidad, es posible señalar que $\langle A_2, h_2\rangle$ es estrictamente más específico que $\langle A_1, h_1\rangle$ y $\langle A_3, h_3\rangle$ lo es con respecto a $\langle A_2, h_2\rangle$. Teniendo en cuenta que $\langle A_2, h_2\rangle$ es un contraargumento más específico que $\langle A_1, h_1\rangle$, $\langle A_2, h_2\rangle$ derrota a $\langle A_1,h_1\rangle$. Por otro lado, $\langle A_3, h_3\rangle$ derrota a $\langle A_2, h_2\rangle$ puesto que $\langle A_3, h_3\rangle$ es un contraargumento más específico que $\langle A_2, h_2\rangle$. Atendiendo a las relaciones de derrota, y al proceso de justificación, es posible señalar que en *MTDR* $\langle A_1,h_1\rangle$ y $\langle A_3,h_3\rangle$ son argumentos justificados, obteniendo el resultado no intuitivo.

Ejemplo 4.2.5 (Ejemplo 4.1.3 reconsiderado)
El ejemplo 4.1.3 puede ser modelado en *MTDR* de la siguiente manera: Si *NMEX(b), NME(b), ME(b), 1M(b), >1/2M(b), <1/2M(b)* representa respectivamente las siguientes proposiciones: *Beth es una nueva empleada de Microsoft en el departamento X, Beth es una nueva empleada de Microsoft, Beth es una empleada de Microsoft, Beth tiene un millón dólares, Beth tiene menos de medio millón de dólares y Beth tiene más de medio millón de dólares,* es posible construir los siguientes argumentos:

$\langle A_1,h_1\rangle$ $\langle\{ME(b)) \rightarrowtail 1M(b)\}, 1M(b)\rangle$

$\langle A_2,h_2\rangle$ $\langle\{NME(b) \rightarrowtail >1/2M(b)\}, >1/2M(b)\rangle$

$\langle A_3,h_3\rangle$ $\langle\{NMEX(b) \rightarrowtail <1/2M(b)\}, <1/2M(b)\rangle$

Es fácil comprobar que en *MTDR* $\langle A_3, h_3\rangle$ es un derrotador de $\langle A_2, h_2\rangle$ y $\langle A_2, h_2\rangle$ es un derrotador de $\langle A_1, h_1\rangle$ de modo que, tanto $\langle A_3, h_3\rangle$ como $\langle A_1, h_1\rangle$, calificarán luego del procedimiento de justificación, como argumentos justificados obteniendo un resultado no intuitivo.

4.3 Cómo interpretar estos resultados

Luego de haber presentado los ejemplos en cada uno de los sistemas considerados y haber constatado que los resultados no intuitivos son obtenidos, cabe la pregunta: ¿Cómo han de interpretarse los resultados precedentes? La razón del interrogante se debe al hecho de que los ejemplos ponen en duda la capacidad de los sistemas argumentativos para modelarlos adecuadamente. El responsable, según Horty (2001), parece ser el restablecimiento. Su sugerencia es simple pero drástica: dado que el restablecimiento permite, no sólo la obtención de resultados problemáticos como en el ejemplo 4.1.2, sino además alcanzar resultados incorrectos como en el ejemplo 4.1.3, es necesario abandonarlo como un principio general. Esta idea no es del todo original. Una indicación similar fue empleada en redes de herencia donde se establece la necesidad de abandonar el restablecimiento por el impacto semántico inadecuado que este principio genera (Touretzky *et al*., 1991). Ahora bien ¿la sugerencia de Horty es adecuada?

Aunque Loui y Stiefvater abordan el problema antes del trabajo de Horty, es instructivo considerar lo dicho por ellos en relación a lo discutido en (Touretzky *et al*., 1991). Ellos señalan que la valoración de los sistemas no debe hacerse en función de

si el restablecimiento es o no inadecuado sino en la capacidad de representación correcta de la información. Sugieren que para impedir el restablecimiento automático, responsable de la obtención del resultado no deseado, simplemente se debe representar de otra manera la información. En particular, frente al ejemplo de Al, proponen que la regla *las gallinas salvajes vuelan*, debe representarse como *las gallinas salvajes vuelan pero no porque son aves*.

En la misma línea Prakken (2002) indica que el problema no es debido al restablecimiento en sí, sino a la incapacidad del lenguaje que no permite bloquear los defaults cuando esto es necesario. Por ello sostiene que en realidad, los ejemplos propuestos por Horty no alcanzan para invalidar el restablecimiento. Para ilustrar su idea en (Prakken, 2002) muestra cómo, mediante un lenguaje de representación con mayor poder expresivo (dotado de cláusulas de anormalidad), pueden obtenerse los resultados adecuados sin necesidad abandonar el restablecimiento. Adicionalmente señala que cuando se modela otro tipo de información que no sea estadística o default, el restablecimiento permite obtener resultados adecuados, de manera que el problema también se encuentra en estrecha relación al tipo de información involucrada.

En una comunicación personal, referida en (Prakken y Vreeswijk, 2002), Pollock señala que el ejemplo 4.1.2 sería representado por él de la siguiente forma:

Ejemplo 4.3.2 (Ejemplo 4.1.2 reconsiderado)

A: *Que Al sea un ave es una razón prima facie para creer que vuela.*

B: *Que Al sea una gallina es una razón prima facie para creer que no vuela.*

C: *Que Al sea una gallina salvaje es una razón prima facie para creer que vuela.*

D: *Que Al sea una gallina es un derrotador por socavamiento para A.*

E: *Que Al sea una gallina salvaje es un derrotador por socavamiento para B.*

Bajo la representación de Pollock, en realidad no hay ningún inconveniente con respecto al restablecimiento. Lo cual sugiere que el problema aparece por la manera en que se modela la información aunque también es instructivo considerar que la derrota por especificidad no es la empleada aquí. Como se observa, la propuesta de Pollock introduce una derrota por socavamiento entre *D* y *A* como entre *E* y *B*, diferente a lo que es usual en *PS*, o *MTDR* y *KT*.

Para Pollock no existen otros tipos de derrota que las de rebatimiento y socavamiento por él propuestas. La derrota por especificidad (en razonamiento default o en estadístico), de la que los ejemplos son un caso, no es más que un ejemplo de las diversas maneras en que la derrota por socavamiento es expresada y, cómo tal, debe apelarse al requerimiento de evidencia total (Pollock, 2001). En función de esto es importante destacar lo señalado por Modgil y Prakken (2013), el socavamiento puede expresar la derrota por especificidad pero para ello es requerido mayor poder expresivo en el lenguaje a fin de que permita bloquear ciertos pasos autorizados por una regla determinada.

Adicionalmente es instructivo considerar el cambio de opinión en Horty sobre el tema. En (Horty, 2012) señala: "…*me basé en situaciones como el ejemplo de Microsoft para argumentar en contra del restablecimiento. Ahora creo que el problema radica, no en el restablecimiento en sí, sino en cómo se formalizan estas situaciones y que, una vez que estén debidamente representados, el restablecimiento puede ser visto como inocuo*…".

Una vez consideradas las interpretaciones de los resultados es claro que hay unanimidad en que el problema no es debido al restablecimiento en sí, sino a la manera en cómo se representa la información. En particular, cuando tal información consiste en el empleo de reglas default o regularidades estadísticas. Ahora bien, lo que parece no ser unánime es en cómo realizar esa representación. Podría hacerse mediante cambios de las representaciones al estilo de Loui y Stiefvater (1992), o mediante cláusulas de anormalidad como lo sugiere Prakken (2002) o a través de la incorporación de un lenguaje lo suficientemente expresivo que sea capaz de expresar derrotas por socavamiento al estilo de Pollock. Sin embargo, es instructivo atender a la advertencia de Prakken (2002). La investigación sobre cómo representar debe estar acompañada con la búsqueda de principios generales que permitan la elección de una representación correcta puesto que se debe evitar *remendar* (o *juguetear* con) la representación frente a ejemplos concretos.

Ahora bien, al parecer, encontrar principios generales de representación es una tarea difícil. Tal vez, una alternativa podría consistir en la identificación de algún criterio formal que permita la obtención de los resultados correctos sin depender de la representación. Pero antes de considerar esta alternativa, que será ampliamente discutida en el capítulo siguiente, será importante considerar el enfoque representacional más en detalle. Para ello será presentado parte de la discusión que puede encontrarse en (Prakken, 2002). Abordar este enfoque será importante, al menos, por dos razones. La primera se debe a que permitirá comprender mejor el problema. La segunda consiste en que la versión representacional podrá ser empleada para valorar la perspectiva que se sugerirá en el capítulo siguiente. En particular, porque al proponerse como alternativa, al menos debe ser capaz de modelar todo lo que la propuesta representacional sea capaz.

4.4 Lenguaje formal y restablecimiento

Prakken (2002) aborda el problema comenzando con una distinción entre restablecimiento directo e indirecto. Para Prakken, Horty (2001) simplemente elabora una crítica a la versión directa del restablecimiento. El restablecimiento es *directo* cuando los argumentos se encuentran en conflicto en sus conclusiones como en el ejemplo 4.2.1. El restablecimiento indirecto en cambio, como en el ejemplo 4.1.1, se da cuando el argumento restablecedor ataca a las premisas del derrotador, o también, cuando un argumento ataca la relación de justificación entre premisa y conclusión.

Partiendo del supuesto de que sólo el restablecimiento directo es cuestionado, Prakken procede a proponer una estrategia de solución señalando primeramente que está de acuerdo con Horty en:

i. los argumento menos específico de los ejemplos 4.1.2 y 4.1.3 deberían ser bloqueados y que

ii. la representación del sistema propuesto por Prakken y Sartor (1996[b]) es incapaz de hacerlo.

Sin embargo, Prakken no está de acuerdo en que las observaciones de Horty impliquen la invalidez del restablecimiento. Según Prakken hay un enfoque alternativo que permite obtener los resultados adecuados frente a los ejemplos problemáticos, pero que valida el restablecimiento y además puede hacer las distinciones pertinentes entre los ejemplos que pasan desapercibidos en la representación usual.

El enfoque consiste en hacer el lenguaje más expresivo a fin de que éste permita bloquear ciertos defaults cuando sea necesario. Una forma de implementarlo puede consistir en el empleo de cláusulas de anormalidad. Por ejemplo, en implicación

preferencial con minimización de los predicados de anormalidad '*Ab*', todos los modelos preferidos en la siguiente teoría y el hecho de que Beth sea una nueva empleada de Microsoft en el departamento X satisfacen que *Beth tiene al menos medio millón*, pero no todos los modelos satisfacen que es millonaria.

r_1: $EM \wedge \neg Ab_1 \rightarrow 1M$

r_2: $NEM \rightarrow EM$

r_3: $NEM \wedge \neg Ab_2 \rightarrow <1/2M$

r_4: $NEM \rightarrow Ab_1$

r_5: $NEMX \rightarrow NEM$

r_6: $NEMX \wedge \neg Ab_3 \rightarrow >1/2M$

r_7: $NEMX \rightarrow Ab_2$

donde *NMEX*, *NME*, *ME*, *1M*, *>1/2M*, *<1/2M* representan respectivamente las siguientes predicados: *nuevo empleada de Microsoft en el departamento X, nuevo empleada de Microsoft, empleada de Microsoft, tener un millón dólares, tener menos de medio millón de dólares y tener más de medio millón de dólares.*

Bajo este enfoque podrían construirse argumentos e integrarlos a un sistema argumentativo dando el resultado esperado. El sistema podría tener un estilo de representación como el siguiente:

A: $\langle \{EM(Beth) \wedge \neg Ab_1 \rightarrow 1M(Beth), EM(Beth)\}, 1M(Beth) \rangle$

B: $\langle \{NEM(Beth) \wedge \neg Ab_2 \rightarrow <1/2M(Beth), NEM(Beth)\}, <1/2M(Beth) \rangle$

C: $\langle \{NEMX(Beth) \wedge \neg Ab_3 \rightarrow >1/2\mathrm{M}(Beth), NEMX(Beth)\}, <1/2M(Beth) \rangle$

D: $\langle \{NEM(Beth) \rightarrow Ab_1, NEM(Beth)\}, Ab_1 \rangle$

E: $\langle \{NEMX(Beth) \rightarrow Ab_2, NEM(Beth)\}, Ab_2 \rangle$

Bajo esta representación, el argumento *D* derrota *A*, *E* derrota a *B* y en consecuencia *C* cuenta como justificado, puesto que *C* es un argumento no derrotado, dando el

resultado esperado ya que la conclusión *Beth es millonaria no es obtenida*. El mismo estilo de representación puede aplicarse al ejemplo 4.1.2 y obtener el resultado correcto.

Según Prakken la razón por la que este enfoque parece mejor que simplemente invalidar el restablecimiento consiste en que si el restablecimiento debe aplicarse o no parece depender de la naturaleza del dominio, el tipo de conocimiento involucrado y el contexto en que se utiliza el conocimiento. Para ilustrar está idea propone el siguiente ejemplo del ámbito moral donde se establecen los motivos para determinar la severidad de un castigo:

Ejemplo 4.4.1

A: *Dado que John robó, le corresponde ser castigado con prisión por un periodo de hasta 6 años.*

B: *Dado a que John robó en situación de necesidad, no debería ser castigado con más de 3 años de prisión.*

C: *Dado a que John robó violentamente, le corresponde ser castigado con más de 4 años de prisión.*

En este ejemplo (al igual que en el de Beth), la conclusión del argumento preferido *C*, es lógicamente más débil que la conclusión del menos preferido. Sin embargo, sostiene Prakken, parece que la pena para un robo violento motivado por causa de pobreza merece una prisión entre 4 y 6 años. Una razón general para este resultado, sigue diciendo Prakken, podría ser que las razones que motivan la severidad del castigo no se bloquean pero si se compensan entre sí. Aunque a primera vista la observación de Prakken sobre este ejemplo parece plausible, no lo es del todo puesto que si se aplica a este caso la estrategia propuesta anteriormente, el argumento *A* no

debería contar como justificado puesto que *John robó en situación de necesidad* constituye un caso no normal para *John robó*.

El siguiente ejemplo también pretende ilustrar la idea de Prakken sobre la dependencia del restablecimiento a la naturaleza del domino, el tipo de conocimiento involucrado y el contexto donde se utiliza el conocimiento.

Ejemplo 4.4.2

A: *John dice que el sospechoso era rubio*

B: *Peter dice que el sospechoso era morocho.*

C: *Louis dice que el sospechoso no era morocho.*

Si se tiene en cuenta que Peter es más confiable que John y que Louis es más confiable que Peter, entonces el argumento *B* derrota a *A* y *C* derrota a *B*. A diferencia de lo que ocurre con el ejemplo anterior y con los ejemplos 4.1.2 y 4.1.3 en este caso parece razonable el restablecimiento de *A,* a pesar de que el argumento menos confiable tiene una conclusión más fuerte que el más confiable.

Este último ejemplo sugiere la idea de que el restablecimiento parece generar problemas cuando se pretende modelar razonamiento default y, como se ha dicho ya, se emplea un lenguaje formal para representarlo que es demasiado débil. Esto se encuentra justificado en el hecho de que a pesar de que la conclusión del argumento preferido sea más fuerte que la del más preferido (como en los ejemplos de Al y Beth), si esto se da en otros dominios (diferentes al de razonamiento default o estadístico) los ejemplos extraños no aparecen y el restablecimiento es correcto (como en el ejemplo 4.4.2).

De manera que cuando se modela información default y estadística, es necesario contar con mayor expresividad, dice Prakken, a fin de bloquear los defaults cuando esto sea conveniente. Por supuesto que esto no quiere decir que vale hacer todos los cambios que se quiera y cuando se quiera, *jugueteando* con el lenguaje formal frente a ejemplos concretos con vistas a obtener el resultado deseado. Por ello, remarca Prakken, el desafío consiste en encontrar principios generales que permitan elegir una correcta formalización.

Por todo lo dicho, Prakken concluye que el problema no se debe al restablecimiento directo en sí sino a la débil expresividad de la representación. Señala adicionalmente que esta conclusión se puede fortalecer si se encuentran otros patrones de razonamiento rebatible que necesiten mayor expresividad como los analizados hasta aquí. Efectivamente los detecta y de esta manera fortalece su posición.

El siguiente ejemplo, también debido a Horty (2001), aunque empleado por él para criticar otro mecanismo empleado en sistemas argumentativos, la separación de las fases de construcción y evaluación de los argumentos, permite destacar lo antes dicho.

Ejemplo 4.4.3

A: *Ana es adinerada porque es una abogada.*

B: *Ana es adinerada porque vive en Brentwood.*

C: *Ana no es adinerada porque es una defensora pública.*

D: *Ana no es adinerada porque renta en Brentwood.*

Los argumentos del ejemplo 4.4.3 suponen la siguiente información default: *Por lo general los abogados son adinerados; por lo general los que viven en Brentwood son adinerados; por lo general los defensores públicos no son adinerados; por lo general los*

que rentan en Brentwood no son adinerados. A su vez supone las siguientes reglas estrictas: *Todos los defensores públicos son abogados; todos los que rentan en Brentwood viven en Brentwood.*

En este ejemplo, el argumento *C* derrota estrictamente a *A* y el argumento *D* derrota estrictamente a *B*. Por otro lado, *A* y *D* se derrotan mutuamente al igual que *C* y *B*. Los sistemas *PS* y *MTDR,* que son escépticos, no concluyen nada sobre si Ana es o no adinerada. Obviamente, cualquier sistema argumentativo crédulo que modele las derrotas y los argumentos al estilo de *PS y MTDR,* obtendrían que *A y B* por un lado y *C, D* por otro, son conjuntos de argumentos aceptables. Sin embargo, en cualquiera de las dos versiones (escéptica o crédula), el resultado es erróneo, al menos desde el punto de vista intuitivo, puesto que dada la información disponible, los únicos argumentos que deberían contar como justificados son justamente *C* y *D*; y deberían ser los únicos crédulamente aceptados.

Prakken señala que esta situación es análoga al caso del restablecimiento. No se trata de que la separación de las fases sea incorrecta, sino que el lenguaje es incapaz de bloquear los defaults cuando ello es requerido. No obstante, si el lenguaje utilizado empleara cláusulas de anormalidad podrían obtenerse los resultados correctos. La representación, podría ser la siguiente. Donde $G(x)$, $D(x)$, $P(x)$, $V(x)$, $R(x)$ representan respectivamente los predicados: *abogada, adinerada, defensora pública, vive en Brentwood y renta en Brentwood*.

A: $\langle\{G(Ana) \wedge \neg Ab_1 \rightarrow D(Ana), G(\text{Ana})\}, D(\text{Ana})\rangle$

B: $\langle\{V(Ana) \wedge \neg Ab_2 \rightarrow D(Ana), V(\text{Ana})\}, D(\text{Ana})\rangle$

C: $\langle\{P(Ana) \wedge \neg Ab_3 \rightarrow \neg D(Ana), P(\text{Ana})\}, \neg D(\text{Ana})\rangle$

D: $\langle\{R(Ana) \wedge \neg Ab_4 \rightarrow \neg D(Ana), R(\text{Ana})\}, \neg D(\text{Ana})\rangle$

E: $\langle \{P(Ana) \rightarrow Ab_1, P(Ana)\}, Ab_1 \rangle$

F: $\langle \{R(Ana) \rightarrow Ab_2, R(Ana)\}, Ab_2 \rangle$

Los argumentos A y B son finalmente derrotados por los argumentos E y F y los argumentos D y C son los únicos argumentos justificados, como es intuitivamente esperable.

Teniendo en cuenta lo dicho hasta aquí, la propuesta de Prakken, que es realizada por él a modo de ilustración para justificar la necesidad de una representación con mayor poder expresivo, es bastante general puesto que no sólo soluciona los problemas frente al restablecimiento sino que es capaz de modelar adecuadamente casos como en el ejemplo 4.4.3. Sin embargo cabe aclarar lo señalado también en el trabajo citado, la investigación debe ser acompañada con la identificación de principios generales que permitan establecer una correcta representación y no elegir una representación que vaya bien frente a ejemplos concretos. Tal tarea no es sencilla. Teniendo en cuenta esto, el objetivo de este trabajo consiste en identificar algún criterio que permita la obtención de los resultados correctos sin depender de la representación o al menos sin que esta dependencia exija la modificación del lenguaje de representación en los diversos formalismos. Este abordaje, a diferencia del cambio o extensión de la representación es más sencillo y práctico de implementar y al parecer es posible detectar algunos indicios que permiten orientar la pesquisa en este terreno. El capítulo siguiente tendrá como fin desarrollar ese modelo alternativo.

4.5 Conclusión

Este capítulo tenía como objetivo presentar una serie de ejemplos que parecen poner en duda un principio ampliamente aceptado en sistemas argumentativos: *el restablecimiento*. Para ello, luego de exponer los ejemplos se instanciaron en tres

sistemas específicos: *PS* y *KT* (ya realizado en Horty, 2001) y *MTDR*. En los tres sistemas, el resultado fue similar: la obtención de resultados intuitivamente inadecuados.

Posteriormente se presentaron diversas respuestas que se han dado al problema. En general todas coinciden que el problema es debido a la deficiente capacidad de expresión del lenguaje de representación. La solución, por tanto, no consiste en el abandono del restablecimiento sino en enriquecer el lenguaje a fin de evitar la obtención de los resultados inadecuados.

Ahora bien, la modificación del lenguaje de representación sin algún principio que permita orientar la elección de tales modificaciones, parece estar destinado a realizar arreglos cada vez que aparece un ejemplo inadecuado. Por ello, Prakken (2002) sugiere que el desafío consiste en identificar principios que permitan elegir una correcta representación. Parece claro que tal tarea sería una empresa difícil. Sin embargo, los ejemplos considerados pueden ser estudiados a fin de detectar posibles causas de la aparición del comportamiento anómalo. El objetivo, tratar de identificar algún criterio que, sin ser dependiente de la representación, permita minimizar el comportamiento anómalo. En el siguiente capítulo se presentarán diversos intentos en este sentido.

Capítulo V: Vías de escape

5.1 Introducción

El restablecimiento es un principio que consiste en determinar que un argumento podrá contar como justificado cuando todos los contraargumentos del argumento en cuestión estén a su vez derrotados. Formulado así, los ejemplos problemáticos sugieren que el restablecimiento no parece un principio general. Sin embargo, los ejemplos deben ser considerados como situaciones que requieren la atención a fin de detectar las causas del comportamiento extraño más que como contraejemplos del restablecimiento. La sugerencia común en la literatura sobre el tema, tal como se ha señalado anteriormente, consiste en indicar que el problema aparece debido a un lenguaje no lo suficientemente adecuado. No obstante, el hecho de que el problema aflore cuando el razonamiento default y la especificidad se encuentran involucrados, podría sugerir alguna manera de limitar la aplicación del restablecimiento en tales casos.

La modelación del razonamiento default no sólo genera problemas en los sistemas argumentativos. Desde el surgimiento de los primeros formalismos para razonamiento no monótono se evidenciaron algunas dificultades. Recuérdese que el razonamiento default es un razonamiento que sustenta conclusiones en base a reglas de la forma *"Por lo general A es B"*, *"Típicamente A es B"* o *"Los A tienden a ser B"* denominadas reglas default en (Reiter, 1980) o reglas rebatibles en (Simari & Loui, 1992). Obviamente el sentido de las expresiones *"por lo general"* o *"típicamente"* se

emplea no en un sentido estadístico sino en un sentido prototípico o de normalidad (Reiter & Criscuolo, 1981).

Uno de los primeros enfoques definidos para un tratamiento formal de tales razonamientos fue el propuesto por Reiter, la Lógica Default (Reiter, 1980). En tal sistema se distingue entre hechos prototípicos (*prototypical facts*) y hechos estrictos (*hard facts*) sobre el mundo tales como '*por lo general los mamíferos no vuelan'* y '*todos los perros son mamíferos'* respectivamente. Los hechos prototípicos, pueden ser vistos como reglas de inferencia llamadas reglas defaults. Las reglas defaults permiten extender plausiblemente, aunque no infaliblemente, las conclusiones que pueden obtenerse a partir de las reglas estrictas.

Reiter supone que es posible pensar una teoría sobre la *realidad* constituida por un conjunto de conocimientos ciertos y un conjunto de conocimientos default. El conjunto de conclusiones que pueden obtenerse a partir de tal teoría se denomina '*extensiones*' del sistema. Las extensiones pueden ser mutuamente excluyentes (múltiples extensiones). Reiter sostiene que las múltiples extensiones de una teoría pueden ser vistas como diversas maneras de *completar* el conocimiento sobre el mundo. Empero, prontamente fue advertido que la interacción entre las reglas default entre sí, y entre reglas default y reglas estrictas, permite la obtención de resultados incorrectos (Reiter & Criscuolo, 1981), i.e. extensiones que dada la información disponible no deberían constituirse como tales.

El primer problema detectado por Reiter y Criscuolo en la Lógica Default fue que las reglas default son entendidas como transitivas. Sin embargo, no necesariamente deben serlo, puesto que ello puede llevar a la obtención de extensiones intuitivamente incorrectas.

Por ejemplo, si se tiene en cuenta las siguientes reglas defaults:

RD-1 *Por lo general los estudiantes universitarios son adultos*

RD-2 *Por lo general los adultos están casados*

RD-3 *Por lo general los estudiantes universitarios no están casados*

En la Lógica Default, que permite transitividad, se obtiene una regla que está en conflicto con la información disponible ya que de RD 1 y RD 2 se obtiene

RD-4 *Por lo general los estudiantes universitarios están casados*

De modo que si se sabe que Ray es un estudiante universitario, entonces, la Lógica Default sanciona dos extensiones, una en la que Ray es adulto y casado y otra en la que Ray es adulto pero no casado. Obviamente, la existencia de múltiples extensiones no es un problema en sí. Sin embargo, a la luz de la información disponible, a menos de que se sepa que efectivamente Ray está casado no parece razonable obtener esta conclusión.

Para evitar situaciones como las anteriores, es decir para bloquear la transitividad default cuando no corresponde, Reiter y Criscuolo (1982) idearon un mecanismo de re-representación mediante el cual las reglas que originan el problema deben ser remplazadas. Atendiendo al ejemplo considerado, la regla siguiente:

RD-5 *Por lo general los adultos que no son estudiantes universitarios están casados*

debe reemplazar a RD-2, lo que permite que la regla RD-4 no pueda obtenerse juntamente con el resultado no intuitivo de que Ray es un adulto casado.

Una situación más grave ocurre cuando las reglas default interactúan con las reglas estrictas. Tal situación permite la derivación transitiva de una regla default no intuitiva. Por ejemplo, si se sabe que

RD-6 *Por lo general las aves vuelan*

RD-7 *Por lo general los pingüinos no vuelan*

RE-1 *Todos los pingüinos son aves*

Es posible derivar de RE-1 y RD-6 la siguiente regla:

RD-8 *Por lo general los pingüinos no vuelan*

De modo que si se sabe que *Tweety es un pingüino,* la Lógica Default sanciona dos extensiones, una en la que *Tweety es ave* y *volador* y otra en la que *Tweety es ave y no vuela*. Ahora bien, dada la información disponible es claro que Tweety es un ave no voladora puesto que es un pingüino. Sin embargo, la lógica de Reiter es incapaz de bloquear RD-8 e impedir la extensión incorrecta. Por ello, Reiter y Criscuolo proponen una re-representación cuando esta situación se da mediante la siguiente reescritura de la regla RD-6:

RD-9 *Por lo general las aves que no son pingüinos son voladoras.*

De esta manera, y como no puede verificarse la propiedad de que Tweety no es un pingüino, no es posible sancionar la extensión de que Tweety sea un ave voladora.

Independientemente de la adecuación o practicidad del enfoque propuesto por Reiter y Criscuolo, hay que destacar que fue uno de los primeros trabajos que reconoció la necesidad de introducir prioridades entre defaults y que ello contribuye a remover modelos contraintuitivos (como la obtención de la conclusión de que Tweety vuele a pesar de saber que es un pingüino normal) sancionados por los formalismos para

razonamiento no monótono clásicos (Reiter, 1980; McCarthy, 1980; McDermott & Doyle, 1980; Moore, 1984)

En tal contexto, surgió la estrategia de emplear la *especificidad* como una relación de prioridad entre razonamientos defaults rivales que permite resolver conflictos causados por situaciones excepcionales. En términos generales, la especificidad puede ser entendida como un principio mediante el cual lo inferido por una subclase anula lo obtenido por una superclase.

Diversos autores han pretendido definir formalmente la especificidad y ya se ha visto en el capítulo III cómo lo hacen Simari y Loui, aunque es importante destacar que el primer significado preciso de especificidad fue dado por Touretzky en (1984). Su definición fue propuesta en el contexto de un sistema de herencias con excepciones. La idea principal consiste en lo siguiente: Si un objeto *A* hereda de un objeto *B* la propiedad "*p*" y también hereda de otro objeto *C,* la propiedad "*no p*", entonces el objeto *A* conservará la propiedad *p* si y sólo si existe un camino, i.e. una prueba, de *A* hacia *C* vía *B* pero no viceversa. *A* conservará la propiedad *p* si y sólo si *B* es una subclase de *C,* porque la subclase es siempre preferida. Un ejemplo típico de ello lo constituye *el triangulo de Tweety*.

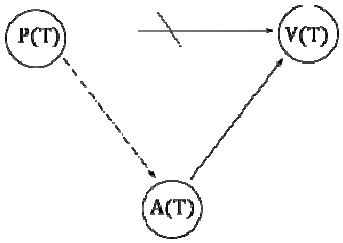


Fig. 5.1. 1: El Triángulo de Tweety

Tweety, a partir del hecho de que es pingüino, hereda la propiedad de que no vuela, mientras que a partir de que es ave, hereda la propiedad de volar. Si bien y dado que

ser pingüino es una subclase de las aves, está será siempre preferida y el sistema de herencia propuesto por Touretzky obtendrá como conclusión *Tweety no vuela* y no obtiene la conclusión de que vuela.

Por otro lado, Poole en (1985) propuso el criterio de especificidad a fin de resolver teorías complementarias en el marco del sistema *Theorist* por él propuesto. En tal sistema, las reglas defaults son utilizadas como posibles hipótesis en una teoría que busca explicar o predecir un estado de cosas. Tal estado de cosas es predecible o explicable si se sigue lógicamente de un conjunto consistente de instancias de las reglas default junto con hechos observables e instancias de un conjunto de sentencias de primer orden. Podría suceder que un estado de cosas *h* se encontrara sustentado por un conjunto *D* pero al mismo tiempo, el sistema podría contar con un conjunto *D'* que sustenta *h'*, donde *h'* consiste en la negación de *h*. Frente a tal situación Poole define el criterio de teoría más específica como mecanismo para determinar si alguna de las teorías es mejor. Una teoría *D* es más específica que *D'* si cada vez que *D'* es aplicable *D* también lo es, y existen casos en que *D* es aplicable pero *D'* no lo es.

Loui, en (1987) propone la definición de especificidad en el marco de un sistema argumentativo y fue empleada como un criterio de comparación de argumentos en conflicto. Si *A* y *B* son argumentos donde las conclusiones *Con*(*A*) y *Con*(*B*) son inconsistentes, y el fundamento de *Con*(*A*) permite deducir el fundamento de *Con*(*B*) entonces *A* es tan específico como *B*, y si *B* no es tan especifico como *A* entonces *A* es más específico que *B*.

En el sistema *MTDR* (Simari y Loui, 1992), tal como se lo ha indicado en el capítulo anterior, la especificidad permite también dirimir argumentos que no pueden ser aceptados conjuntamente. La idea general consiste en establecer que un argumento *A*

es mejor que otro *B* cuando *A* se base en mayor evidencia que *B* o que *A* emplee menos reglas que *B*. Intuitivamente, un argumento *A* será más específico que *B* cuando todo lo que activa a *A* activa también a *B* pero algo que activa a *B* no activa a *A*.

Un estudio general de la especificidad en sistemas que modelan razonamiento default puede encontrarse en (Benferhat, 2000) y un estudio sobre los principios que gobiernan el empleo de la especificidad en (Moinard, 1990).

Además del estudio de la especificidad en sistemas para razonamiento no monotónico es posible identificar la apelación a tal principio en el contexto de la filosofía de la ciencia. En particular en la propuesta de Hempel (1965) para resolver explicaciones estadísticas ambiguas. Es importante destacar que la ambigüedad estadística no solo afecta a las explicaciones inductivo estadística, sino también al uso predictivo de los enunciados estadísticos.

La especificidad consiste en un principio, ya sea aplicado a razonamiento default o razonamiento estadístico, que permite dirimir situaciones conflictivas. Ahora bien, parece ser que su empleo, que ha permitido resolver problemas en otros terrenos, trae aparejada la aparición de situaciones problemáticas en los sistemas argumentativos cuando el restablecimiento es aplicado.

El presente capítulo tiene como objetivo detectar posibles vías de escape al problema expuesto en el capítulo anterior suponiendo la validez del restablecimiento y la adecuación del lenguaje de representación en los formalismos en cuestión. De manera que las vías de escape sugeridas pretenderán conservar a los formalismos tanto como sea posible en su versión original pero que sean al mismo tiempo capaces de evitar las situaciones anómalas.

La primera vía de escape parte del supuesto de que el problema expresado en los ejemplos 4.1.2 y 4.1.3 aparece cuando los argumentos *restablecedor* y *restablecido* se relacionan por especificidad (un principio de comparación ampliamente utilizado en la modelación de razonamiento default) y sus conclusiones se implican lógicamente. Por ello, teniendo en cuenta el rasgo señalado, se procedió a definir una relación de derrota alternativa, denominada *derrota-S_1* entre argumentos consistentes que verifican tales rasgos.

La segunda vía de escape parte del supuesto de que el problema es debido al hecho de que los argumentos justificados contraintuitvamente por restablecimiento no verifican una propiedad que será denominada máxima especificidad. De manera que el objetivo consiste en introducir una relación que permita derrotar a aquellos argumentos que no son máximamente específicos. Tal derrota, llamada derrota por socavamiento (*undermining defeat*) se basa en la intuición subyacente al principio de máxima especificidad propuesto por Hempel (1965).

La tercera propuesta es un refinamiento de la idea de máxima especificidad definida en la propuesta anterior. A partir de ella, se determina un criterio de preselección de argumentos. Aquellos argumentos que no satisfacen el requerimiento son simplemente descartados. Posteriormente, los argumentos que hayan sido preseleccionados son comparados y luego se seleccionan a los argumentos justificados.

Las diversas vías de escape estructuran el capítulo en tres partes. En la sección 5.2 se presenta el mecanismo formal consistente en la introducción de la relación *derrota-S_1* entre argumentos consistentes. Se presentan ejemplos para valorar la propuesta y se señalan fortalezas y debilidades del abordaje. En la sección 5.3 se muestra la

estrategia consistente en la definición de una relación de derrota denominada '*socavamiento*' que permite derrotar a aquellos argumentos que no son máximamente específicos. Se valoran los resultados que pueden obtenerse con ella. Finalmente se propone una última estrategia, en la sección 5.4, en la que se preseleccionan argumentos mediante la satisfacción de un criterio llamado *máxima x-especificidad* y se define la justificación de argumentos en base a un conjunto de argumentos determinado excluyendo a aquellos que no satisfacen el requisito. Se comparan los resultados con las propuestas anteriores. Finalmente en la sección 5.5 se concluye.

5.2 Derrota entre argumentos consistentes

Prakken (2002) brindó las bases como para solucionar los problemas señalados en (Horty, 2001) mediante un recurso representacional. Aunque esta propuesta es aceptable, el objetivo del presente trabajo está más inclinado a dar una solución no dependiente de la modificación del lenguaje de representación. Bajo esta óptica, una propuesta puede hacerse con vistas a evitar la obtención de los resultados incorrectos en los ejemplos 4.1.2 y 4.1.3. Esto puede hacerse mediante la introducción de una relación de derrota adicional entre argumentos consistentes pero que estén relacionados de una manera particular.

Considérese el ejemplo 4.1.3 nuevamente desde un punto de vista práctico. Supóngase que Pedro quiere convencer a Juan acerca de que Beth es millonaria y lo hace argumentando sobre la base de que ella es una empleada de Microsoft (argumento *A*). Pedro podría decirle a Juan que ella posee menos de medio millón porque aunque es una empleada de Microsoft, es una empleada nueva. Ahora bien, si Pedro eligiera el argumento *C* para defender a *A* en contra del argumento *B* no parece una estrategia que permitirá ganar el debate, al menos, de manera directa. No sólo porque la conclusión de *C* (*Beth tiene al menos medio millón*) es más débil que la tesis inicial,

sino porque *C* invoca una excepción de una excepción entre la clase de los empleadas de Microsoft, que fue justamente la principal evidencia presentada por Pedro para sustentar la conclusión de que Beth es millonaria. Por ello *C* no sólo derrota a *B* sino que además socaba al argumento *A*.

De esta manera es posible señalar que *C* constituye una especie de derrota contra *A*. Obviamente, no una derrota en el sentido usual. Por ello, la pregunta inmediata es ¿cuáles son las condiciones para que tal derrota se dé? Para responderla es importante observar que el argumento *C* es estrictamente más específico que *A* y a su vez, la conclusión de *A* implica la conclusión de *C,* rasgo que puede detectarse también el ejemplo 4.1.2. Atendiendo a ese rasgo la respuesta será que habrá una relación de derrota entre argumentos consistentes cuando se verifiquen las condiciones indicadas. Con vistas a identificar esta relación será denominada como '*derrota-S_1*'.

En términos generales la *derrota-S_1* puede ser definida como una relación entre argumentos tal que si *A* y *B* son argumentos y *A* es un argumento estrictamente más específico que *B* y la conclusión de *B* implica la conclusión de *A*, se dirá que *A derrota-S_1 B.*

Con vista a valorar la propuesta esta será implementada en el contexto de *DeLP* aunque claramente puede ser aplicada a otros formalismos. La presentación general de *DeLP* se encuentra en la sección 3.5 del capítulo III.

Definición 5.2.1
(derrota-S_1)
Sean $\langle T_1,h_1\rangle$ y $\langle T_2,h_2\rangle$ dos estructuras de argumentos, se dirá que $\langle T_1,h_1\rangle$ *derrota-s_1* a $\langle T_2,h_2\rangle$ si y sólo si

i. $\langle T_1,h_1\rangle \succ \langle T_2,h_2\rangle$, y

ii. $\Pi_G \cup \{h_2\} \vdash h_1$

Donde $\succ$ es la relación de especificidad definida en *DeLP* (definición 3.5.9) y Π_G es el conjunto de reglas estrictas del programa.

Ahora bien, atendiendo a la definición 5.2.1, *DeLP* parece ser capaz de hacer frente a los ejemplos 4.1.2 y 4.1.3 obteniendo los siguientes resultados.

Ejemplo 5.2.1 (Ejemplo 4.1.2 reconsiderado)

$\langle A_1,h_1\rangle$: $\langle \{Gallina\_Salvaje(Al), Vuela(Al) \prec Ave(Al)\}, Vuela(Al)\rangle$

$\langle A_2,h_2\rangle$: $\langle \{Gallina\_Salvaje(Al), \neg Vuela(Al) \prec Gallina(Al)\}, \neg Vuela(Al)\rangle$

$\langle A_3,h_3\rangle$: $\langle \{Gallina\_Salvaje(Al), Vuela(Al) \prec Gallina\_Salvaje(Al)\}, Vuela(Al)\rangle$

Nuevamente, las relaciones de derrota usuales harán que $\langle A_3,h_3\rangle$ sea un derrotador propio para $\langle A_2,h_2\rangle$ y $\langle A_2,h_2\rangle$ un derrotador propio para $\langle A_1,h_1\rangle$ porque $\langle A_2,h_2\rangle \succ \langle A_1,h_1\rangle$ y $\langle A_2,h_2\rangle$ contraargumenta $\langle A_1,h_1\rangle$, y $\langle A_3,h_3\rangle \succ \langle A_2,h_2\rangle$ y $\langle A_3,h_3\rangle$ contraargumenta a $\langle A_2,h_2\rangle$. Pero incorporando *la derrota-s*$_1$, $\langle A_1,h_1\rangle$ no resultará justificado. Puesto que $\langle A_3,h_3\rangle \succ \langle A_1,h_1\rangle$ y claramente, $\Pi_G \cup \{h_1\} \vdash h_3$. De modo que $\langle A_3,h_3\rangle$ *derrota-s*$_1$ $\langle A_1,h_1\rangle$, permitiendo de esta manera obtener el resultado esperado, $\langle A_3,h_3\rangle$ como el único argumento justificado.

Ejemplo 5.2.2 (Ejemplo 4.1.3 reconsiderado)

$\langle A_1,h_1\rangle$: $\langle \{ME(Beth), 1M(Beth) \prec ME(Beth)\}, 1M(Beth)\rangle$

$\langle A_2,h_2\rangle$: $\langle \{NME(Beth), >1/2M(Beth) \prec NME\ (Beth)\}, >1/2M(Beth)\rangle$

$\langle A_3,h_3\rangle$: $\langle \{NEMX(Beth), <1/2M\ (Beth) \prec NMEX(Beth)\}, <1/2M\ (Beth)\rangle$

$\langle A_3,h_3\rangle$ es un derrotador propio para $\langle A_2,h_2\rangle$ y $\langle A_2,h_2\rangle$ un derrotador propio para $\langle A_1,h_1\rangle$ porque $\langle A_2,h_2\rangle \succ \langle A_1,h_1\rangle$ y $\langle A_2,h_2\rangle$ contraargumenta $\langle A_1,h_1\rangle$, y $\langle A_3,h_3\rangle \succ \langle A_2,h_2\rangle$ y $\langle A_3,h_3\rangle$ contraargumenta a $\langle A_2,h_2\rangle$. Incorporando la *derrota-s1* a *DeLP*, $\langle A_1,h_1\rangle$ no constituye un argumento justificado igual que en el ejemplo 4.1.2. Puesto que $\langle A_3,h_3\rangle \succ \langle A_1,h_1\rangle$ y $\Pi_G \cup (h_1) \vdash h_3$. De modo que $\langle A_3,h_3\rangle$ *derrota-s1* $\langle A_1,h_1\rangle$. En este ejemplo, *DeLP* puede obtener el resultado deseable, $\langle A_3,h_3\rangle$ es el único argumento justificado.

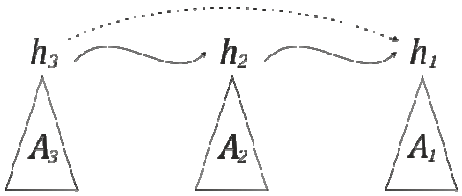


Fig. 5.2. 1: Representación común de los ejemplos 4.1.2 y 4.1.3. bajo S1. Las flechas sólidas representan la *derrota propia* y la flecha punteada a *S1*

En los ejemplos precedentes, S_1 permite obtener los resultados esperados sin involucrar un incremento en el lenguaje y tampoco invalidando el restablecimiento. Ahora bien, ¿podrá considerarse S_1 como una relación de derrota general? Es decir, S_1 ¿no es acaso una propuesta *ad hoc* que simplemente permite resolver casos similares al de los ejemplos 4.1.2 y 4.1.3? Dado que la propuesta se encuentra basada más bien en un rasgo común de los ejemplos problemáticos y no se cuenta con una justificación de porqué el problema aparece cuando tales condiciones se dan, la propuesta parece *ad hoc*. Sin embargo, el abordaje podría ser valorado positivamente atendiendo al resultado que permite obtener frente al siguiente ejemplo similar al propuesto en (Horty, 2012).

Ejemplo 5.2.3

Imagínese que un biólogo se encuentra estudiando la distribución de las aves en una remota zona de islas donde ha identificado una especie de canarios llamada *canarios*

del cabo. Por un particular comportamiento de la especie, los nidos se encuentran distribuidos en su mayoría aunque no únicamente en la isla Florida. Una subespecie de estos canarios, llamados *canarios del cabo rojo,* por su parte, tiene, por lo general, sus nidos distribuidos en la isla Florida o en la isla Desierta. Ahora bien, considerando a un individuo particular de tal especie, llamado Frank, que se sabe es un canario del cabo rojo ¿Qué podrá concluir el biólogo con respecto de la ubicación del nido de Frank? ¿Qué conclusiones podrían obtenerse en los sistemas argumentativos? A partir de la anterior información podrían construirse los siguientes argumentos.

A: *El nido de Frank se encuentra en la isla Florida porque Frank es un canario del cabo.*

B: *El nido de Frank se encuentra en la isla Florida o en la isla Desierta porque Frank es un canario del cabo rojo*

Bajo la óptica de los sistemas argumentativos en general, ambos argumentos estarán justificados puesto que no media entre ellos una relación de conflicto, necesaria para activar la relación de derrota. No obstante, intuitivamente, es esperable que el único argumento justificado sea *B*. Ahora bien, si se tiene en consideración la relación de derrota definida anteriormente (S_1) el argumento *A* se encuentra derrotado por *B* dado que *B* es estrictamente más específico que *A* y la conclusión de *A* implica la conclusión de *B*.

Aunque el comportamiento de S_1 frente al ejemplo anterior hace pensar que constituye una buena herramienta para hacer frente a casos como ese, S_1 sanciona argumentos no justificados cuando en realidad no necesariamente deberían estarlo:

Ejemplo 5.2.4

A: *Timoteo vuela porque es un ave.*

B: *Timoteo vuela porque es un tero.*

Dado que *B* es más específico que *A* y la conclusión de *A* implica la conclusión de *B* se puede aplicar la derrota por S_1. De modo que *B* derrota a *A*. Sin embargo, mientras que en los otros ejemplos parece tener sentido tal derrota, en este caso no lo es. En particular porque el hecho de que Timoteo sea un tero no es una evidencia que muestra que Timoteo sea un ave no normal, al menos con respecto a volar. Esto podría sugerir la siguiente modificación de S_1.

Definición 5.2.2

(derrota-S_1 revisada)

Sean $\langle T_1,h_1\rangle$ y $\langle T_2,h_2\rangle$ dos estructuras de argumentos, se dirá que $\langle T_1,h_1\rangle$ *derrota-*s_1 a $\langle T_2,h_2\rangle$ si y sólo si:

i. $\langle T_1,h_1\rangle \succ \langle T_2,h_2\rangle$, y

ii. $\Pi_G \cup \{h_2\} \vdash h_1$ y $\Pi_G \cup \{h_1\} \nvdash h_2$

□

La condición *ii*. ahora sólo exige que el argumento más general implique la conclusión del más específico dejando fuera los casos dónde la conclusión de ambos argumentos es equivalente. Aunque la definición 5.2.2 puede tener un comportamiento adecuado frente al ejemplo 5.2.4, el ejemplo 4.1.2 no podría ser modelado correctamente. Por otro lado, el ejemplo 4.4.3, que puede ser modelado correctamente en la propuesta de Prakken, tampoco puede ser abordado correctamente ni por S_1 ni por S_1 revisada.

La propuesta no parece lo suficientemente general, al menos como para proponerse como alternativa a la versión representacional, pero es interesante el resultado obtenido frente al ejemplo 5.2.3.

5.3 Socavamiento entre argumentos

La propuesta anterior (definición 5.2.1) permite obtener los resultados esperados frente a los ejemplos 4.1.2 y 4.1.3. A pesar de ello, la misma no es demasiado tolerante dado que derrota argumentos innecesariamente como en el ejemplo 5.2.4. Por otro lado tampoco es lo suficientemente general, al menos no tan general como la propuesta representacional de Prakken dado que no puede dar cuenta de ciertos casos, como el ejemplo 4.4.3. A su vez, S_1 es notablemente una propuesta *ad hoc*. La modificación ensayada (definición 5.2.2) tampoco ayuda en este aspecto. Cabe resaltar que S_1 permite obtener un resultado interesante frente a situaciones como en el ejemplo 5.2.3. Con vistas a obtener otra propuesta alternativa y que subsane los inconvenientes de S_1, se define la siguiente.

La segunda vía de escape parte del supuesto de que la raíz del problema está en que la relación de derrota definida en los sistemas argumentativos considerados permite aceptar ciertos argumentos que, como se verá más adelante, no son *máximamente específicos*. La estrategia se inspira en el principio de máxima especificidad propuesto por Hempel para hacer frente al problema de la ambigüedad estadística. Básicamente consistirá en introducir un *requerimiento de máxima especificidad* como criterio de demarcación para la aceptación de argumentos. Este criterio puede ser usado para filtrar a aquellos argumentos que no son máximamente específicos con respecto a las conclusiones que sustentan y que justamente son los responsables de llevar a explicaciones irrelevantes o a sustentar conclusiones incorrectas como en los ejemplos 4.1.2 y 4.1.3.

Tal como se dijo, el enfoque se basa en la noción de *máxima especificidad* propuesta por Hempel como recurso para resolver el problema de la ambigüedad estadística. Con vistas a comprender mejor la idea, la misma será definida a continuación. Para iniciar la discusión se expone el siguiente ejemplo propuesto en (Hempel, 1965) que permite ilustrar el problema de la ambigüedad estadística.

Ejemplo 5.3.1

A: *Supóngase que John sufre de una grave infección por estreptococo y es tratado con penicilina. Adicionalmente supóngase que la probabilidad de recuperarse de una grave infección por estreptococos tratada con penicilina es cercana a 1. De modo que es prácticamente seguro o muy probable que John se recuperará.*

Tal como lo advierte Hempel (1965), la ley estadística invocada en este caso afirma la recuperación por los efectos de la penicilina sólo para un alto porcentaje de las infecciones por estreptococos pero no para todas ellas. Por caso considérese la siguiente situación.

Ejemplo 5.3.2

B: *Supóngase que John tiene una infección por estreptococos pero de una variedad resistente a la penicilina. Adicionalmente, téngase en cuenta que la probabilidad de recuperación de una grave infección por estreptococos resistente a penicilina es cercana a 0, de modo que la probabilidad de la no recuperación será cercana a 1. Por lo tanto, es prácticamente seguro o muy probable que John no se recuperará.*

El razonamiento formulado, *B,* es un razonamiento rival para *A*, puesto que mientras que la conclusión *A* sanciona una conclusión que establece que es prácticamente seguro o muy probable que John se recuperará, la conclusión de *B* afirma que es prácticamente seguro o muy probable que John no se recuperará. También podría formularse otro ejemplo similar al anterior y que sea también un razonamiento rival para el primero.

Ejemplo 5.3.3

*B** *Supóngase que John tiene una infección por estreptococos y John es un octogenario de corazón débil. Adicionalmente, téngase en cuenta que la probabilidad de recuperación de una grave infección por estreptococos en octogenario de corazón débil es muy pequeña, cercana a 0, de modo que la probabilidad de la no recuperación será cercana a 1. Por lo tanto, es prácticamente seguro o muy probable que John no se recuperará.*

Es importante notar que tanto *B** como *B* son argumentos rivales para *A* y en todos los casos sustentan su conclusión en premisas verdaderas. De modo que los anteriores ejemplos ilustran el hecho de que para una explicación probabilística propuesta (o para una predicción basada en enunciados estadísticos) que confieren la casi seguridad de un suceso particular, habrá a menudo un razonamiento rival de la misma forma probabilística y con premisas igualmente verdaderas que confiere la casi seguridad de la no producción del mismo hecho.

Tal situación sugiere el siguiente planteo: ¿En cuál de los razonamientos antagónicos se puede confiar racionalmente para la explicación o la predicción?

Según Hempel, la decisión sobre la aceptabilidad de una explicación o una predicción probabilística propuesta deberá tomarse en base al *requisito de los elementos de juicio totales.* Este requisito, que es una regla de aplicación de la lógica inductiva más que un principio lógico, exige que se tengan en cuenta todos los enunciados aceptados en una situación dada, cuando se quieren obtener inferencias estadísticas. Tales enunciados aceptados será representado en un conjunto K.

Por ejemplo, la explicación de la recuperación de John basada en la información de que tuvo una infección por estreptococos y fue tratado con penicilina, y de que la probabilidad estadística de recuperación en tales casos es muy alta, *es inaceptable* si se incluye la información adicional de que los estreptococos que afectan a John son resistentes a la penicilina o de que es un octogenario de corazón débil y que en estas clases de referencias la probabilidad de recuperación es muy pequeña.

Obviamente, tal como lo sugiere Hempel, sería deseable que una explicación *aceptable* se basara en un enunciado de probabilidad estadística perteneciente a la *más restringida clase de referencia* de la cual sea miembro el hecho particular en consideración, según la información total disponible. De modo que, si en K se cuenta no solamente con la información de que John tuvo una infección por estreptococos y que fue tratado con penicilina, sino que además es un octogenario de corazón débil (y si K no brinda ninguna información más específica), una explicación aceptable de la respuesta al tratamiento de John debe estar soportada en una base estadística que formulase la probabilidad de esta respuesta en la más restringida clase de referencia a la cual la información total disponible asigna la enfermedad de John, es decir, la clase de las infecciones por estreptococos que sufren los octogenarios de corazón débil.

Una explicación estadística será por tanto aceptable cuando satisfaga el requisito de máxima especificidad (RME). Si E^* es una explicación estadística

$$E^* \qquad \begin{array}{l} p(G, F) = r \\ \hline\hline Fb \\ Gb \end{array} r$$

y s es la conjunción de las premisas y si K es el conjunto de todos los enunciados aceptados en el tiempo t, donde k es una proposición equivalente a K, se dirá que tal explicación es *máximamente específica* si $s \cdot k$ implica que b pertenece a una clase F_1, entonces $s \cdot k$ debe también implicar un enunciado que especifique la probabilidad estadística de G en F tal que $p(G,F_1)=r_1$ donde r_1 es igual r, a menos que el enunciado de probabilidad citado sea simplemente un teorema de la teoría matemática de la probabilidad.

Lo que la máxima especificidad requiere es que si algo puede ser inferido a partir de una explicación máximamente específica de una clase F teniendo en consideración la evidencia total K, también puede ser inferido a partir de alguna subclase F' de F con la misma probabilidad.

Si se consideran los ejemplos 4.1.2 y 4.1.3 a la luz del requerimiento de máxima especificidad podría pensarse que aquellos argumentos que no son máximamente específicos no deberían considerarse como argumentos aceptables. Ahora bien, ya que la extrapolación de tal principio a la argumentación rebatible no puede hacerse de manera directa, puesto que las inferencias no son obtenidas con medidas probabilísticas ¿cómo saber que un argumento es máximamente específico?

La propuesta consistirá en definir la *máxima especificidad* sobre una clase G cuando lo que permite inferirse de tal clase, al menos no sea contradictorio con la conclusión obtenida por cualquier subclase H de G en consideración de la evidencia total. Esto en términos de los sistemas argumentativos exigirá que para considerar a un argumento como máximamente específico, tal argumento debe ser libre de contraargumentos más específicos. Esto es así porque los contraargumentos más específicos son indicativos de que la subclase permite inferir una conclusión contradictoria a lo que puede inferirse en base a la clase.

Retomando el ejemplo 4.1.3, si el argumento *A*

A: *Dado que Beth es una empleada de Microsoft y teniendo en cuenta que tales empleadas tienden a ser millonarios es posible concluir que Beth es millonaria.*

fuese máximamente específico, entonces lo que se puede inferir a partir de que Beth es una empleada de Microsoft, al menos no debe ser contradictorio con la conclusión obtenida a partir del dato de que Beth es una nueva empleada de Microsoft, donde claramente las nuevas empleadas de Microsoft son una subclase de los empleados de Microsoft.

Por otro lado, si el argumento *B,* a saber,

B: *Dado que Beth es una nueva empleada de Microsoft y teniendo en cuenta que tales empleadas por lo general poseen menos de medio millón es posible concluir que Beth posee menos de medio millón.*

fuese un argumento máximamente específico entonces no debería poder inferirse de una subclase una conclusión contradictoria, como efectivamente sucede con el argumento C:

C: *Dado que Beth es una nueva empleada de Microsoft en el departamento X y teniendo en cuenta que tales empleadas por lo general poseen al menos medio millón se puede concluir que Beth posee al menos medio millón.*

puesto que *los nuevos empleados de Microsoft en el departamento X* son una subclase de *los empleados nuevos de Microsoft.*

Como se habrá notado, la evidencia o datos '*ser una nueva empleada de Microsoft en el departamento X' y* '*ser una nueva empleada de Microsoft'* es lo que permite detectar si un argumento no es máximamente específico en el ejemplo de *Beth.* Esto datos constituirán lo que a partir de ahora será denominado '*evidencia socavadora',* en el sentido de que tal evidencia permite que un argumento más específico infiera una conclusión que contradice lo sostenido por un argumento determinado. Ahora es posible decir que un argumento *A* será considerado máximamente específico cuando no exista evidencia socavadora para *A*.

Atendiendo a este análisis es posible observar que el único argumento que satisface el criterio de ser máximamente específico, en el ejemplo 4.1.3, es el argumento *C.* El mismo análisis podría realizarse para el ejemplo 4.1.2 obteniéndose el mismo resultado.

Lo anterior parece sugerir la idea de que el problema aparece en los sistemas considerados en el capítulo anterior porque no cuentan con un mecanismo capaz de impedir que los argumentos que no son máximamente específicos cuenten como justificados. La intuición puede verse fortalecida si se considera el ejemplo 4.4.2 donde el restablecimiento no lleva a la aceptación de un argumento problemático y en tal ejemplo cada argumento satisface la condición de ser máximamente específico. El ejemplo 4.1.1 también verifica la propiedad. Por otro lado, si se tiene en cuenta el

ejemplo 4.4.3, que en realidad no es un resultado contraintuitivo con respecto al restablecimiento pero supone un desafío a los sistemas argumentativos, los argumentos: *Ana es adinerada porque es una abogada* y *Ana es adinerada porque vive en Brentwood* son claramente argumentos que no son máximamente específicos.

Con vistas a ilustrar la propuesta en un sistema argumentativo específico, la idea será implementada en *DeLP*. Obviamente es posible que sea aplicada a otros formalismos pero se escoge ése por su capacidad y facilidades expresivas relativas al fin que propuesto.

Para definir los argumentos máximamente específicos en *DeLP* se considerará a la información representada en Π como la información total disponible. Los argumentos máximamente específicos en este sistema serán aquellos que al menos no cuenten con *derrotadores propios* dado que estos son indicativos de que una subclase permite inferir una conclusión contradictoria de lo que puede inferirse en base a la clase de referencia.

Antes de presentar la definición será conveniente recordar que la información no rebatible se encuentra contenida en el conjunto Π. Este conjunto está particionado en dos subconjuntos disyuntos Π_F y Π_G. En Π_F se encuentran los hechos y en Π_G reglas estrictas. También es importante recordar que la unión de los conjuntos Π_G y T, donde T es un subconjunto de Δ, de reglas rebatibles, no pueden inferir una conclusión sin un subconjunto Π_F que permite activar las reglas estrictas y rebatibles.

Atendiendo a lo dicho, se procede a brindar la definición de lo que sido llamado '*evidencia socavadora*'.

Definición 5.3.1

(evidencia socavadora)

Sea $\mathcal{P}$ = (Π,Δ) un *delp*, y sea F un subconjunto de Π_F, se dice que F es *evidencia socavadora* de $\langle T,h\rangle$ si $F \cup T' \cup \Pi_G \vdash h'$ para algún derrotador propio $\langle T',h'\rangle$ de $\langle T,h\rangle$ (i.e. F "activa" algún derrotador propio del argumento).

□

En la definición 5.3.1 se considera que un subconjunto F del conjunto de hechos Π_F es evidencia socavadora de un argumento cualquiera cuando tal subconjunto permite activar a algún derrotador propio de tal argumento, i.e. cuando tal subconjunto de hechos permite activar a un contraargumento más específico.

En el siguiente ejemplo se ilustra el comportamiento de la definición.

Ejemplo 5.3.1 (ejemplo 4.2 reconsiderado)

Π_F={*Gallina_Salvaje*(*Al*)}

Π_G={*Gallina*(*x*) ⟵*Gallina_Salvaje*(*x*);*Ave*(*x*) ⟵*Gallina*(*x*)}

Δ={¬*Vuela*(*x*) ⇁*Gallina*(*x*); *Vuela*(*x*) ⇁*Ave*(*x*); *Vuela*(*x*) ⇁*Gallina_Salvaje*(*x*)}

$\langle T_1,h_1\rangle$: $\langle${*Gallina_Salvaje*(*Al*), *Vuela*(*Al*) ⇁ *Ave*(*Al*)}, *Vuela*(*Al*)$\rangle$

$\langle T_2,h_2\rangle$: $\langle${*Gallina_Salvaje*(*Al*), ¬*Vuela*(*Al*) ⇁ *Gallina*(*Al*)}, ¬*Vuela*(*Al*)$\rangle$

$\langle T_3,h_3\rangle$: $\langle${*Gallina_Salvaje*(*Al*), *Vuela*(*Al*) ⇁ *Gallina_Salvaje*(*Al*)}, *Vuela*(*Al*)$\rangle$

Dado que *Gallina_Salvaje*(*Al*) implica *Gallina*(*Al*) y *Gallina*(*Al*) permite activar un derrotador propio de $\langle T_1,h_1\rangle$, i.e. $\langle T_2,h_2\rangle$, *Gallina*(*Al*) es evidencia socavadora para $\langle T_1,h_1\rangle$. Adicionalmente, *Gallina_Salvaje*(*Al*) es evidencia socavadora de $\langle T_2,h_2\rangle$ dado que activa $\langle T_2,h_2\rangle$ y al mismo tiempo a un derrotador propio $\langle T_3,h_3\rangle$.

Tal cómo anteriormente se ha dicho, un argumento será máximamente específico cuando sea libre de evidencia que lo socave. Formalmente en *DeLP*:

Definición 5.3.2
(máxima especificidad)
Sea $\boldsymbol{\mathcal{P}}$ = (Π,Δ) un *delp*. Se dice que una estructura de argumento $\langle T, h\rangle$ es *máximamente especifica* (c.r. a h) en $\boldsymbol{\mathcal{P}}$ si y sólo si no existe evidencia socavadora para $\langle T, h\rangle$ en $\boldsymbol{\mathcal{P}}$.

□

Ejemplo 5.3.2 (ejemplo 4.1.2 reconsiderado)
Claramente sólo
$\langle T_3,h_3\rangle$: $\langle\{Gallina\_Salvaje(Al), Vuela(x) \prec Gallina\_Salvaje(x)\}, Vuela(Al)\rangle$
es el único argumento máximamente específico puesto que $Gallina\_Salvaje(Al)$ es evidencia socavadora para

$$\langle\{Gallina\_Salvaje(Al), Vuela(x) \prec Ave(x)\}, Vuela(Al)\rangle$$

y para

$$\langle\{Gallina\_Salvaje(Al), \neg Vuela(x) \prec Gallina(x)\}, \neg Vuela(Al)\rangle$$

Requerir *máxima especificidad* (MS) como una condición para un argumento justificado, implica rechazar cualquier argumento que tenga algún derrotador propio. Es importante notar que no importa si el derrotador propio esté o no derrotado. Para rechazar argumentos que no son máximamente específicos sólo basta con que cuenten con derrotadores propios dado que será una manera de detectar la evidencia socavadora. En el ejemplo 4.1.3 sólo el argumento C satisface MS con respecto a $>1/2M(Beth)$ (*Beth posee al menos medio millón*) pero ni A ni B satisfacen MS porque $NEMX(Beth)$ constituye evidencia socavadora para ambos. Lo visto en esto último y lo señalado en el ejemplo anterior motiva el siguiente lema en *DeLP*:

Lema 5.3.1. *Sean* ⟨*T, h*⟩ *y* ⟨*T', h'*⟩ *dos estructuras de argumento tales que* ⟨*T, h*⟩ *es un derrotador propio de* ⟨*T', h'*⟩*. Si H es evidencia socavadora para* ⟨*T, h*⟩ *entonces H es evidencia socavadora para* ⟨*T', h'*⟩*.*

□

Prueba. Sea ⟨*T, h*⟩ un derrotador propio de ⟨*T', h'*⟩, y sea *H* la evidencia socavadora para ⟨*T, h*⟩. Entonces *H* activa algún derrotador propio ⟨*S, j*⟩ de ⟨*T, h*⟩. Dado que ⟨*S, j*⟩ es más específico que ⟨*T, h*⟩, H activa ⟨*T, h*⟩ y dado que ⟨*T, h*⟩ es más específico que ⟨*T', h'*⟩, *H* también activa a ⟨*T', h'*⟩. Entonces, por definición 5.3.1, H es evidencia socavadora para ⟨*T', h'*⟩.

■

Como puede advertirse, la existencia de un contraargumento más específico es una condición suficiente para que exista *evidencia socavadora*.

Aunque hasta ahora el enfoque parece prometedor es importante destacar que si el mismo no es expresado en términos de una relación de derrota, la propuesta no puede ser implementada en un sistema argumentativo. Esto es así, porque el proceso de justificación de argumentos se hace apelando a las relaciones de derrota, a menos que se cambie el proceso de justificación. No obstante, el objetivo de la identificación de vías de escape se encuentra basado en la idea de conservar los formalismos tanto como sea posible. Por ello, una vez definida la noción de '*evidencia socavadora*' y '*argumento máximamente específico*' puede ahora introducirse una derrota alternativa, denominada '*derrota por socavamiento'*.

Definición 5.3.3
(derrota por socavamiento)

Sea $\mathcal{P}$ = (Π,Δ) un *delp*. Sean ⟨*T, h*⟩ y ⟨*T', h'*⟩ dos estructuras de argumento. Se dice que ⟨*T, h*⟩ es un *derrotador por socavamiento* de ⟨*T',h'*⟩ si y sólo si para cualquier subconjunto de hechos $F \subseteq \Pi_F$, si $F \cup T \cup \Pi_G \vdash h$ entonces $F \cup S \cup \Pi_G \vdash j$ para algún derrotador propio ⟨*S, j*⟩ de ⟨*T', h'*⟩ (i.e. si *F* activa ⟨*T,h*⟩ entonces *F* también activa algún derrotador propio de ⟨*T', h'*⟩). También se definirá $def_{und}(Args) =_{df} \{(A,B) : A,B \in Args$ y *A* es un derrotador por socavamiento de *B*}. Donde *Args* es un conjunto de estructuras argumentativas.

□

La definición 5.3.3 establece que un argumento *A* será un derrotador por socavamiento de *C* cuando el conjunto de hechos que activan a *A* también activan a un argumento *B*, y *B* es un contraargumento más específico de *C.* Obviamente que *C* podría ser el mismo argumento *A*.

La *derrota por socavamiento* propuesta no es una derrota por rebatimiento, ya que puede ser el caso que la aceptación conjunta de dos argumentos no lleve a contradicción, pero que exista una relación por socavamiento. En el ejemplo 4.1.3 el argumento *C* es un derrotador por socavamiento del argumento *A*, pero *C* no es un derrotador por rebatimiento de *A*. Esto es así porque, por el lema 5.3.1 y la definición 5.3.3, la evidencia socavadora se propaga mediante encadenamiento de derrotadores propios. Esto será enunciado en el lema 5.3.2.

Por otro lado, esta derrota también puede ser vista como un caso de *undercutting defeater* (Pollock, 1995) al menos indirectamente, dado que su aceptación implica el uso de la evidencia total que brinda razones que hacen la conclusión del argumento derrotado no inferible.

Lema 5.3.2 *Sea ⟨T,h⟩ un derrotador propio de ⟨T',h'⟩ entonces para cualquier derrotador propio ⟨S, j⟩ de ⟨T,h⟩, ⟨S,j⟩ es un derrotador por socavamiento de ⟨T',h'⟩.*

□

Prueba. Inmediato por definición 5.3.3 y lema 5.3.1.

■

Una consecuencia interesante del lema 5.3.2 consiste en que, en ciclos de derrotas propias en *DeLP*, la sucesión de derrotas por socavamiento son indicativas de la evidencia socavadora para todos los argumentos involucrados en el ciclo incluyéndose a sí mismos como se verá más adelante.

Los lemas 5.3.1 y 5.3.2 llevan al siguiente teorema:

Teorema 5.3.1. Sea *Args* un conjunto de estructuras argumentativas y sea $def_{prop}(Args) =_{df} \{(A,B) : A,B \in Args$ y A es un derrotador propio de $B\}$ y $def_{prop}(Args)^{tr}$ la clausura transitiva de $def_{prop}(Args)$. Entonces, $def_{und}(Args) = def_{prop}(Args)^{tr}$. Donde $def_{prop}(Args) =_{df} \{(A,B) : A,B \in Args$ y A es un derrotador propio de $B\}$.

□

Prueba. Inmediato por lema 5.3.1 y 5.3.2

■

Los ejemplos 4.1.2 y 4.1.3 pueden ser modelados correctamente puesto que la noción de derrota por socavamiento permite rechazar los argumentos problemáticos. Esto puede ser ilustrado en la figura 5.3.1, como una representación común de ambos ejemplos donde el único argumento finalmente no derrotado es el esperable.

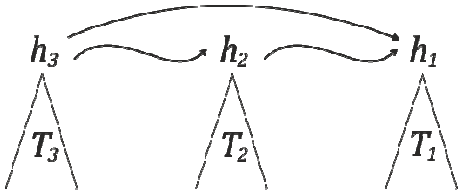


Fig. 5.3. 1: Derrota por socavamiento.

En el teorema 5.3.1 subyace un principio que permite aplicar el enfoque tanto al sistema *PS* como a *KT.* Lo único necesario será entonces, para definir la derrota por socavamiento en tales sistemas, propagar la derrota mediante encadenamientos de derrotadores ya sea mediante la versión estricta de la derrota *por rebatimiento* en *PS* (definición 3.2.5) o la propagación de la noción de derrota en *KT*.

Anteriormente se dijo que la *derrota por socavamiento* tiene un resultado interesante cuando es aplicada en *DeLP*. En particular en lo que respecta a los ciclos de derrotas propias entre argumentos. A continuación se presentará el nuevo comportamiento de *DeLP* frente a los casos de derrota recíproca, ciclos impares y pares cuando la derrota por socavamiento es considerada.

Dos argumentos se dicen *recíprocamente derrotados s*i y sólo si se da la siguiente situación: Sean $\langle T,h\rangle$ y $\langle S,j\rangle$ dos estructuras de argumento tales que $\langle S,j\rangle$ es un derrotador propio de $\langle T',h'\rangle$ y $\langle T,h\rangle$ es un derrotador propio de $\langle S',j'\rangle$ donde $\langle T',h'\rangle$ es un subargumento propio de $\langle T,h\rangle$ y $\langle S',j'\rangle$ lo es de $\langle S,j\rangle$. Mediante la incorporación de derrota por socavamiento las derrotas reciprocas llevan inevitablemente a la autoderrota de los argumentos en consideración. Esto se ilustra en la siguiente figura:

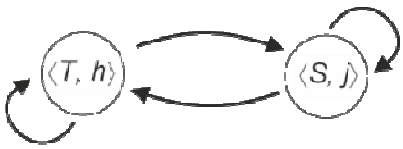


Fig. 5.3. 2: Derrota recíproca

Ante esta situación, ⟨*T, h*⟩ socava a ⟨*S, j*⟩ y dado que ⟨*S, j*⟩ es un derrotador de socavamiento para ⟨*T, h*⟩ en ⟨*T', h'*⟩, entonces ⟨*T, h*⟩ es un derrotador por socavamiento para ⟨*T, h*⟩. Lo mismo para ⟨*S, j*⟩ de manera que ambos argumentos no pueden estar justificados porque están autoderrotados. Como en *DeLP* no es esperable la autoderrota, puesto que los argumentos son estructuras consistentes, será preciso definir un proceso de justificación acorde a esta situación. No obstante, esto será sencillo de realizar mediante un juego argumentativo que será presentado luego.

Ahora considere la siguiente situación: Sea Λ = [⟨T_1,h_1⟩,⟨T_2,h_2⟩,⟨T_3,h_3⟩,⟨T_1,h_1⟩] *una línea de argumentación circular* en *DeLP,* tal que el ciclo impar es definido por derrotas propias. Bajo la noción de derrota por socavamiento, cada argumento derrotaría a cada uno de los argumentos que pertenece a la línea y cada argumento, finalmente, se derrotaría a sí mismo.

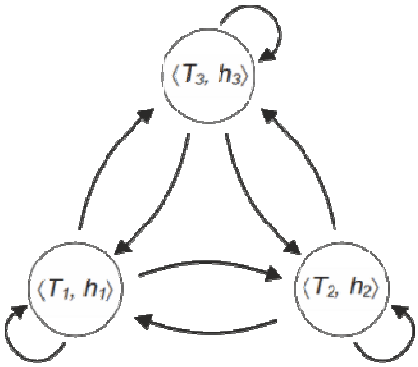


Fig. 5.3. 3: Ciclos impares

Por otro lado, si la línea fuese par, como por ejemplo, Λ = [⟨T_1,h_1⟩,⟨T_2,h_2⟩,⟨T_3,h_3⟩,⟨T_4,h_4⟩,⟨T_1, h_1⟩] entonces el resultado puede ser ilustrado en la figura 5.3.4.

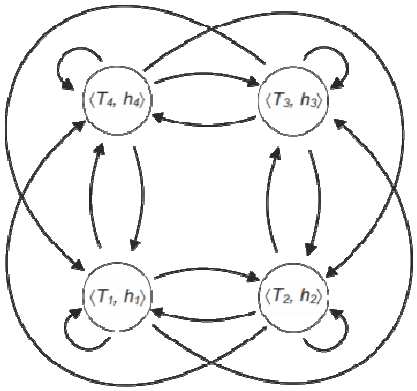


Fig. 5.3. 4: Derrotas en ciclos pares

Dado que el proceso de justificación definido en *DeLP* no está definido para manipular argumentos autoderrotados exigirá la introducción de un proceso de justificación alternativo al originalmente propuesto. Tal proceso será definido en base a un juego argumentativo. Primeramente será necesario definir un marco argumentativo asociado a *DeLP* donde se establezcan las relaciones de derrota que puedan darse entre los argumentos que pertenecen al conjunto *Args.*

Definición 5.3.4

(marco argumentativo asociado a *DeLP*)

Dado un *delp* $\mathcal{P}$, el *marco argumentativo asociado a* $\mathcal{P}$ es un par $\langle Args, derrota \rangle$ donde *Args* es el conjunto de todas las estructuras de argumento obtenidas a partir de $\mathcal{P}$ y $derrota = \bigcup DEF(Args)$, donde $DEF(Args) = \{def_1, \ldots, def_k\}$, es el conjunto que contiene cualquier criterio de derrota $def_i \subseteq Args \times Args$ $(1 \leq i \leq n)$ definido en *Args*.

Se asume que $def_{prop}(Args) \in DEF(Args)$ al igual que $def_{und}(Args) \in DEF(Args)$.

□

Definición 5.3.5 **[Bodanza et al., 2012]**

(juego argumentativo)

Un *juego argumentativo* en un marco argumentativo es un juego extensivo de suma cero en el que:

1. Existen dos jugadores, i y $-i$, quienes juegan los roles de **P** y **O** respectivamente.
2. Una *historia* en el juego es cualquier secuencia $A_0, A_1, A_2,\ldots,A_{2k},A_{2k+1},\ldots$ de elecciones de argumentos en *Args* hecha por los jugadores en el juego. A_{2k} corresponde a **P** y A_{2k+1} a **O**, para $k = 0,1,\ldots$ En cualquier historia, A_0 es el argumento que **P** intenta defender.
3. En una historia, las elecciones del jugador i en el nivel $k > 0$ son $C_i(k) = \{A \in Args: \exists B \in C_{-i}(k-1), (A,B) \in derrota\}$.
4. Una historia de longitud finita $K, A_0,\ldots,A_K$ es terminal si A_K corresponde al jugador i ($j = i$ o $j = -i$) y $C_{-j}(k+1) = \varnothing$.
5. Los pagos son determinados en las historias terminales: el pago para **P** en $A_0,\ldots,A_K$ es 1 (gana) si K es par (i.e. **O** no puede responder al último argumento de **P**) sino −1 (pierde). A su vez, el pago en $A_0,\ldots,A_K$ para **O** es 1 si K es impar, sino -1.

□

Un juego en el que **P** intenta defender un argumento A pude entenderse como un árbol en el que A es la raíz. Cada nodo no terminal en el nivel l consiste de una historia $A_0,\ldots,A_l$ y sus hijos son todas las historias en $A_0,\ldots,A_l, A_{l+1}$. Los nodos terminales son las historias terminales. Con vistas a representar un criterio de justificación escéptico se agregará la restricción siguiente:

Juego escéptico: El proponente **P** no tiene permitido jugar un argumento que ha sido previamente jugado por cualquier jugador.

Definición 5.3.6 **[Bodanza et al., 2012]**

(estrategia)

Una *estrategia* para un jugador *i* es una función que asigna un elemento $A_{l+1} \in C_i(l)$, a cada historia no terminal $A_0,\ldots,A_l$ donde A_l corresponde al jugador $-i$. Una estrategia del jugador *i* se denomina *estrategia ganadora* para *i* si para cada estrategia elegida por $-i$, la historia terminal produce un pago 1 para el jugador *i*.

□

Definición 5.3.7 **[Bodanza et al., 2012]**

(justificación escéptica)

Dado un marco argumentativo ⟨*Args*, *derrota*⟩ un argumento *A* está *justificado escépticamente* si y sólo si **P** tiene una estrategia ganadora para defender a *A* en un juego escéptico en ⟨*Args*, *derrota*⟩.

□

Dotando a *DeLP* de las anteriores definiciones, permite resolver adecuadamente los ejemplos 4.1.2 y 4.1.3. Adicionalmente está propuesta no introduce derrotas innecesarias como lo hace S_1 en ejemplos como 5.2.4, destacando un aspecto favorable. Sin embargo, la propuesta no es capaz de dar cuenta del ejemplo de Ana:

Ejemplo 5.3.3 (ejemplo 4.4.3 reconsiderado)

A: *Ana es adinerada porque es una abogada.*

B: *Ana es adinerada porque vive en Brentwood.*

C: *Ana no es adinerada porque es una defensora pública.*

D: *Ana no es adinerada porque renta en Brentwood.*

Como puede observarse, claramente *D* y *C* son máximamente específicos mientras que *A* y *B* no lo son, y aunque *C* es un derrotador por socavamiento para *A* y *D* lo es para *B*, no existe alguna estrategia ganadora para **P** en caso de pretender defender *C* o *D*. Supóngase que **P** pretende establecer que *C* está justificado de modo que **P** jugará con el protocolo escéptico. **P** juega *C*, **O** puede jugar *B,* **P** puede únicamente jugar *D* para defender a *C* de *B*, de modo que **P** juega *D,* pero **O** juega *A* así que **P** no cuenta con jugadas que le permitan defender a *C* puesto que no puede repetir ningún argumento jugado por algún jugador. Obviamente lo mismo ocurre si pretende defender a *D*.

Dado que en este ejemplo puede observarse que argumentos máximamente específicos, *C* y *D,* no pueden contar como justificados a pesar de que sus contraargumentos han sido socavados por estos, el proceso de justificación debe ser refinado mediante la introducción de una exigencia adicional además del protocolo escéptico empleado. Esto trae aparejado el siguiente cuestionamiento ¿cuál es la exigencia adicional? y ¿qué justifica su introducción?

Obviamente la restricción debe exigir que cada jugador tenga permitido jugar únicamente argumentos máximamente específicos. La razón que puede justificar tal regla puede consistir en que los jugadores deben jugar siempre sus mejores argumentos, y estos serán los máximamente específicos. De lo contrario, estarán jugando argumentos para los que existe evidencia socavadora, o lo que es lo mismo, reglas aplicadas a casos para los que se cuenta con información excepcional disponible en el conjunto de conocimientos. Esta restricción será expresada como una regla adicional que estipulará que una vez que se socava un argumento se debe finalizar el juego.

(1) El juego *finaliza* si, en cualquier nivel *k*, un jugador *i* mueve un argumento *A* tal que el argumento *B* movido en el nivel *k–1* por el jugador *–i* es tal que *A* es un derrotador por socavamiento de B (*i* gana).

(2) **P** no tiene permitido mover un argumento que ha sido jugado por cualquier jugador en la misma historia.

La regla (1) dice que una vez que un derrotador por socavamiento es jugado, el juego finaliza (el jugador que movió el argumento no máximamente especifico pierde). La intención de esta regla consiste en obligar a los jugadores a usar sólo argumentos MS. La regla (2), asegura un juego escéptico. Se denominará a este protocolo PU.

Por otro lado, una forma obvia de obtener el mismo comportamiento en marcos argumentativos donde no se definan *undermining defeater* pero si derrotas al estilo de los *derrotadores propios* de *DeLP* consistirá en remplazar la primera regla del protocolo PU como sigue:

(1') El juego *finaliza* si, en el nivel *k*, el jugador *i* mueve el argumento *A* tal que el argumento *B* movido en el nivel $k - 1$ por el jugador $-i$ es tal que existe una secuencia $A_1,\ldots, An$ donde $A1 = A$, $A_n = B$ y $(A_h, A_{h+1}) \in def_{prop}$ para cualquier h, $1 \leq h<n$ (*i* gana).

Este protocolo será llamado PP. Entonces, el teorema 5.3.1 claramente asegura el mismo comportamiento bajo el protocolo PU que bajo el protocolo PP. Tal situación es interesante puesto que, aunque *DeLP* no se encuentre dotado de una derrota por socavamiento como la expuesta en la definición 5.3.3 es capaz de obtener los mismos resultados.

Ejemplo 5.3.4 (ejemplo 4.4.3 reconsiderado)

Bajo el protocolo PP o el protocolo PU, el ejemplo de Ana puede modelarse correctamente dado que: Si **P** juega *C*, **O** puede jugar *B*, luego **P** jugando *D* gana el juego ya que introduce un derrotador por socavamiento para *B*, señalando que **O** no ha jugado un argumento máximamente específico. Por otro lado, si *D*(**P**)–*A*(**O**)–*C*(**P**) se obtiene el mismo resultado. De manera que *D* y *C* están justificados. Ahora bien, si se pretende justificar *A*, la historia puede ser expresada de la siguiente manera: *A*(**P**)–*D*(**O**)–B(**P**)–C(**O**) de manera que **O** gana y también podría darse que A(P)–C(O) obteniendo el mismo resultado, *A* no está justificado bajo el protocolo PP. Lo mismo en caso de querer defender *B*.

La propuesta permite modelar correctamente los ejemplos considerados hasta aquí, aunque, debe acompañarse necesariamente con una modificación en el protocolo para modelar casos como en el ejemplo 4.4.3. De lo contrario, argumentos que son considerados máximamente específicos no son justificados. Por otro lado, el sistema no es capaz de dar una solución adecuada al ejemplo de Frank ya que la propuesta sigue asumiendo la derrota basada en conflicto. En suma, aunque el enfoque es bastante general aún resta un aspecto del que S_1 puede dar cuenta correctamente.

Por otro lado es importante señalar que la semántica *básica,* que es escéptica, de un marco argumentativo no pueden expresar este protocolo. Tampoco otra semántica escéptica llamada '*ideal*' (Dung et al., 2007) es capaz de capturar esta noción, ya que la misma es definida en base a la intersección de las extensiones crédulas. De modo que el protocolo PU no tiene una expresión semántica. En (Dimopoulos et al., 2009) fue definida una *semántica estable* a fin de modelar adecuadamente el ejemplo 4.4.3 y al igual que en Alessio (2012). Tal vez, estas semánticas podrían usarse como base para identificar una semántica escéptica que capture el protocolo PU. Habrá que o bien

definir una semántica que lo capture o bien podría pensarse que, en realidad, el enfoque no es capaz de detectar el problema adecuadamente.

5.4 Depuración de argumentos

En la propuesta anterior se brindó un acercamiento a la idea de *máxima especificidad* como criterio que un argumento debe satisfacer. Se advirtió que en los contraejemplos del restablecimiento aparecen argumentos que no verifican la propiedad. Con vistas a implementar la máxima especificidad se recurrió a definir una relación de derrota que intente capturarla. A pesar de haber obtenido resultados prometedores, la misma no parece del todo convincente ya que, aunque soluciona algunos problemas, requiere de mecanismos adicionales para poder impedir otros resultados problemáticos, como el ejemplo de 4.4.3, y aún así, tampoco es capaz de dar cuenta de casos similares al ejemplo de Frank.

Ahora bien, si se pudiese definir un enfoque que fuera capaz de solucionar todas las anomalías consideradas y a la vez dar una respuesta a casos como el ilustrado por el ejemplo de Frank se estaría frente a una alternativa superadora que al mismo tiempo conserva tanto como sea posible a los sistemas en su versión original.

Con vistas a obtener una propuesta general, podría pensarse en la integración de los últimos dos enfoques presentados. A pesar de ello, su aplicación directa llevaría a la introducción de diversas relaciones de derrota entre los mismos argumentos (A derrotaría a B por S_1 y al mismo tiempo por socavamiento) lo cual resultaría extraño. No obstante, las ideas intuitivas de ambos enfoques pueden ser herramientas que permitan la construcción de una propuesta alternativa. Con vistas a identificar tal abordaje se presentará, a continuación, un nuevo análisis de los ejemplos presentados a lo largo del capítulo.

Considérese nuevamente el ejemplo 4.1.2. En tal ejemplo existe la posibilidad de comparar por un lado el argumento "*Al vuela porque es un ave*" con "*Al no vuela por ser una gallina*" y por otro "*Al no vuela por ser una gallina*" con "*Al vuela por ser una gallina salvaje*". Obviamente, entre tales argumentos se ha señalado usualmente, en los sistemas argumentativos, que uno rebate a otro. No obstante, si se consideran, por ejemplo, sólo a la información que permite construir los argumentos "*Al vuela porque es un ave*" y "*Al no vuela por ser una gallina*" puede pensarse que en realidad el primer argumento no es razonable a la luz de la evidencia disponible. En realidad el hecho de saber que Al es una gallina es una razón que invalida la aplicación de la regla "*por lo general las aves vuelan*" al menos al caso de Al. Esto es así porque la información disponible permite saber que Al no es un ave prototípica puesto que si fuese una subclase normal no debería negar una de las propiedades que las aves normales tienen, volar. Esto es claramente consistente con la propuesta representacional de Prakken (2002).

Por otro lado, si ahora se considera a los argumentos "*Al no vuela por ser una gallina*" y "*Al vuela por ser una gallina salvaje*" el anterior análisis puede aplicarse. En realidad, el hecho de que Al sea una gallina salvaje, y que tales gallinas son excepcionales con respecto a las gallinas no voladoras, invalida la aplicación de la regla "*las gallinas no vuelan*".

Ahora bien, si se tienen en cuenta los tres argumentos simultáneamente es claro que el argumento "*Al vuela por ser una ave*" no puede estar habilitado a emplear la regla "*las aves vuelan*" porque se dispone de información que señala que Al es una de esas aves no normales. Pero al mismo tiempo, el argumento para "*Al no vuela porque es una gallina*" tampoco tiene autorizada la aplicación de la regla "*las gallinas no vuelan*" porque esa regla vale para las gallinas normales, y justamente, Al no es una gallina

normal, pero como se ha dicho tampoco es un ave normal, puesto que Al es una excepción de la excepción. Nuevamente, que Al sea una excepción de la excepción no significa que sea un ave normal.

Con respecto al ejemplo de Timoteo las cosas son un poco distintas. El hecho de que Timoteo sea un tero, no constituye evidencia acerca de que sea un ave excepcional. En este caso no tiene sentido bloquear la aplicación de la regla *"por lo general las aves vuelan"* puesto que la aplicación de tal regla es correcta. Por ello, este tipo de argumentos no debe ser invalidado por la propuesta que se construya.

En el ejemplo de Frank, a pesar de que sean argumentos consistentes como en el ejemplo de Timoteo, la cuestión es diferente. El argumento para aceptar que el nido de Frank se encuentra en la isla Florida basado en "*Frank es un canario del cabo"* no es razonable a la luz del hecho: *"Frank es un canario del cabo rojo"*, ya que tales canarios son excepcionales, en relación a su especie (los canarios del cabo), con respecto al lugar donde hacen sus nidos. El hecho de que Frank sea una canario del cabo rojo y atendiendo a que estos canarios pueden tener sus nidos en la isla Florida o la isla Desierta, se debería impedir la aplicación de la regla *"los canarios del cabo tienen sus nidos en la isla Florida"* al menos para el caso de Frank.

Ahora bien, si se supone que existe otra variedad de canarios del cabo, p.e. *los canarios del cabo cuello azul,* que por lo general hacen sus nidos en la isla Florida y dado un canario del cabo cuello azul, llamado Fred, es posible construir dos argumentos consistentes que establecerán que el nido de Fred esté en la isla Florida. Frente a esta nueva situación (similar al ejemplo de Timoteo), el hecho de que Fred sea un canario del cabo cuello azul no inválida la aplicación de la regla *"los canarios del cabo hacen*

sus nidos en la isla Florida", puesto que los canarios del cabo cuello azul son canarios del cabo normales con respecto al lugar donde hacen sus nidos.

Considérese el ejemplo de Beth nuevamente. Dado que Beth no es una empleada normal de Microsoft, el argumento para sostener que Beth es millonaria no parece razonable. Por otro lado, el hecho de que Beth sea una nueva empleada de Microsoft y, que además, trabaje en el departamento X, es una razón que permite inhabilitar la regla "*por lo general los empleadas nuevas de Microsoft poseen menos de medio millón*". Esto es así porque tal información señala que Beth no es una empleada nueva *normal* de Microsoft, al menos en lo que respecta a sus ganancias. Por lo tanto, la regla que habilita la conclusión de tal argumento no debería ser aplicable. Por caso, si se contara con algún argumento que sustentara la conclusión de que Beth, en realidad, no es una empleada no normal (i.e. que Beth no es una empleada nueva o que Beth no es una empleada nueva en el departamento X) el argumento basado en la regla "*por lo general los empleadas de Microsoft son millonarios*" podría autorizar a inferir la conclusión de que Beth es millonaria. Pero no es el caso aquí.

El ejemplo de Ana parece comportarse de igual manera que en los casos anteriores. El hecho de que Ana sea una defensora pública inhabilita la aplicación de la regla *"por lo general los abogados son adinerados"* puesto que los defensores públicos son excepcionales con respecto a ello. De manera que tal argumento no debería contarse como habilitado para sustentar su conclusión. A su vez, el hecho de que rente en Brentwood es una buena razón para negar que Ana sea una residente de Brentwood estándar, ya que en general éstos, los que son adinerados, son propietarios.

Tras este análisis se puede observar que todos los ejemplos problemáticos, sean o no contraejemplos con respecto al restablecimiento, comparten un rasgo común, son

argumentos que sustentan conclusiones basadas en reglas que *no deberían ser aplicables* atendiendo a la evidencia total. Adicionalmente, dado que en los sistemas considerados en el capítulo anterior no cuentan con algún mecanismo que permita detectarlos, tales argumentos impiden la justificación de argumentos que deberían estarlo, como en el ejemplo de Ana, o se permite la justificación de argumentos que no deberían contar como tales, como en los ejemplos de *Al*, *Beth* y *Frank*. Obviamente, la propuesta representacional, al estar dotada de mayor expresividad, si puede hacerlo puesto que cuenta con un predicado de anormalidad o excepcionalidad que evidencia que los argumentos en cuestión lo violan.

Atendiendo a lo planteado anteriormente cabe presentar también aquí una falacia que permitirá comprender la cuestión de una manera interesante. Tal falacia es llamada '*falacia de la exclusión*'.

La falacia de la exclusión ocurre cuando cierta evidencia relevante que puede socavar un argumento inductivo no es considerada en tal argumento. Esta falacia atenta contra el principio de la evidencia total, es decir, toda la información relevante no es tenida en cuenta (Chakraborti, 2006). El siguiente ejemplo, similar al propuesto en Downes (1995), constituye un caso de la falacia de la exclusión. Nótese la similitud con los ejemplos discutidos recientemente.

Ejemplo 5.4.1

A. *Dado que Jones es de Alberta y teniendo en cuenta que la mayoría de los habitantes de Alberta votan al Partido A, probablemente Jones vote al Partido A.*

El argumento del ejemplo 5.4.1 parece razonable, pero existe información relevante que ha sido excluida. La información omitida consiste en que Jones vive en Edmonton, la Capital de Alberta. Obviamente, que Jones viva en Edmonton no parece ser una información importante que invalide el argumento anterior. Sin embargo, el problema aparece a la luz de la siguiente información: *por lo general las personas que viven en Edmonton votan al Partido B o C.*

Obviamente, si se analiza el ejemplo, como es usual en sistemas argumentativos, se dirá que el argumento *A* del ejemplo es derrotado (por rebatimiento) por el siguiente, por estar sustentado en información más específica:

Ejemplo 5.4.2

B. *Jones es de Edmonton, y la mayoría de los que viven en Edmonton votan al Partido B o C, probablemente Jones vote al Partido B o C.*

No obstante, el argumento *A* es un argumento falaz y la prueba de que un argumento es falaz en el sentido de la falacia de la exclusión consiste en señalar la evidencia omitida, en este caso '*Jones es de Edmonton*' y mostrar que tal evidencia cambia la conclusión del argumento '*Jones votará al Partido B o C*'. Debe destacarse que no es suficiente que se muestre que cierta información ha sido excluida, sino que hay que mostrar que la incorporación (o la exclusión) de tal evidencia *cambian* la conclusión del argumento.

Teniendo en cuenta la falacia de la exclusión parece claro que los argumentos problemáticos en todos los ejemplos discutidos adolecen de la misma característica. Son argumentos que excluyen información tal, que si es tenida en cuenta, la conclusión de los argumentos podría cambiar. Esto vale tanto para el caso de Al, como

el de Beth pero también incluye al ejemplo de Ana e incluso al de Frank. Esto justamente se encuentra expresado o subyace a la necesidad de emplear cláusulas de anormalidad que permiten explicitarlo, tal como lo señala Prakken (2002).

Lo anterior permite destacar que las relaciones de derrota, definidas como es usual en los sistemas argumentativos, no logran evitar la obtención de argumentos que no deberían contar como justificados o que interfieren en la justificación de argumentos razonables. Esto es así porque la derrota solamente señala qué argumento es mejor cuando ciertos argumentos no pueden ser aceptados conjuntamente. Adicionalmente, el proceso de justificación supone que todos los argumentos en comparación son *adecuados* i.e. al menos no son falaces.

Si los argumentos construidos son argumentos falaces, no necesariamente se debe acusar a la relación de derrota como demasiado pobre, ni tampoco al restablecimiento, sino al hecho de que no se cuenta con un mecanismo capaz de detectar tales situaciones y establecer la manera de cómo deben ser tratados. La versión representacional de Prakken es, tal vez, una manera de hacerlo.

Por otro lado es interesante, en este contexto, tener en cuenta el principio de máxima especificidad discutido en la sección anterior. Retomando lo dicho, se definió tal principio con vistas a impedir que, entre los argumentos justificados, hayan algunos para los que existe evidencia socavadora, i.e. evidencia que permita la construcción de contraargumentos más específicos, indicando con esto que se tratan de argumentos no máximamente específicos. Como es posible notar, tal criterio permite, en principio, también detectar a los argumentos que son falaces, en el sentido de la falacia de la exclusión, puesto que la falacia consiste en omitir información que socava el argumento en cuestión cuando ésta es considerada.

Ahora bien y atendiendo a la reflexión realizada en párrafos anteriores es importante notar que el principio de máxima especificidad propuesto en la sección 5.3 requerirá de una modificación, al menos para que sea capaz de bloquear la aplicación de ciertas reglas a pesar de que no haya conflicto, como en el ejemplo de Frank, puesto que en tal situación la omisión o explicitación de cierta información lleva a cambiar la conclusión, aunque por cierto, no es un cambio por la conclusión complementaria.

Requerimiento de máxima especificidad: *Un argumento satisface el principio de máxima especificidad cuando no exista al menos un contraargumento más específico* (ejemplos 4.1.2; 4.1.3; 4.4.3) *o cuando no exista un argumento más específico cuya conclusión es implicada por tal argumento pero no viceversa* (ejemplo 5.2.3).

Con vistas a no confundir el requerimiento recién expuesto con el requisito de máxima especificidad definido en la sección 5.3 se empleará el nombre de *'máxima x-especificidad'* para el caso de la presente sección. Ahora bien, si un argumento no satisface este requerimiento, y según lo que se ha visto ya, tal argumento podría ser considerado como un argumento falaz, un argumento que no es razonable a la luz de la evidencia total disponible.

El requerimiento de máxima x-especificidad podría ser implementado de la misma manera que se hizo con el requisito de máxima especificidad, mediante una derrota. Un argumento *A x-socava* a un argumento *B* cuando *A* es estrictamente más especifico que *B* y, o bien la conclusión de *A* y *B* implican contradicción o bien la conclusión de *B* implica la conclusión de *A* pero no viceversa. Pero, tal noción de derrota, aunque modelaría bien los ejemplos 4.1.2 y 4.1.3 como también el ejemplo de Timoteo y de Frank, seguiría permitiendo que ciertos argumentos interfieran en la justificación de argumentos máximamente específicos, como en el caso ilustrado por el ejemplo de

Ana (*ejemplo 4.4.3*). Nuevamente la versión representacional aventajaría a tal propuesta.

De modo que, y hasta aquí, la pretensión de definir un mecanismo formal o criterio que

i. *sea capaz de obtener los resultados adecuados frente a los ejemplos contraintuitivos que genera el restablecimiento,*

ii. *conserve los formalismos tanto como sea posible, y que*

iii. *permita modelar todo lo que la propuesta representacional puede modelar*

parece destinado al fracaso. La versión representacional parece más general dado que no sólo brinda una adecuada respuesta frente a los ejemplos 4.1.2 y 4.1.3 sino que puede resolver las situaciones ilustradas por el ejemplo 4.4.3. Adicionalmente podría extenderse para modelar adecuadamente el ejemplo de Frank.

Ahora bien, si se advierte nuevamente que los argumentos que no satisfacen el requerimiento de máxima x-especificidad pueden considerarse como argumentos falaces, tales argumentos podrían excluirse del proceso de justificación mediante una fase de preselección de los argumentos. Obviamente las razones que motivan tal abordaje son bastante claras. Sólo aquellos argumentos adecuados, i.e. argumentos que al menos no sean falaces, pueden considerarse como buenos candidatos para elegir, entre ellos, cuál o cuáles se encuentran justificados.

La idea enunciada evitaría que los argumentos que no verifiquen la propiedad de *máxima x-especificidad* interfieran en la justificación de los argumentos. Este proceso de filtrado permitirá que, a la larga, los sistemas cuenten solamente con argumentos máximamente x-específicos como justificados, es decir con argumentos para los que no existe información que invalide la aplicación de ciertas reglas de inferencia, o lo

que es lo mismo, con argumentos que se sustentan en la más restringida clase de referencia teniendo en cuenta la información total disponible. Por otro lado serán excluidos aquellos argumentos que son inaceptables a la luz de la información total disponible, argumentos que sustentan su conclusión en base a cierta información que omite evidencia tal que si fuese tenida en cuenta el argumento no sería razonable.

La idea general para el nuevo enfoque consistirá en definir una fase adicional para un sistema argumentativo donde se realizará una depuración de aquellos argumentos inadecuados mediante su exclusión. Esto permitirá que ciertos argumentos no se entrometan en el proceso de justificación.

Esquemáticamente un sistema argumentativo contará con una fase de *construcción de argumentos* (*fase 1*), una fase de *depuración de argumentos* (*fase 2*), una fase de *comparación de argumentos a favor o en contra de ciertas conclusiones* (*fase 3*) y una fase de *selección de argumentos justificados* (*fase 4*). El criterio de depuración estará basado en la noción de argumento máximamente x-específico. Con vistas a brindar una definición formal de la idea, será implementada en *DeLP,* aunque claramente la idea general puede ser fácilmente adaptada a otros formalismos como se sugerirá más adelante.

Definición 5.4.1
(x-especificidad)
Sea $\boldsymbol{\mathcal{P}}$ = (Π, Δ) un *delp* donde Π_G es el conjunto de todas las reglas estrictas (sin incluir hechos). Sea $\boldsymbol{\mathcal{F}}$ el conjunto de todos los literales para los que existe una derivación rebatible a partir de $\boldsymbol{\mathcal{P}}$ ($\boldsymbol{\mathcal{F}}$ será considerado un conjunto de hechos). Sean $\langle T_1, h_1\rangle$ y $\langle T_2, h_2\rangle$ dos argumentos obtenidas a partir de $\boldsymbol{\mathcal{P}}$. Se dirá que $\langle T_1, h_1\rangle$ es *estrictamente más x-específico* que $\langle T_2,h_2\rangle$, notado como $\langle T_1, h_1\rangle \succ_{\text{x-esp}} \langle T_2, h_2\rangle$, si y sólo si:

i. Para todo $H \subseteq \mathcal{F}$: Si $\Pi_G \cup H \cup T_1 \vdash h_1$ y $\Pi_G \cup H \nvdash h_1$ entonces $\Pi_G \cup H \cup T_2 \vdash h_2$, y

ii. Existe al menos un $H' \subseteq \mathcal{F}$ tal que $\Pi_G \cup H' \cup T_2 \vdash h_2$ y $\Pi_G \cup H' \nvdash h_2$ y $\Pi_G \cup H' \cup T_1 \nvdash h1$, y

iii. o bien $\Pi_G \cup \{h_1, h_2\} \vdash \bot$ o bien $\Pi_G \cup \{h_2\} \vdash h_1$ y $\Pi_G \cup \{h_1\} \nvdash h_2$.

□

i. dice que para todo conjunto *H*, si *H* permite derivar h_1 a partir de $\Pi_G \cup T_1$, donde al menos una regla de es utilizada puesto que no existe una derivación estricta para h_1 a partir de $\Pi_G \cup H$, entonces ese mismo conjunto H permite obtener una derivación rebatible para h_2 a partir de $\Pi_G \cup T_2$. La condición *ii.* establece que existe al menos un conjunto *H'* que permite que haya una derivación de h_2 pero no de h_1. *iii*. establece que la comparación ha de hacerse entre argumentos tales que sus conclusiones sean contradictorias o la del más general implique la conclusión del más específico.

Si un argumento es más x-específico que otro, ello no significa que el argumento en cuestión sea máximamente x-específico, la siguiente definición puntualiza la exigencia.

Definición 5.4.2
(máxima x-especificidad)
Una argumento $\langle T_1, h_1 \rangle$ se dice *máximamente x-específico* con respecto a h_1 si y sólo si no existe ningún argumento $\langle T_2, h_2 \rangle$ tal que $\langle T_2, h_2 \rangle$ sea estrictamente más x-específico que $\langle T_1, h_1 \rangle$.

□

Bajo estas nociones es posible volver a considerar nuevamente los ejemplos contraintuitivos con respecto al restablecimiento: El ejemplo de Al (4.1.2) y el ejemplo

de Beth (4.1.3). Como también aquellos que permiten valorar la generalidad de la propuesta: los ejemplos 4.4.3 (el ejemplo de Ana), 5.2.3 (*el ejemplo de Frank*) y 5.2.4 (*el ejemplo de Timoteo*).

En el ejemplo 4.1.2 el único argumento máximamente x-específico es *C*, por su parte en el ejemplo 4.1.3 sólo *C* es máximamente x-específico. En el ejemplo 4.4.3, únicamente lo son *C* y *D*. En el ejemplo 5.2.3 sólo *B*, mientras que en el ejemplo 5.2.4, *A* y *B* son máximamente x-específicos. Todos coinciden con lo esperable intuitivamente. Por otro lado, los argumentos restantes no son máximamente x-específicos con respecto a la conclusión que sustentan.

Es importante notar que la propiedad de argumento *máximamente x-específico* es derrotable. La incorporación de nueva información puede permitir la construcción de un argumento que retracte tal estado, i.e. construir un argumento estrictamente más x-específico que el argumento en cuestión.

Resta establecer la manera de implementar el requerimiento de máxima x-especificidad. Como se dijo anteriormente, la propuesta consistirá en incorporar una fase adicional de depuración de los argumentos que no son máximamente x-específicos. Vale recordar que en general los sistemas argumentativos se caracterizan por un proceso constituido de las siguientes fases o etapas:

***Fase 1*:** Construcción de argumentos

***Fase 2*:** Definir el marco argumentativo, i.e. establecer las relaciones de derrota entre argumentos.

Fase 3: seleccionar los argumentos justificados.

La fase adicional se incorporará antes de definir el marco argumentativo. Esquemáticamente:

***Fase 1**:* *Construcción de los argumentos*.

***Fase 2**:* *Selección de los argumentos máximamente x-específicos*.

***Fase 3**:* *Marco argumentativo con los argumentos seleccionados en el punto anterior*.

***Fase 4**:* *Selección de los argumentos justificados*.

Obviamente la selección de los argumentos máximamente x-específicos y la depuración de aquellos que no lo son, debe hacerse antes del proceso de comparación puesto de lo contrario, tales argumentos interferirán en el proceso de justificación. De modo que en la fase 2 se obtendrá el conjunto de los argumentos máximamente x-específicos de la siguiente manera:

Definición 5.4.3
(conjunto de argumentos máximamente x-específicos)
Dado el conjunto *Args* de todas las estructuras argumentativas que pueden construirse en base a un programa $\mathcal{P}$, se define: $Args^* =_{df} \{\langle T, h\rangle \in Args : \langle T, h\rangle$ es máximamente x-específico c.r. a $h\}$.

□

El marco argumentativo, la *fase 3*, por tanto será definido en base al conjunto *Args**.

Definición 5.4.4
(marco argumentativo)
Un *marco argumentativo* asociado *a* $\mathcal{P}$ es un par

$$MA_{\mathcal{P}} = \langle Args^*, derrota\rangle$$

donde $Args^*=_{df}\{\langle T,h\rangle \in Args$: A es máximamente x-específico c.r. a $h\}$ y *derrota* es una relación entre los miembros de $Args^*$.

□

La fase 4 podrá consistir simplemente en la implementación de un juego escéptico como el definido en la definición 5.3.5 bajo el protocolo escéptico, i.e. **P** no puede jugar un argumento ya jugado por algún jugador. Aunque no habría inconvenientes de hacerlo mediante el proceso de justificación definido para *DeLP*.

Hasta aquí parece que se ha obtenido un criterio lo suficientemente general como para resolver los problemas considerados. Sin embargo, la definición de x-especificidad parece que debería ser modificada en un detalle adicional. Si se considera el siguiente ejemplo abstracto basado en *DeLP* se evidenciará el inconveniente:

Ejemplo 5.4.3

$\Pi_F = \{e_1, e_2, e_3\}$

$\langle A_1,\neg q\rangle$: $\langle\{r \prec e_3;\ \neg q \prec e_2 \wedge r\ \},\ \neg q\rangle$

$\langle A_2,\neg p\rangle$: $\langle\{q \prec e_2;\ \neg p \prec e_1 \wedge q\ \},\ \neg p\rangle$

$\langle A_3, h\rangle$: $\langle\{p \prec e_1;\ h \prec p\ \},\ h\rangle$

Según la definición de máxima x-especificidad, $\langle A_1,\neg q\rangle$, $\langle A_2,\neg p\rangle$, $\langle A_3,h\rangle$ son argumentos máximamente x-específicos. No obstante, un subargumento propio de $\langle A_2,\neg p\rangle$ y un subargumento propio de $\langle A_3,\ h\rangle$ no son máximamente x-específicos. Sería deseable que un argumento $\langle A_1,h_1\rangle$ cuente como máximamente x-específico cuando no exista al menos un argumento $\langle A_2,h_2\rangle$ que sea más x-específico que cualquier subargumento de $\langle A_1,h_1\rangle$. Esto podría ser considerado como un principio de composicionalidad de la

máxima x-especificidad. Tal idea lleva a la siguiente revisión de la noción de máxima x-especificidad.

Definición 5.4.5
(máxima x-especificidad revisada)
Una argumento $\langle T_1,h_1\rangle$ se dice *máximamente x-específico* con respecto a h_1 si y sólo si no existe un argumento $\langle T_2,h_2\rangle$ tal que $\langle T_2,h_2\rangle$ sea estrictamente más específico que $\langle T, h\rangle$ para cualquier $\langle T, h\rangle$ tal que $\langle T, h\rangle$ es un subargumento de $\langle T_1,h_1\rangle$.

□

Esta nueva versión, finalmente, satisface todos los ejemplos, incluyendo el último propuesto (5.4.3).

Es importante recordar que ya sea en *DeLP,* como en *PS* o *KT,* existen otros tipos de derrotas posibles además de la basada en especificidad. Por ello, en la construcción del marco argumentativo, podrán existir derrotas asimétricas independientemente de que las derrotas estrictas basadas en especificidad sean inexistentes, puesto que el proceso de preselección elimina tales relaciones entre los argumentos. De hecho, para cualquier par de argumentos *A* y *B* que pertenezcan a *Args*,* si *A* derrota a *B, A* no es más específico que *B*. Esto es así porque si *A* fuese más específico que *B*, *B* no sería máximamente x-específico, es decir, *B* no pertenece a *Args**, lo cual es contradice el supuesto.

Por otro lado, y teniendo en cuenta el principio de composicionalidad de la máxima x-especificidad, la propuesta permite también obtener resultados interesantes frente a la ocurrencia de ciclos o derrota reciprocas cuando la especificidad y la derrota a premisas se encuentran involucradas.

Los ciclos pueden ser ilustrados con el siguiente ejemplo definido en el lenguaje de *MTDR.*

Ejemplo 5.4.4

$\langle A_1, \neg q\rangle$: $\langle \{e_3 \succ\!\!- r;\ e_2 \wedge r \succ\!\!- \neg q\}, \neg q\rangle$

$\langle A_2, \neg p\rangle$: $\langle \{e_2 \succ\!\!- q;\ e_1 \wedge q \succ\!\!- \neg p\}, \neg p\rangle$

$\langle A_3, \neg r\rangle$: $\langle \{e_1 \succ\!\!- p;\ e_3 \wedge p \succ\!\!- \neg r\ \}, \neg r\rangle$

Donde $\Pi_F = \{e_1, e_2, e_3\}$. En *MTDR,* $\langle A_1, \neg q\rangle$ es un derrotador propio de $\langle A_2, \neg p\rangle$ y $\langle A_2, \neg p\rangle$ es un derrotador propio de $\langle A_3, \neg r\ \rangle$ mientras que $\langle A_3, \neg r\rangle$ es un derrotador propio de $\langle A_1, \neg q\rangle$. La figura 5.4.1 ilustra la situación.

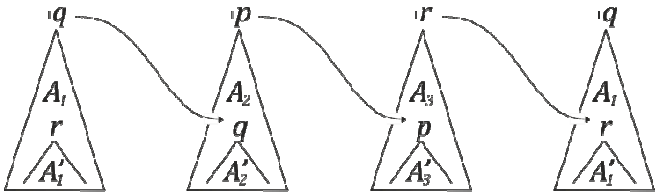


Fig. 5.4. 1: Ciclo de derrotas impar

Frente a esta situación es posible advertir que los argumentos involucrados en el ciclo, no verifican la propiedad de ser máximamente x-específicos. Por lo tanto, ninguno de tales argumentos pertenecerá a *Args*.* Lo mismo sucede frente a las derrotas recíprocas o derrotas cruzadas (Prakken, 1993).

Ejemplo 5.4.5

$\langle A_1, \neg p\rangle$: $\langle \{e_1 \succ\!\!- r;\ e_2 \wedge r \succ\!\!- \neg p\}, \neg p\rangle$

$\langle A_2, \neg r\rangle$: $\langle \{e_2 \succ\!\!- p;\ e_1 \wedge p \succ\!\!- \neg r\ \}, \neg r\rangle$

En el ejemplo 5.4.5, el argumento $\langle A_1, \neg p \rangle$ no es máximamente x-específico con respecto a la conclusión $\neg p$ y lo mismo puede verificarse para $\langle A_2, \neg r \rangle$, de manera que $\langle A_1, \neg p \rangle \notin Args^*$ y $\langle A_2, \neg r \rangle \notin Args^*$.

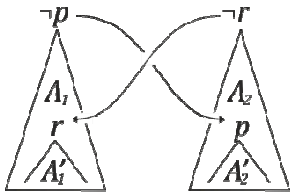


Fig. 5.4. 2: Derrotas recíprocas

Además de ciclos impares de derrotas, también los hay pares, como se ilustra en la figura 5.4.3. En tal figura se supone que el argumento A_1 es un contraargumento más específico del argumento A_2' (subargumento de A_2). El argumento A_2 es un contraargumento más específico para A_3' (subargumento de A_3) y A_3 de A_4' (subargumento de A_4). A su vez, A_4 es un contraargumento más específico de A_1'. En este caso, también, todos los argumentos involucrados en el ciclo son descartados en el proceso de selección de argumentos puesto que ninguno es máximamente x-específico.

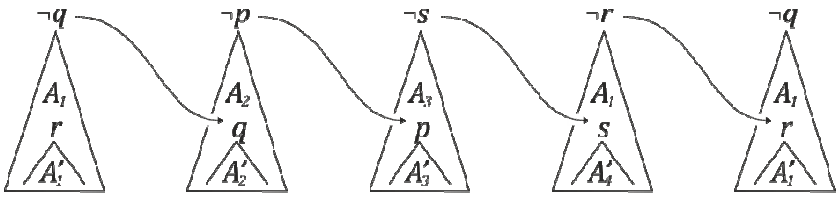


Fig. 5.4. 3: Ciclo de derrotas par

Lo anterior permite evidenciar que la máxima x-especificidad brinda una solución alternativa para el tratamiento de ciclos de derrotas o derrotas recíprocas.

El proceso de preselección de argumentos puede aplicarse a diferentes sistemas. En caso de que se esté interesado en implementar la propuesta, por ejemplo, en el sistema *PS* se debe definir una relación de preferencia estricta para argumentos cuyas reglas se relacionan por especificidad. En las definiciones 5.4.6 a 5.4.8 se propone una posible manera de expresar la máxima x-especificidad para el sistema *PS*. Obviamente, la misma merece un estudio aparte, pero es útil para ilustrar la idea general.

Definición 5.4.6
(argumento preferido por x-especificidad)

Sea T una teoría constituida por dos conjuntos de reglas D y S, donde S es un conjunto de reglas estrictas y D es un conjunto de reglas rebatibles. Sea (T, <) una teoría ordenada, si $r<r'$ entonces la regla r' es preferida a r. Sean A_1 y A_2 dos argumentos y sea $>_{esp}$ una preferencia por especificidad dada en T. Se dirá que *A_1 es estrictamente preferido por x-especificidad a A_2* si y sólo si:

i. Existe al menos una regla $r_a \in A_1$ y una regla $r_b \in A_2$ tal que $r_a >_{esp} r_b$, y

ii. No existe ninguna $r'_b \in A_2$ y $r'_a \in A_1$ tales que $r'_b >_{esp} r'_a$, y

iii. o bien $S \cup \{Con(r_a), Con(r_b)\} \vdash \bot$ o bien $S \cup \{Con(r_b)\} \vdash r_a$ y $S \cup \{Con(r_a)\} \nvdash r_b$

□

Contando con la preferencia por x-especificidad será muy fácil definir la idea de argumento máximamente x-específico para el sistema *PS*.

Definición 5.4.7
(argumento máximamente x-específico)

Un argumento A_1 será *máximamente x-específico* con respecto a su conclusión si y sólo si no existe ningún argumento A_2 que sea estrictamente preferido por x-especificidad a A_1.

□

A partir de tal definición podrá construirse un conjunto de argumentos refinado por la satisfacción de la definición 5.4.7 y entre ellos definir las relaciones de derrota restantes.

Definición 5.4.8

(conjunto de argumentos máximamente x-específicos)

El conjunto *Args* de argumentos máximamente x-específicos* es definido como $Args^* =_{df}$ $\{A_i \in Args : A_i$ es máximamente x-específico con respecto a la conclusión que sustenta$\}$.

□

Una vez que el conjunto *Args** se encuentra disponible, el sistema *PS* podrá obtener el conjunto de argumentos justificados según la siguiente definición

Definición 3.2.8

(conjunto $JustArgs^*_\Gamma$)

Sea Γ una teoría ordenada. Entonces se define la siguiente secuencia de subconjuntos de $Args^*_\Gamma$:

- $F_\Gamma^0 = \varnothing$
- $F_\Gamma^{i+1} = \{A \in Args^*_\Gamma \mid$ A es aceptable c.r. a $F_\Gamma^i\}$

Entonces el conjunto $JustArgs^*_\Gamma$ en base de Γ es $\bigcup_{i=0}^{\infty}(F_\Gamma^i)$.

□

Algo similar a lo anterior podría equipar la propuesta de Kowalski y Toni (1996) a fin de que ciertos conjuntos de hipótesis se vean descartados, en una fase previa al proceso de comparación.

Finalmente, después de una serie de versiones preliminares y tras haber ajustado la propuesta final, es momento de considerar si constituye una solución *ad hoc*. Obviamente, dado que la propuesta ha sido definida en base a los problemas detectados la respuesta parece afirmativa. Sin embargo es bastante general, en primer lugar porque el enfoque no sólo resuelve el problema detectado frente a los ejemplos 4.1.2 y 4.1.3 sino que además puede dar cuenta de otros casos como 4.4.3 y 5.2.3. Adicionalmente, la propuesta no derrota argumentos innecesariamente (ejemplo 5.2.4). A su vez, brinda una solución alternativa para el tratamiento de ciclos de derrotas o derrotas recíprocas.

Una tarea ha quedado pendiente, implementar la estrategia de manera adecuada para el sistema *PS* (Prakken & Sartor, 1996[b]) y *KT* (Kowalski y Toni, 1996). Una posible vía de avance se ha bosquejado. Resta evaluarla en el contexto general de cada sistema. No obstante parecería deseable más bien definir un mecanismo en términos abstractos que permita la instanciación del mismo en diversos sistemas. Esto podría hacerse en investigaciones futuras. Por otro lado, también sería interesante valorar la implementación en un marco más general de *DeLP*, en particular estudiar cómo afecta el proceso de preselección en *PreDeLP*, dado que en tal sistema se incorpora el empleo de presunciones (Martínez et al., 2012).

5.5 Conclusión

El capítulo partió planteando la búsqueda de una respuesta alternativa a la sugerida en Horty (2001) y Prakken (2002). La primera, expuesta en el capítulo anterior, se

basaba en que el restablecimiento es una idea errónea y debe evitarse. Esta posición fue ilustrada en el capítulo precedente mediante la presentación de ejemplos y la instanciación de los mismos en tres sistemas argumentativos (*PS*, *KT* y *SL*). Se observó cómo los resultados intuitivamente incorrectos eran obtenidos. La segunda, responsabiliza al lenguaje formal y su incapacidad para evitar la aparición de los casos problemáticos.

De manera que el objetivo del capítulo fue identificar un criterio, independiente del lenguaje de representación, que permita la minimización de los resultados anómalos, suponiendo la validez del restablecimiento. Para ello, se analizaron los ejemplos propuestos en Horty (2001). Con vistas a valorar tal criterio, fue útil estudiar la estrategia representacional formulada por Prakken (2002), presentada al finalizar el capítulo IV. Como resultado se obtuvieron tres vías de escape posibles.

La primera vía de escape consistió en una estrategia basada en la introducción de una relación de derrota entre argumentos consistentes, cuando éstos se relacionan por especificidad y la conclusión del argumento más general implica la conclusión del más específico. Se detectó que tal propuesta era poco tolerante puesto que, aunque puede modelar correctamente los ejemplos 4.1.2 (*el ejemplo de Al*) y 4.1.3 (*el ejemplo de Beth*), derrota ciertos argumentos sin necesidad (*el ejemplo 5.2.4, de Timoteo*). Pensando una estrategia de solución inmediata a tal situación se advirtió que la misma puede evitar las derrotas innecesarias pero no puede modelar adecuadamente el ejemplo 4.1.2. Adicionalmente, se constató que tampoco es capaz de dar cuenta de casos ilustrados por el ejemplo 4.3.3 (*el ejemplo de Ana*), ejemplo correctamente modelado en la versión representacional de Prakken. Sin embargo, un resultado interesante fue obtenido, la modelación correcta del ejemplo 5.3.3 (*el ejemplo de*

Frank). Como resultado final se puede concluir que el enfoque no puede proponerse como una estrategia general y que la versión representacional claramente lo supera.

Posteriormente se brindó un acercamiento alternativo basado en las nociones de *evidencia socavadora* y *máxima especificidad*. La propuesta permite dar cuenta de los ejemplos 4.1.2 y 4.1.3, evita la aparición de derrotas innecesarias, y con una modificación en el proceso de justificación, puede modelar adecuadamente el ejemplo 4.3.3. Sin embargo, el resultado obtenido en la propuesta anterior frente al ejemplo 5.2.3, no puede obtenerse.

Finalmente se propuso una estrategia de preselección de argumentos o de depuración mediante un refinamiento de la noción de máxima especificidad sugerida por la propuesta anterior. En esta alternativa se propuso la incorporación de una fase adicional en el proceso argumentativo. La fase adicional consiste, simplemente, en la selección de *argumentos máximamente x-específicos* y el descarte de aquellos que no satisfacen el requerimiento. Cuando tal conjunto ha sido determinado, se construye el marco argumentativo correspondiente y luego, mediante algún proceso de selección (en este caso se propuso un juego argumentativo escéptico) se fija el conjunto de argumentos justificados. Como resultado, se pueden modelar correctamente a los ejemplos 4.1.2 y 4.1.3, se evitan derrotas innecesarias (ejemplo 5.2.4) y se obtienen los resultados esperados frente a los ejemplos 4.3.3 y 5.2.3.

Si bien las tres estrategias pueden ser valoradas en función de cuán generales son, todas ellas cuentan con un recurso formal que no apela al incremento del lenguaje formal pero que puede solucionar no sólo los contraejemplos al restablecimiento, sino que además permite resolver otros casos problemáticos. Obviamente, la última propuesta parece más simple de implementar y a su vez, es más general que las

anteriores. Finalmente resta en considerar que una tarea ha quedado pendiente, definir la estrategia de manera adecuada para que sea implementada en los sistemas *KT* y *PS*, aunque una dirección de cómo hacerlo se ha indicado. Adicionalmente se ha señalado que la propuesta merece un estudio en el contexto de *PreDeLP*, dado que en tal sistema se incorpora el empleo de *presunciones* y la noción de máxima x-especificidad supone la apelación a *hechos*.

Capítulo VI: Conclusión

Los sistemas argumentativos son modelos formales que pretenden representar información tentativa y potencialmente contradictoria mediante la construcción y comparación de argumentos a favor o en contra de ciertas conclusiones. Estos sistemas pueden ser caracterizados en base a rasgos y etapas comunes. En el capítulo II esto fue realizado en respuesta a tres interrogantes: *Cómo se construyen los argumentos, cómo se relacionan y cómo se determina el estado final de los mismos.* En el capítulo III fueron expuestos diversos sistemas argumentativos específicos que instancian las características comunes de los sistemas argumentativos en formalismos concretos.

En el capítulo IV se mostraron ejemplos que ponen en duda una de los principios ampliamente aceptados en argumentación rebatible: *el restablecimiento.* A su vez, tales ejemplos fueron modelados en sistemas particulares y se evidenció que tales sistemas no son capaces de dar los resultados esperados.

Diferentes respuestas fueron dadas en la literatura frente a la problemática identificada. La primera respuesta, simple pero drástica, fue dada por Horty (2001). Dado que el restablecimiento permite la obtención de resultados problemáticos (*como en el ejemplo de Al*) y resultados incorrectos (*como en el ejemplo de Beth*) es necesario abandonarlo como un principio general. La idea ya había sido discutida en el terreno de las redes de herencia (Touretzky et al., 1991) y se había obtenido la misma sugerencia. Por ello, Loui y Stiefvater (1992) abordaron el problema frente a lo sugerido por Touretzky y otros. Ellos sostienen que para impedir el restablecimiento

incorrecto, simplemente se debe representar de otra manera la información. Prakken (2002) crítica la sugerencia de Horty e indica que el problema no es debido al restablecimiento en sí, sino a la incapacidad del lenguaje para bloquear ciertos defaults cuando esto es necesario. Por ello sostiene que, en realidad, los ejemplos propuestos por Horty no alcanzan para invalidar el restablecimiento. Horty (2012), cambiando su postura, también adhiere a que la cuestión radica no en el restablecimiento en sí sino en cómo se formalizan los casos considerados. Según él, una vez que los ejemplos se encuentran debidamente representados, el restablecimiento puede ser visto como inocuo.

Un ensayo representacional propuesto por Prakken (2002) fue expuesto al final del capítulo IV. La estrategia permite resolver los problemas ocasionados por los contraejemplos del restablecimiento y además abordar adecuadamente otros. De modo que hay varias razones que justifican este enfoque. Sin embargo, es importante destacar una de las advertencias que hace Prakken en el mismo trabajo: el desafío consiste en identificar principios que permitan elegir una correcta representación. Parece claro que la tarea es una empresa ardua. Por ello, en la presente tesis se pretendió como objetivo estudiar la posibilidad de detectar algún criterio formal, independiente del lenguaje, que pudiera neutralizar los casos problemáticos.

Diversas propuestas se han realizado en el capítulo V. Con vistas a valorarlas, fue instructivo tener siempre presente los resultados obtenidos por el enfoque representacional de Prakken puesto que si un enfoque alternativo es definido, sería deseable que al menos permita resolver todo lo que tal enfoque es capaz de hacer.

La primera estrategia consistió en introducir una derrota adicional, denominada S_1. Esta idea estaba basada en la intuición de que el problema aparece cuando

restablecedor y restablecido se relacionan en dos aspectos: Especificidad e implicación de las conclusiones. De modo que se procedió a definir una relación de derrota entre argumentos que se relacionan por ambos aspectos. Tal abordaje es capaz de dar cuenta de los resultados extraños señalados pero no es lo suficientemente general como para que constituya una propuesta definitivamente convincente.

La siguiente vía de escape, también se asentó en la introducción de una derrota adicional, llamada *derrota por socavamiento*. La justificación de la introducción de tal derrota se encuentra basada en el hecho de que aquellos argumentos que se encuentran justificados cuando no deberían estarlo, al menos desde el punto de vista intuitivo, son argumentos que no verifican la propiedad de ser máximamente específicos. Un argumento es considerado como máximamente específico cuando no existe ningún argumento rival para tal argumento que sea más específico. La propuesta resultante, al igual que S_1, permite obtener resultados interesantes pero se advirtió que no es lo suficientemente general, al menos como para proponerse como alternativa al enfoque representacional.

Finalmente mediante un refinamiento de las intuiciones subyacentes a las anteriores propuestas se sugirió un criterio de preselección de los argumentos previa satisfacción de una propiedad: *la máxima x-especificidad.* La idea general consiste en definir una fase adicional para un sistema argumentativo donde se realiza una depuración de aquellos argumentos inadecuados mediante su exclusión, i.e. aquellos argumentos que no son máximamente x-específicos. Esto evita que ciertos argumentos se entrometan en el proceso de justificación. Esquemáticamente la propuesta establece que un sistema argumentativo contará con una fase de *construcción de argumentos*, una fase de *depuración de argumentos*, una fase de

comparación de argumentos a favor o en contra de ciertas conclusiones y una fase de *justificación de argumentos.*

Tal como se ha dicho, el criterio de depuración está basado en la noción de *argumento máximamente x-específico* capturado en la definición 5.4.4. Tal definición sostiene que un argumento *A* será considerado máximamente x-específico cuando no exista ningún argumento *B* tal que *B* sea más x-específico que *A* y, o bien *B* sea un contraargumento de *A* o bien la conclusión de *A* implique la conclusión de *B* pero no viceversa.

Una vez que los argumentos máximamente x-específicos han sido detectados se procede a su recolección en un conjunto de argumentos, notado como *Args**, tal que *Args**$\subseteq$ *Args.* El conjunto *Args** está constituido únicamente por aquellos argumentos que satisfacen la propiedad de máxima x-especificidad. Los restantes argumentos *Args–Args** son simplemente descartados. Entre los miembros de *Args** se definen las relaciones de derrota y se procede a seleccionar los argumentos justificados. Es decir, el marco argumentativo es definido en base a *Args** y no en base a *Args.*

Como resultado, la propuesta permite modelar correctamente a los ejemplos propuestos por (Horty, 2001) que ponen en duda el restablecimiento. Adicionalmente el enfoque es capaz de obtener los resultados esperados frente a otro ejemplo discutido también en (Horty, 2001 y Prakken, 2002) donde, según el enfoque propuesto, los argumentos que no son máximamente x-específicos interfieren en la justificación de argumentos máximamente x-específicos (*El ejemplo de Ana*). A su vez, permite resolver un ejemplo propuesto en (Horty, 2012) donde se obtienen resultados no intuitivos entre argumentos consistentes (*El ejemplo de Frank*).

Trabajos futuros

La propuesta final, consistente en la depuración de argumentos, es interesante porque permite obtener los resultados esperados y a su vez resuelve otros inconvenientes adicionales dando razones a favor de su generalidad. Pero, dado que fue instanciada únicamente en *DeLP* resta poder o bien expresar el criterio en un marco abstracto que permita la instanciación del criterio en diversos sistemas que modelan razonamiento default, tales como (Prakken & Sartor, 1996[b]; Kowalski & Ton, 1996, García & Simari, 2004; Dung & Son, 2000; Martínez et al., 2012), o bien poder definir diversas maneras específicas de expresar el criterio de *máxima x-especificidad* en cada sistema.

Esta última estrategia tiene como punto fuerte que el hecho de definir concretamente la máxima x-especificidad como criterio de preselección permite atender a aspectos específicos de los sistemas a fin de prever la obtención de resultados inadecuados. Sin embargo, frente a cada sistema particular será necesario un estudio de la implementación de la estrategia. Por esto, la búsqueda de la definición del principio en el contexto de un sistema abstracto parece tratarse de una estrategia mejor. Para ello será importante poder seleccionar o definir un sistema de un nivel de relativa abstracción donde los argumentos cuenten con alguna estructura que permita capturar la noción de especificidad basada en conflicto e implicación, lo que podría hacerse en el contexto de los sistemas propuestos en (Baláž, M. & Frtúz, J., 2012) o (Martínez, D. C., García, A. J., & Simari, G. R. 2006; 2008) o (Prakken, 2010).

Bibliografía

Alessio, C.A. (2013). Un sistema argumentativo para razonamiento default: discusión y propuesta. En Algañaraz V. et al (comp.) *II Encuentro de Jóvenes Investigadores: consolidando espacios del quehacer científico en San Juan*.

Alessio, C.A. (2012). Una propuesta para modelar razonamiento default en sistemas argumentativos. En Milone, R. A. y Torres, J.M. *Ciencia y Metafísica: Selección de Trabajos*, Ed. Facultad de Filosofía y Letras - UNCuyo.

Alessio, C.A. (2012). Aceptabilidad extendida para marcos argumentativos abstractos. En Salvatico, L., Bozzoli, M. y Pesenti, L. (Eds): *Epistemología e Historia de la Ciencia: Selección de trabajos de las XXII Jornadas*.

Alessio, C.A. (2010). Extensiones no justificadas en Lógica Default. En García, P. & Massolo, A. (Comp.) *Epistemología e historia de la ciencia: selección de trabajos de las XX Jornadas*, (pp. 13-19).

Alessio, C.A. (2009). Razonamiento rebatible: origen, historia, estrategias. *Revista Cuadernos de la Universidad*, *42*: 15-35.

Amgoud, L., Bodenstaff, L., Caminada, M., McBurney, P., Parsons, S., Prakken, H., van Veenen, J. & Vreeswijk, G.A.W. (2006). Final review and report on formal argumentation system. Deliverable D2.6, ASPIC IST-FP6-002307

Amgoud, L., & Cayrol, C. (1997). Integrating preference orderings into argument-based reasoning. In *Qualitative and Quantitative Practical Reasoning* (pp. 159-170). Springer Berlin Heidelberg.

Amgoud, L., & Cayrol, C. (1998, July). On the acceptability of arguments in preference-based argumentation. In *Proceedings of the Fourteenth conference on Uncertainty in artificial intelligence* (pp. 1-7). Morgan Kaufmann Publishers Inc.

Amgoud, L. & Cayrol, C. (2002). A model of reasoning based on the production of acceptable argument. *Annals of Mathematics and Artificial Intelligence, 34* (1-3), 197–215.

Amgoud, L. Cayrol, C., Lagasquie-Schiex, M.C. (2004). On the bipolarity in argumentation frameworks. En Didier Dubois, Christopher A. Welty, Mary-Anne Williams (Eds.): *Principles of Knowledge Representation and Reasoning: Proceedings of the Ninth International Conference (KR2004*) (pp 1–9). AAAI Press.

Amgoud, L., Cayrol, C., Lagasquie-Schiex, M. C., & Livet, P. (2008). On bipolarity in argumentation frameworks. *International Journal of Intelligent Systems*, *23*(10), 1062-1093.

Amgoud, L., Maudet, N. & Parsons, S. (2000). Modelling dialogues using argumentation. *In Proceedings of the Fourth International Conference on MultiAgent Systems* (pp. 31–38). IEEE. Boston, MA.

Annis, D. (1973). Knowledge and Defeasibility. *Philosophical Studies.* 24:199-203

Augusto, J.C. (1998). *Razonamiento Rebatible Temporal*. Tesis Doctoral. Universidad Nacional de Sur, Bahía Blanca, Argentina.

Baader, F. & Hollunder, B. (1993). How to prefer more specific defaults in terminological default logic. *Technical Report RR-92-58,* DFKI

Baláž, M. & Frtúz, J. (2012). Abstract Argumentation with Structured Arguments. *In Proc. NMR12.*

Barker, J. (1976). What You Don't Know Won't Hurt You. *American Philosophical Quarterly*, 13: 303-308.

Baroni, P., Caminada, M. & Giacomin, M. (2011). An introduction to argumentation semantics. *The Knowledge Engineering Review, 26*(04): 365-410.

Baroni, P., & Giacomin, M. (2009). Semantics of abstract argument systems. En Simari, G. & Rahwan, I. (Eds.) *Argumentation in Artificial Intelligence* (pp. 25-44). Springer US.

Baroni, P., Giacomin, M. & Guida, G. (2005). Scc-recursiveness: a general schema for argumentation semantics. *Artificial Intelligence*, *168* (1-2): 162–210.

Barker, J. (1976). What You Don't Know Won't Hurt You. *American Philosophical Quarterly, 13*: 303-308.

Bench-Capon, T.M. (2003). Persuasion in practical argument using value-based argumentation frameworks. *Journal of Logic and Computation, 13*(3):429–448.

Bench-Capon, T. J. M., Atkinson, K. & Chorley, A. (2005). Persuasion and value in legal argument. *J. Log. Comput.*, *15*(6):1075–1097.

Bench-Capon, T.M., Coenen, F., & Orton, P. (1993). Argument based explanation of the british nationality act as a logic program. *Computers, Law and AI, 2* (1): 53-66.

Bench-Capon, T.M. & Leng, P. (1994). Developing heuristics for the argument based explanation of negation in logic programs. *Technical report, First Workshop on Computational Dialectics,* Seattle, USA.

Bench-Capon, T.M., Lowes, D., & McEnery, A. (1991). Using Toulmin's Arguments Schema to Explain Logic Programs. *Knowledge Based Systems 4*, (3): 177-183.

Benferhat, S. (2000). Computing specificity in default reasoning. En Gabbay, D.M. & Smets, P. (Eds.) *Handbook of Defeasible Reasoning and Uncertainty Management Systems* (pp. 147-177). Springer Netherlands.

Bergmann, M. (1997). *Internalism, Externalism, and Epistemic Defeat.* PhD Dissertation: University of Notre Dame.

Besnard, P., & Hunter, A. (2001). A logic-based theory of deductive arguments. *Artificial Intelligence*, *128*(1-2):203–235.

Besnard, P., & Hunter, A. (2008). *Elements of argumentation*. Cambridge: MIT press.

Bochman, A. (2003). Collective argumentation and disjunctive logic programming. *J. Log. Comput., 13*(3): 405–428.

Bodanza, G.A. & Alessio, C.A. (2010) Sobre la aceptabilidad de argumentos en un marco argumentativo con especificidad. *Actas de la II Conferencia Internacional Lógica, Argumentación y Pensamiento Crítico* (pp. 74-81). CEAR. Santiago, Chile.

Bodanza, G.A. & Alessio, C.A. (2014). Reinstatement and the Requirement of Maximal Specificity in Argument Systems. *Logic, Language, Information, and Computation*: 81-93.

Bodanza, G.A. & Simari, G.R. (1995). Argumentación rebatible con bases disyuntivas. En *I Congreso Argentino en Ciencias de la Computación* (pp. 313-324).

Bodanza, G. A., Tohmé, F., & Simari, G. R. (2012). Argumentation Games for Admissibility and Cogency Criteria. En Verheij, B., Szeider, S., Woltran, S. (Eds.) *Computational Models of Argument. Proceedings of COMMA 2012* (pp. 153–164). IOS Press.

Bondarenko, A., Dung, P. M., Kowalski, R. A., & Toni, F. (1997). An abstract, argumentation-theoretic approach to default reasoning. *Artificial intelligence*, *93*(1), 63-101.

Boutilier, C. (1992). What is a default priority? En *Canadian Conference on AI,* (pp 140-147)

Brewka, G. (1994) Adding priorities and specificity to default logic. *En Proceedings of the European Workshop on Logics in Artificial Intelligence* (pp. 247-260). Springer.

Brewka, G. (1994). Reasoning about priorities in default logic. Proceedings of the Twelfth National Conference on Artificial Intelligence (pp. 940–945). *AAAI*.

Brewka, G. & Eiter,T. (1999). Preferred answer sets for extended logic programs, *Artificial Intelligence, 109*: 297–356.

Caminada, M (2004). *For the sake of the Argument. Explorations into argument-based reasoning*. Doctoral dissertation: Free University Amsterdam.

Caminada, M. (2006). On the issue of reinstatement argumentation, in M. Fisher, W van der Hoek, B. Konev, and A. Lisitsa (eds.), *Logics in Artificial Intelligence, 10th European Conference, JELIA 2006* (pp- 111-123). Springer.

Caminada, M. (2006b). Semi-stable semantics. En Dunne, P. & Bench-Capon, T.J.M. (Eds.) *Computational Models of Arguments* (pp. 121–130). IOS Press.

Caminada, M. (2007). An Algorithm for Computing Semi-stable Semantics. En Mellouli, K. (Ed.) *Symbolic and Quantitative Approaches to Reasoning with Uncertainty* (pp. 222–234. Springer.

Caminada, M. (2007b). Comparing two unique extension semantics for formal argumentation: Ideal and eager. In *Proceedings of the 19th Benelux Conference on Artificial Intelligence,* (pp. 81–87).

Caminada, M. & Amgoud, L. (2007). On the Evaluation of Argumentation Formalisms. *Artificial Intelligence 171*(5-6): 286–310.

Caminada, M. & Gabbay, D. (2009). A logical account of formal argumentation, *Studia Logica, 93*(2-3): 109-145.

Caminada, M. and Wu, Y. (2008). Towards an Argument Game for Stable Semantics. *En Computational Models of Natural Argument 2008,* Toulouse, France.

Cayrol, C. (1995) On the relation between argumentation and non-monotonic coherence based entailment. En *Proceedings of the 14th International Joint Conference on Artificial Intelligence* (pp. 1443–1448). Morgan Kaufmann.

Cayrol, C., Doutre, S. & Mengin, J. (2003). On Decision Problems related to the preferred semantics for argumentation frameworks. *Journal of Logic and Computation*, *13*(3), 377–403.

Chandler, J. (2013). Defeat Reconsidered. *Analysis*, *73*(1): 49-51.

Chakraborti, C. (2006). *Logic: Informal, symbolic and inductive*. Prentice Hall of India, New Delhi

Chesñevar, C. I., Maguitman, A. G., & Loui, R. P. (2000). Logical models of argument. *ACM Computing Surveys (CSUR)*, *32*(4): 337-383.

Chesñevar, C. I., Maguitman, A. & Simari, G. (2006). Argument-Based Critics and Recommenders: A Qualitative Perspective on User Support Systems. *Journal of Data and Knowledge Engineering*, *59*(2): 293–319.

Chisholm, R. (1989). *Theory of Knowledge.* 3rd Edition, New Jersey: Prentice-Hall. 1st Edition: 1966.

Clark, K. L. (1978). Negation as failure. En Hervé Gallaire & Jack Minker (Eds.) *Logic and Data Bases*. (pp. 293-322). Springer US.

Clark, P. (1990) Representing knowledge as arguments: Applying expert system technology to judgmental problem-solving. En Addis, T. R. & Muir, R. M. (Eds.) *Research and Development in Expert Systems VII* (pp. 147–159). Cambridge University Press.

Coste-Marquis, S., Devred, C. & Marquis, P. (2005). Prudent semantics for argumentation frameworks. In *Proceedings of the 17th International Conference on Tools with Artificial Intelligence* (pp. 568–572). IEEE Computer Society.

Das, S., Fox, J., & Krause, P. (1996). A unified framework for hypothetical and practical reasoning (1): theoretical foundations. In *Proccedings of the International Conference on Formal and Applied Practical Reasoning* (pp. 58-72). Berlin: Spinger Verlag.

de Kleer, J. (1986). An assumption based TMS. *Artificial Intelligence*, *28*(2): 127-162.

Delrieux, C. (1995). *Incorporando razonamiento plausible en los sistemas de razonamiento revisable.* Tesis de Maestría: Universidad Nacional del Sur.

Delgrande, J. & Schaub, T. (1994) A general approach to specificity in default reasoning. En *Proc. 4th International Conference on Knowledge Representation and Reasoning* (pp. 147-158). Morgan Kaufmann, San Mateo, CA.

Delgrande, J.P. & Schaub, T.H. (1997). Compiling specificity into approaches to nonmonotonic reasoning, *Artificial Intelligence*, *90*: 301–348.

Dimopoulos, Y., Moraitis, P., & Amgoud, L. (2009). Extending argumentation to make good decisions. En *Algorithmic Decision Theory* (pp. 225-236). Springer Berlin Heidelberg.

Downes, S. (1995). Stephen's guide to the logical fallacies. *Electronic document*.

Doyle, J. (1979). Truth Maintenance Systems. *Artificial Intelligence, 12*: 231-272.

Dunne, P.E. & Bench-Capon, T.J.M. (2003). Two Party Immediate Response Disputes: Properties and Efficiency. *Artificial Intelligence Journal*, *149*(2): 221–250.

Dung, P.M. (1995). On the acceptability of arguments and its fundamental role in nonmonotonic reasoning, logic programming and *n* person games. *Artificial intelligence*, 77(2): 321-357.

Dung, P.M., Mancarella, P. & Toni, F. (2007). Computing ideal skeptical argumentation. *Artificial Intelligence Journal*, *171* (10–15):642–674.

Dung, P.M., & Son, T.C. (1995). Nonmonotonic inheritance, argumentation and logic programming. En *Logic Programming and Nonmonotonic Reasoning* (pp. 316-329). Springer Berlin Heidelberg.

Dung, P.M. & Son, T.C. (1996). An argumentation-theoretic approach to default reasoning with specificity. En *Proc. 5th International Conference on Knowledge Representation Reasoning* (pp. 407–418). Morgan Kaufmann, San Mateo, CA.

Dung, P.M., & Son, T.C. (2000).Default reasoning with specificity. En *Computational Logic—CL 2000* (pp. 792-806). Springer Berlin Heidelberg.

Dung, P.M., & Son, T.C. (2001). An argument-based approach to reasoning with specificity. *Artificial Intelligence 133*: 35–85

Dung, P.M. & Thang, P.M. (2014). Closure and Consistency In Logic-Associated Argumentation. *Journal of Artificial Intelligence Research, 49*: 79–109.

Elkan, C. (1990). A Rational Reconstruction of Nonmonotonic Truth Maintenance Systems. *Artificial Intelligence*, 43: 219-234.

Etherington, D. (1987) Formalizing Nonmonotonic Reasoning Systems. *Artificial Intelligence* 31: 41–85.

Etherington, D.W. & Reiter, R. (1983). On Inheritance Hierarchies With Exceptions. En Genesereth, M.R. (Ed.): *Proceedings of the National Conference on Artificial Intelligence* (pp. 104-108). AAAI Press.

Falappa, M.A., García, A.J. & Simari, G.R (2004) Belief Dynamics and Defeasible Argumentation in Rational Agents. *In Proceedings of NMR 2004, section Belief Change* (pp 164–170).

Falappa, M.A., Kern-Isberner, G. & Simari, G.R. (2002). Belief Revision, Explanations and Defeasible Reasoning. *Artificial Intelligence Journal*, *141*:1–28.

Forbus, K. y de Kleer, J. (1993). *Building Problem Solvers*. MIT Press, Cambridge, Massachusetts

Fox, J. & Parsons, S. (1997). On using arguments for reasoning about action and values. En *Proceedings of the AAAI Spring Symposium on Qualitative Preferences in Deliberation and Practical Reasoning,* Stanford*.*

Fox, J., Krause, P. & Elvang-Goransson, M. (1993). Argumentation as a general framework for uncertain reasoning. In *Proccedings of the 9th Conference on Uncertainty in Artifial Intelligence* (pp. 428-434)*.* Morgan Kaufmann Publishers Inc.

Freeman, K. & Farley, A.M. (1996). A model of argumentation and its application to legal reasoning. *Artificial Intelligence and Law, 4*: 163–197.

Fuentes Bravo, C., & Santibáñez Yãnez, C. (2014). Toulmin: razonamiento, sentido común y derrotabilidad. *Kriterion: Revista de Filosofía*, *55*(130): 531-548.

García, A.J. (1997). *La Programación en Lógica Rebatible: su definición teórica y computacional*. Tesis de Maestría en Ciencias de la Computación: Universidad Nacional del Sur, Bahía Blanca, Argentina.

García, A.J. (2000). *Programación en Lógica Rebatible: lenguaje, semántica operacional y paralelismo*. Tesis de Doctor en Ciencias de la Computación: Universidad Nacional del Sur, Bahía Blanca, Argentina.

García, A. J., & Simari, G. R. (2004). Defeasible logic programming: An argumentative approach. *Journal of Theory and Practice of Logic Programming*, *4* (1-2): 95-138.

Geffner, H., & Pearl, J. (1992). Conditional entailment: bridging two approaches to default reasoning. *Artificial Intelligence*, *53* (2): 209-244.

Geerts, P. & Vermeir, D. (1993). A nonmonotonic reasoning formalism using implicit specificity information. En *Proceedings 2th International Conference on Logic Programming, NonMonotonic Reasoning Conference* (pp. 380–396). MIT Press, Cambridge, MA.

Gelfond, M. & Son, T.C. (1998). Prioritized default theory. En *Selected Papers from the Workshop on Logic Programming Knowledge Representation* (164-223). Springer, Berlin.

Goldszmidt, M. & Pearl, J. (1996). Qualitative probabilities for default reasoning, belief revision, and causal modeling, *Artificial Intelligence 84* (1–2): 57–112.

Gómez, S.A. & Chesñevar, C.I. (2004). Integrating defeasible argumentation and machine learning techniques. *Technical report*: Universidad Nacional del Sur.

Gordon, T.F. (1994). The Pleadings Game: an exercise in computational dialectics. *Artificial Intelligence and Law*, 2: 239–292.

Gordon, T.F. (1995). *The Pleadings Game. An Artificial Intelligence Model of Procedural Justice*. Dordrecht etc.: Kluwer Academic Publishers.

Gordon T.F. & Karacapilidis, N. (1997). The Zeno argumentation framework. En *Proceedings of the Sixth International Conference on Artificial Intelligence and Law* (pp. 10–18). New York: ACM Press.

Hanks S. & McDermott, D. (1987). Nonmonotonic Logic and Temporal Projection. *Artificial Intelligence,* 33: 379–412.

Hansson, S.O. (2007). La formalización en la filosofía. *Astrolabio, 4:* 43-60.

Hart, H.L.A. (1949). The ascription of responsibility and rights. *Proceedings of the Aristotelean Society* (pp. 171–194). Reprinted in *Logic and Language.* First Series, ed. A.G.N. Flew, (pp. 145–166). Oxford: Basil Blackwell, 1951.

Hempel, C.G. (1965). *La explicación científica. Estudios sobre la filosofía de la ciencia* (Aspects of scientific explanation and others essays in the philosophy of science), traducción de M. Frassineti de Gallo, Néstor Míguez e Irma Ruiz Aused, Barcelona, Piados, 1996.

Hempel, C. (1968). Maximal Specificity and Lawlikeness in Probabilistic Explanation. *Philosophy of Science* 35(2): 116–133.

Horty, J. (1994). Some Direct Theories of Nonmonotonic Inheritance. In: Gabbay, D., Hobber, C., Robinson, J. (eds.) *Handbook of Logic in Artificial Intelligence and Logic Programming. Nonmonotonic Reasoning and Uncertain Reasoning*, vol. 3 (pp. 111–187). Oxford University Press.

Horty, J.F. (2001). Argument construction and reinstatement in logics for defeasible reasoning. *Artificial Intelligence and Law*, *9*(1): 1-28.

Horty, J.F. (2007). Reasons as Defaults. *Philosophers' Imprint*, Nr. 3, Vol. 7: 1-28.

Horty, J.F. (2012). *Reasons as Defaults*. Oxford University Press.

Horty, J.F., Thomason, R.H., & Touretzky, D.S. (1990). A skeptical theory of inheritance in nonmonotonic semantic networks. *Artificial intelligence*, *42*(2): 311-348.

Jakobovits, H. & Vermeir, D. (1999). Dialectic Semantics for Argumentation Frameworks. *Proceedings of the Seventh International Conference on Artificial Intelligence and Law* (pp. 53–62). ACM.

Kakas, A. & Mancarella, P. (2013). On the semantics of abstract argumentation. *Journal of Logic and Computation*, *23*(5): 991-1015.

Klein, P. (1976). Knowledge, Causality, and Defeasibility. *The Journal of Philosophy*, *73*: 792-812.

Konolige, K. (1987). On the Relation Between Default and Autoepistemic Logic. *Artificial intelligence*, *35*(3): 343-382.

Konolige, K. (1988). Defeasible argumentation in reasoning about events. En Ras, Z.W. & Saitta, L. (Eds.) *Methodologies for Intelligent Systems* (pp. 380–390). Amsterdam: Elsevier.

Konolige, K. (1988b). Hierarchic autoepistemic theories for nonmonotonic reasoning, En Shrobe, H. E., Mitchell, T.M. & Smith, R.G. (Eds.) *Proceedings of the 7th National Conference on Artificial Intelligence* (pp. 439-443). St. Paul, MN: AAAI Press.

Koons, R. (2014). Defeasible Reasoning, En Edward N. Zalta (ed.) *The Stanford Encyclopedia of Philosophy*.

Kowalski, R. A., & Toni, F. (1996). Abstract argumentation. *Artificial intelligence and law*, *4*(3-4): 275-296.

Kraus, S., Lehmann, D. & Magidor, M. (1990). Nonmonotonic reasoning, preferential models and cumulative logics. *Artificial Intelligence 44*: 167–207.

Krause, P., Ambler, S.J., Elvang-Gøransson, M. & Fox, J. (1995). A logic of argumentation for uncertain reasoning. *Computational Intelligence 11*:113–131.

Lehrer, K & Paxson, T. (1969). Knowledge: Undefeated Justified True Belief. *Journal of Philosophy*, *66*: 225-37.

Lifschitz, V. (1985). Computing Circumscription. En Joshi, A.K. (Ed.) *Proceedings of the 9th International Joint Conference on Artificial Intelligence* (pp. 121-127). Los Angeles, CA: Morgan Kaufmann.

Lifschitz, V. (1995). Closed-world databases and circumscription. *Artificial Intelligence, 27*: 229-235.

Loui, R.P. (1987). Defeat among arguments: a system of defeasible inference. *Computational intelligence*, *3*(1): 100-106.

Loui, R.P. (1998). Process and policy: resource-bounded non-demonstrative reasoning. *Computational Intelligence*, *14*:1–38.

Loui, R.P. & Norman, J. (1995). Rationales and argument moves. *Artificial Intelligence and Law, 3*:159–189.

Loui, R.P., Norman, J., Olson, J. & Merrill, A.. (1993). A design for reasoning with policies, precedents, and rationales. *Proceedings of the Fourth International Conference on Artificial Intelligence and Law* (pp. 202–211). New York: ACM Press.

McCarthy, J. (1986). Applications of circumscription to formalizing common-sense knowledge. *Artificial Intelligence*, *28*(1): 89-116.

Marek, W. & Truszczynski, M. (1993). *Nonmonotonic logic: context-dependent reasoning.* Springer.

Martínez, D. C., García, A. J., & Simari, G. R. (2006). On the Acceptability in Abstract Argumentation Framework with an Extended Defeat Relation. En Dunne, P. & Bench-Capon, T.J.M. (Eds.) *Computational Models of Arguments* (pp. 273–278). IOS Press.

Martínez, D. C., García, A. J., & Simari, G. R. (2008). An abstract argumentation framework with varied-strength attacks. En Brewka, G. & Lang, J. (Eds.) *Principles of Knowledge Representation and Reasoning: Proceedings of the*

Eleventh International Conference (pp. 135-144). Sydney, Australia: AAAI Press.

Martínez, M. V., García, A. J., & Simari, G. R. (2012). On the Use of Presumptions in Structured Defeasible Reasoning. En Verheij, B., Szeider, S. & Woltran, S. (Eds.) *Computational Models of Argument* (pp. 185-196). Vienna, Austria: IOS Press.

McAllester, D. (1990). Truth Maintenance. En Smith, R. y Mitchell, T., (Eds.) *Proceedings of the 8th National Conference on Artificial Intelligence* (pp. 1109-1116). AAAI Press.

McCarthy, J. (1980). Circunscription – A form of Non-monotonic Reasoning. *Artificial Intelligence, 13*(1-2): 27-39.

McDermott, D. (1982). Nonmonotonic logic II: Nonmonotonic modal theories. *Journal of the ACM, 29*(1): 33-57.

McDermott, D. & Doyle, J. (1980). Non-monotonic Logic I. *Artificial Intelligence, 13* (1-2): 41-72.

Minsky, M. (1975). A Framework for Representing Knowledge. En Winston, P.H. (Ed.) *The Psychology of Computer Vision* (pp. 34-57). McGraw-Hill, NY.

Moinard, Y. (1990).Preference by Specificity in Default Logic. En *ESPRIT Project DRUMS RP11st Workshop*, Marseille, France.

Modgil, S. (2006). Value Based Argumentation in Hierarchical Argumentation Frameworks. En: Dunne, P., Bench-Capon, T. (eds.) *Computational Models of Argument* (pp. 297–308). IOS Press.

Modgil, S. (2006). Hierarchical argumentation. En Fisher, M., Hoek, W., Konev, B. & Lisitsa, A. (Eds.) *Logics in Artificial Intelligence, 10th European Conference* (319-332). Liverpool, UK: Springer.

Modgil, S. & Caminada, M. (2009). Proof Theories and Algorithms for Abstract Argumentation Frameworks. En I. Rahwan, G. R. Simari (eds.) *Argumentation in Artificial Intelligence* (pp. 105-129). Springer US.

Modgil, S., & Prakken, H. (2013). A general account of argumentation with preferences. *Artificial Intelligence*, *195*: 361-397.

Moore, R. (1984). Possible-world semantics for autoepistemic logic. En *Proceedings of the Non-Monotonic Reasoning Workshop* (pp. 344-354). New Paltz, USA: AAAI Press.

Možina, M., Žabkar, J. & Bratko, J. (2007). Argument based machine learning. *Artificial Intelligence*, *171*: 922–937.

Nute, D. (1988[a]). Defeasible Reasoning. En Fetzer, J.H. (Ed.) *Aspects of Artificial Intelligence* (pp. 251-288). Norwell, MA: Kluwer Academic Publisher.

Nute, D. (1988[b]). Defeasible Logic. En Gabbay, D. (Ed.) *Handbook of logic in Artificial intelligence and Logic Programming* (pp. 355-395). Oxford University Press.

Nute, D. (1988[c]). Defeasible reasoning: a philosophical analysis in Prolog. *In Aspects of Artificial Intelligence* (pp. 251-288). Springer Netherlands.

Nute, D. (1988[d]). Defeasible reasoning and decision support systems. *Decision Support Systems, 4*: 97-110

Parsons, S., Sierra, C., & Jennings, N. (1998). Agents that reason and negotiate by arguing. *Journal of Logic and Computation, 8*: 261-292.

Pearl, J. (1990, March). System Z: A natural ordering of defaults with tractable applications to nonmonotonic reasoning. En *Proceedings of the 3rd Conference on Theoretical Aspects of Reasoning about Knowledge* (pp. 121-135). Morgan Kaufmann Publishers Inc.

Plantinga, A. (1986). The Foundations of Theism: A Reply. *Faith and Philosophy,* 3 (3): 310-312.

Plantinga, A. (1993a). *Warrant and Proper Functions.* NY: Oxford University Press.

Plantinga, A. (1993b). *Warrant: The current debate*. NY: Oxford University Press.

Plantinga, A. (1995). Reliabilism, Analyses, and Defeater. *Philosophy and Phenomenological Research*, *55*: 427-464.

Pollock, J.L. (1974). *Knowledge and Justification*. Princeton.

Pollock, J.L. (1987). Defeasible reasoning. *Cognitive science*, *11*(4): 481-518.

Pollock, J.L. (1992). How to reason defeasible. *Artificial Intelligence*, *57*(1): 1-42.

Pollock, J.L. (1995). *Cognitive carpentry: A blueprint for how to build a person*. MIT Press.

Pollock, J.L. (2001). Defeasible reasoning with variable degrees of justification. *Artificial Intelligence*, *133*(1): 233-282.

Poole, D. (1984). A Logical System for Default Reasoning. *In Proc. AAAI Workshop on Non-Monotonic Reasoning* (pp. 373-384). NY: AAAI Press.

Poole, D. (1985). On the Comparison of Theories: Preferring the Most Specific Explanation. En Joshi, A.K. (Ed.): *Proceedings of the 9th International Joint Conference on Artificial Intelligence* (pp. 144-147). Los Angeles, CA. Morgan Kaufmann.

Poole, D. (1988) A logical framework for default reasoning. *Artificial Intelligence, 36*(1):27-48.

Poole, D. (1989). What the Lottery Paradox tells us about default reasoning. En Brachman, R.J., Levesque, H.J. & Reiter, R. (Eds.) *Proceedings of the 1st International Conference on Principles of Knowledge Representation and Reasoning* (pp. 333-340). Toronto, Canada: Morgan Kaufmann

Poole, D. (1991). The effect of knowledge on belief: conditioning, specificity and the lottery paradox in default reasoning. *Artificial Intelligence, 49*: 281-307.

Prakken, H. (1993). *Logical Tools for Modelling Legal Arguments*. PhD thesis, Vrije University, Amsterdam (Holanda).

Prakken, H. (1993b). An argumentation framework in default logic. *Annals of Mathematics and Artificial Intelligence, 9*:91–132.

Prakken, H. (1995). A semantic view on reasoning about priorities. *Proceedings of the Second Dutch/German Workshop on Nonmonotonic Reasoning*, (pp. 152–159).

Prakken, H. (1997). *Logical Tools for Modelling Legal Argument. A Study of Defeasible Reasoning in Law.* Dordrecht etc.: Kluwer Law and Philosophy Library.

Prakken, H. (2002). Intuitions and the Modelling of Defeasible Reasoning: Some Case Studies. En Benferhat, S., Giunchiglia, E. (eds.) *9th International Workshop on Non-Monotonic Reasoning* (pp. 91-102). Tolouse, France.

Prakken, H. (2005). Coherence and flexibility in dialogue games for argumentation. *Journal of logic and computation*, *15*(6): 1009-1040.

Prakken, H. (2005b). A study of accrual of arguments, with applications to evidential reasoning. En *The Tenth International Conference on Artificial Intelligence and Law, Proceedings of the Conference* (pp. 85–94). Bologna, Italy. ACM Press.

Prakken, H. (2005c). AI & Law, Logic and Argument Schemes. *Argumentation*, 19: 303-320.

Prakken, H. (2009). An abstract framework for argumentation with structured arguments. *Technical Report UU-CS-2009-019*, Department of Information and Computing Sciences, Utrecht University, Utrecht.

Prakken, H. (2010). An abstract framework for argumentation with structured arguments. *Argument and Computation*, *1*(2): 93-124.

Prakken, H. (2011). An overview of formal models of argumentation and their application in philosophy. *Studies in Logic*, 4(1): 65-86.

Prakken, H. & Sartor, G. (1995). On the relation between legal language and legal argument: assumptions, applicability and dynamic priorities. In *Proceedings of the 5th international conference on Artificial intelligence and law* (pp. 1-10). ACM Press.

Prakken, H. & Sartor, G. (1996[a]). A dialectical model of assessing conflicting arguments in legal reasoning. *Artificial Intelligence and Law 4*: 331-368.

Prakken, H. & Sartor, G. (1996[b]). A system for defeasible argumentation, with defeasible priorities. En Gabbay, D.M. & Ohlbach, H.J. (Eds.) *Practical*

Reasoning, International Conference on Formal and Applied Practical Reasoning (pp. 510-524). Bonn, Germany: Springer Berlin Heidelberg.

Prakken, H. & Sartor, G. (1997). Argument-based extended logic programming with defeasible priorities. *Journal of applied non-classical logics*, *7*(1-2): 25-75.

Prakken, H. & Sartor, G. (1998). Modelling reasoning with precedents in a formal dialogue game. *Artificial Intelligence and Law,* 6:231–287.

Prakken, H., & Sartor, G. (2009). A logical analysis of burdens of proof. En Kaptein, H. Prakken, H. & Verheij, B (Eds.) *Legal evidence and proof: Statistics, stories, logic* (pp. 223-253). Aldershot: SSRN.

Prakken, H., & Vreeswijk, G.A.W. (2000). Logics for defeasible argumentation. En Gabbay, D. M. & Franz, G. (Eds.) *Handbook of philosophical logic,* Vol. IV (pp. 219-318). Springer Netherlands.

Rahwan, I.,& Simari, G.R. (eds.) (2009). *Artificial Intelligence*. Dordrecht Heidelberg London New York: Springer.

Reiter, R. (1978). On closed world data bases. En Hervé Gallaire & Jack Minker (Eds.) *Logic and Data Bases*. (pp. 55-76). Springer US.

Reiter, R. (1980). A logic for default reasoning. *Artificial intelligence*, *13*(1); 81-132.

Reiter, R., & Criscuolo, G. (1981).On Interacting Defaults. En Hayes, P.J. (Ed.) *Proceedings of the 7th International Joint Conference on Artificial Intelligence* (pp. 270-276). Vancouver, BC, Canada: William Kaufmann.

Rescher, N. (1977). *Dialectics, a Controversy-Oriented Approach to the Theory of Knowledge.* State University of New York Press, Alabny.

Ross, W.D. (1930). *The Right and the Good*. Oxford: Oxford University Press.

Swain, M. (1974). Epistemic Defeasibility. *The American Philosophical Quarterly, 11* (1): 15-25

Sartor, G. (1994). A formal model of legal argumentation. *Ratio Juris, 7*: 212–226.

Starmans, R.J.C.M. (1996). *Logic, Argument, and Commonsense.* Doctoral Dissertation: Tilburg University.

Shoham, Y. (1988) *Reasoning about Change. Time and Causation from the Standpoint of Artificial Intelligence*. Cambridge, MA: MIT Press.

Simari, G. R. (2011). A brief overview of research in argumentation systems. En Benferhat, S. & Grant, J. *Scalable Uncertainty Management* (pp. 81-95). Dayton, OH, USA: Springer Berlin Heidelberg.

Simari, G.R., Chesñevar, C.I., García, A.J., Pierce, B.C., & Turner, D.N. (1994). The role of dialectics in defeasible argumentation. *Anales de la XIV Conferencia Internacional de la Sociedad Chilena para Ciencias de la Computación, Chile.*

Simari, G.R. & Loui, R.P. (1992). A mathematical treatment of defeasible reasoning and its implementation. *Artificial intelligence*, *53*(2): 125-157.

Toulmin, S.E. (1958). *The Uses of Argument*. Cambridge: Cambridge University Press.

Touretzky, D.S. (1984). Implicit Ordering of Defaults in Inheritance Systems. En Brachman, R.J. (Ed.) *Proceedings of the National Conference on Artificial Intelligence.* (pp. 322-325). Austin, TX: AAAI Press

Touretzky, D.S (1986). *The Mathematics of Inheritance Systems*. Morgan Kaufmann, San Mateo, CA.

Touretzky, D.S., Horty, J.F. & Thomason, R.H. (1987). A clash of intuitions: The current state of nonmonotonic multiple inheritance systems. En McDermott, J. (Ed.): *Proceedings of the 10th International Joint Conference on Artificial Intelligence* (pp 476-482). Milan, Italy: Morgan Kaufmann.

Touretzky, D.S., Thomason, R.H., & Horty, J.F. (1991). A skeptic's menagerie: Conflictors, preemptors, reinstaters, and zombies in nonmonotonic inheritance. En Mylopoulos, J. & Reiter, R. (Eds.) *Proceedings of the 12th International Joint Conference on Artificial Intelligence* (pp. 478-485). Sydney, Australia: Morgan Kaufmann Publishers Inc.

Verheij, B. (1996a). *Rules, Reasons, Arguments. Formal Studies of Argumentation and Defeat.* Doctoral Dissertation: University of Maastricht.

Verheij, B. (1996b). Two approaches to dialectical argumentation: admissible sets and argumentation stages. En Meyer, J. J.Ch. & van der Gaag, L.C. (Eds.) *Proceedings of the Eighth Dutch Conference on Artificial Intelligence* (pp. 357–368). Utrecht.

Vreeswijk, G.A.W. (1993[a]). The Feasibility of Defeat in Defeasible Reasoning. En Allen, J.F., Fikes, R. & Sandewall, E. (Eds.): *Proceedings of the 2nd International Conference on Principles of Knowledge Representation and Reasoning* (pp. 526–534). San Mateo, CA: Morgan Kaufmann Publishers Inc.

Vreeswijk, G.A.W. (1993[b]). *Studies in Defeasible Argumentation.* Doctoral dissertation: Free University Amsterdam.

Vreeswijk, G.A.W. (1993[c]). Defeasible dialectics: a controversy-oriented approach towards defeasible argumentation. *Journal of Logic and Computation, 3*:317–334.

Vreeswijk, G.A.W. (1995). The computational value of debate in defeasible reasoning. *Argumentation, 9*: 305–342.

Vreeswijk, G.A.W. (1997). Abstract argumentation systems. *Journal of Artificial Intelligence*, *90*: 225-279.

Vreeswijk, G.A.W. (2000). Representation of formal dispute with a standing order. *Artificial Intelligence and Law*, *8*(2-3): 205-231.

Vreeswijk, G. A. W. & Prakken, H. (2000). Credulous and skeptical argument games for preferred semantics. En Ojeda-Aciego, M., de Guzmán, I.P., Brewka, G. & Moniz Pereira, L. (Eds.) *Logics in Artificial Intelligence, European Workshop, JELIA Proceedings* , (pp. 239–253). Springer.

Walton, D. (2011). How to refute an argument using artificial intelligence. In: M. Koszowy (ed.), *Argument and computation, Studies in Logic, Grammar and Rhetoric, 23*(36): 123–154.

Wheeler, G. (2014). Defeat Reconsidered and Repaired. *The Reasoner 8* (2): 15.

Wu, Y. & Caminada, M. (2010). A Labeling-Based Justification Status of Arguments. *Studies in Logic, 3* (4): 12-29

Wu, Y (2012). *Between Argument and Conclusion, Argument-based Approaches to Discussion, Inference and Uncertainty.* Doctoral dissertation: University of Luxembourg. Luxembourg.