EPISTEMOLOGÍA E HISTORIA DE LA CIENCIA

SELECCIÓN DE TRABAJOS DE LAS XVIII JORNADAS
VOLUMEN 14 (2008)

Horacio Faas
Hernán Severgnini

Editores

ÁREA LOGICO-EPISTEMOLÓGICA DE LA ESCUELA DE FILOSOFÍA
CENTRO DE INVESTIGACIONES DE LA FACULTAD DE FILOSOFÍA Y HUMANIDADES
UNIVERSIDAD NACIONAL DE CÓRDOBA

Argumentos que se atacan a sí mismos en marcos argumentativos

*Hipólito M. Hasrun**

Introducción

La lógica clásica es monótona. Esto significa que si cierta información (por ejemplo, ciertas premisas) permite derivar una conclusión, esa conclusión podrá seguir infiriéndose aunque se agregue nueva información (por ejemplo, nuevas premisas). En las lógicas no monótonas, por otra parte, la adición de nueva información puede llevar a una retractación y algunas conclusiones que era posible inferir, sostener o defender, ya no podrán inferirse. Los sistemas de razonamiento rebatible (*default reasoning*) y la lógica *default* o "por defecto" como, por ejemplo, los de Bondarenko et al. (1997), Lin y Shoham (1989), Loui (1987), Nute (1987), Pollock (1987; 1992; 1994), Poole (1985; 1988), Prakken (1993) Reiter (1980), Simari y Loui (1992), etc., permiten modelar razonamiento no monótono. Las propiedades de cada sistema dependerán de las lógicas subyacentes y del lenguaje que utilicen para modelar. En todos estos sistemas es crucial el conflicto entre argumentos, que es lo puede llevar a la retractación de una conclusión. Estos conflictos tendrán lugar en la medida en que el formalismo lo permita, es decir, dependerá de cómo se construyen los argumentos dentro de cada sistema.

Un enfoque más abstracto es el de marcos argumentativos (*argumentation frameworks*) como el de Dung (1993; 1995). En este enfoque la estructura y construcción de los argumentos es irrelevante. La propiedad de no monotonía reside en su carácter dialéctico: argumentos que una vez fueron sostenidos pueden dejar de serlo ante la aparición de nuevos argumentos conflictivos. En este sistema los conflictos entre argumentos se limitan a una clase de "ataques" (*attacks*), definidos como una relación binaria sobre el conjunto primitivo de argumentos. Esta relación es primitiva y no define el estatus de los argumentos; que un argumento ataque a otro no implica que el primero sea preferible al segundo. La evaluación de los argumentos dependerá de cuán defendibles sean dentro de cada marco. Dependiendo de las condiciones que un agente (que puede ser más o menos crédulo, escéptico, cauto, etc.) pueda requerir para considerar sostenible (o defendible ante los ataques) un argumento o conjunto de argumentos, se definen las diferentes *extensiones*.

Dentro de los sistemas de argumentación rebatible se presta especial atención a la posibilidad de la construcción de argumentos que se atacan o se derrotan a sí mismos (*self-defeating arguments*), como discuten Prakken, H. y Vreeswijk, G. (2.002), ya que suelen generar resultados antiintuitivos. En los marcos argumentativos, al no considerarse la estructura y construcción de los argumentos y al no imponerse restricciones a la relación de ataque, es posible construir marcos que contengan argumentos que se atacan a sí mismos. En estos marcos, suele ocurrir que las extensiones no se corresponden con lo que intuitivamente un agente podría sostener.

Tras exponer las nociones centrales del enfoque de P. M. Dung, se analizará, también de forma abstracta (esto es, independientemente del lenguaje y de la lógica subyacente), la noción

* Universidad Nacional del Sur

de ataque. Considerando que puede haber dos relaciones de ataque (una de ellas necesariamente simétrica pero la otra no), se propondrá que los argumentos que se atacan a sí mismos pueden ser, en su versión más sencilla, de dos clases. Esto posibilitará un refinamiento en el enfoque. Finalmente se analizará, a modo de ejemplo, un marco que contiene un argumento que se ataca a sí mismo. Se mostrará que las extensiones de Dung no satisfacen lo que sería intuitivamente esperable, y que este resultado varía merced al refinamiento propuesto.

Marcos argumentativos (Dung)

Para exponer el enfoque de los marcos argumentativos es menester introducir las definiciones de cada uno de los términos.

Definición 1. [Dung (1995:326)] (Marco argumentativo)

Un *marco argumentativo* AF es un par $AF = <AR, attacks>$, donde AR es un conjunto de argumentos y *attacks* es una relación binaria sobre AR.

AR es la información de la que se dispone. Se interpretará '(X,Y)' como 'el argumento X ataca al argumento Y'. Por otra parte, se dirá que un conjunto de argumentos S ataca a un argumento A, si A es atacado por algún argumento contenido en S.

La relación de ataque no es aquí una relación de derrota: que el argumento A ataque al B no significa que deba creerse A o que deba dejar de creerse B. No evalúa los argumentos. Serán los distintos ataques del marco los que determinarán el éxito de los ataques y de los argumentos.

Definición 2 [Dung (1995:326)] (Conjunto libre de conflictos)

Un conjunto de argumentos S se dice *libre de conflictos* si no existen argumentos A y B en S tales que A ataca a B.

Esta propiedad de estar libre de conflicto puede relacionarse con la noción de consistencia. se espera de un agente racional que sea consistente en sus creencias. En términos de los marcos argumentativos, se espera que los argumentos que el agente sostiene (los que cree, los que resultan airosos ante los conflictos) no se ataquen entre sí.

Definición 3 [Dung (1995:326)] (Argumento aceptable)

Un argumento $A \in AR$ se dice *aceptable* con respecto a un conjunto de argumentos S si, y sólo si, para todo argumento $B \in AR$: si B ataca a A, entonces S ataca a B.

Definición 4 [Dung (1995:326)] (Conjunto admisible)

Un conjunto de argumentos libre de conflictos S es *admisible* si, y sólo si, cada argumento A *en* S es aceptable con respecto a S.

Hasta aquí, las nociones mínimas. A continuación, utilizando las nociones de conjunto admisible y argumento aceptable, se definen las extensiones.

Definición 5 [Dung (1995:327)] (Extensión preferida -*preferred*-)

Una *extensión preferida* de un marco argumentativo AF es un conjunto máximamente admisible (con respecto a la inclusión de conjuntos) de AF.

Es decir, una extensión preferida es un conjunto admisible S tal que no existe en el marco un conjunto que contenga a S y sea también admisible. Como $\varnothing$ es un conjunto admisible y está incluido en todo otro conjunto, resulta que todo marco tiene al menos una extensión preferida.

Definición 6 [Dung (1995:328)] (Extensión estable -*stable*-)

Un conjunto de argumentos S libre de conflicto es una *extensión estable* si, y sólo si, S ataca a todo argumento que no pertenece a S.

No todos los marcos tienen extensión estable. Para definir la extensión *grounded*, es preciso definir antes la función característica

Definición 7 [Dung (1995:328)] (Función característica)

La *función característica* de un marco argumentativo AF, denotada F_{AF}, es una función F_{AF}: $P(AR) \rightarrow P(AR)$ tal que para todo conjunto S de argumentos $F_{AF}(S) = \{A \mid A$ es aceptable con respecto a $S\}$.

Es decir, la función característica es una función que va del conjunto potencia de AR al conjunto potencia de AR: toma cada subconjunto S de AR y le asigna el conjunto de todos los argumentos de AR que son aceptables en S.

Definición 8 [Dung (1995:329)] (Extensión fija -*grounded*-)

Se llama *extensión fija* de un marco argumentativo AF al menor punto fijo de la función característica F_{AF}.

Cuando la función característica F_{AF} se aplica a un conjunto S y el resultado es el mismo conjunto S, se dice que el conjunto S es un punto fijo de la función. El menor de esos puntos fijos será la extensión fija.

Definición 9 [Dung (1995:329)] (Extensión completa -*complete*-)

Un conjunto admisible de argumentos S se denomina *extensión completa* si, y sólo si, todo argumento aceptable con respecto a S pertenece a S.

Estas nociones se aplicarán al ejemplo motivador que se propone para mostrar algunos problemas.

Ejemplo 1

Sea el marco argumentativo $M=<\{A,B\}, \{(A,A),(A,B)\}>$

Extensiones: Preferida=Fija=Completa=$\varnothing$. Sin extensión estable.

El ejemplo que se pretende discutir aquí es sencillo. Se tiene un marco con dos argumentos A y B tales que A se ataca a sí mismo y a B. En otras palabras, se tiene un argumento B cuyo único conflicto es ser atacado por un argumento que se ataca a sí mismo. Lo intuitivo aquí sería que del conflicto resultara vencedor B, ya que su atacante es "defectuoso" o difícil de sostener. Sin embargo, como se ve, todas las extensiones del marco o dan $\varnothing$ o no existen. Esto es lo que resulta antiintuitivo.

La propuesta entonces es analizar las posibilidades de que A se ataque a sí mismo.

Conflictos entre argumentos

En principio, hay dos clases de conflicto o ataque entre argumentos [Pollock (1990), Prakken y Vreeswijk (2002)]. La primera clase se da cuando las conclusiones de los argumentos son contradictorias o contrarias. La segunda clase se da cuando la conclusión de un argumento es la negación de alguna inferencia efectuada en otro argumento.

Definición 10 (Ataque refutador -rebbutting-)

Un argumento A ataca refutando a un argumento B (A refuta a B) si la conclusión de A es la negación de B.

Un ataque refutador es necesariamente simétrico: A refuta a B y viceversa. Sin embargo, no todo ataque simétrico es necesariamente refutador. Como se verá lo relevante del análisis es que algunos ataques son necesariamente simétricos y otros no.

Definición 11 (Ataque por socavación -undercutting-)

Un argumento A ataca por socavación a un argumento B (A es un atacante por socavación de B) si la conclusión de A contradice alguna regla de inferencia utilizada por B.

Entonces un argumento A puede entonces atacarse a sí mismo de dos maneras: el argumento A_1 ataca refutando a un subargumento A_2 (y por ser el ataque simétrico, A_2 ataca refutando a A_1) o bien el argumento A_1 ataca por socavación a A_2.

Ejemplo 1 revisado

El ejemplo 1 puede ahora revisarse: el ataque que A dirije contra sí será un ataque por socavación o por refutación. En el primer caso, una parte de A atacará a otra. En el segundo caso, la primera y la segunda parte se atacarán mutuamente. Se dividirá para el análisis a A en dos argumentos A_1 y A_2. Y se analizarán los posibles casos de ataque entre A_1, A_2 y B. Siguiendo el ejemplo original y las definiciones que se han dado de ataque, el ataque a B es un ataque por socavación (B no ataca a A, es decir, el ataque no es simétrico y, por ende, no puede ser refutador).

Ejemplo 1'

Se tienen tres casos posibles para reconstruir el marco M si se considera que el ataque que se autodirije A es un ataque refutador (o bien puede interpretarse que se trata de ataques cruzados por socavación): A_1 y A_2 se atacan y a) sólo A_1 ataca a B; b) sólo A_2 ataca a B; y c) tanto A_1 como A_2 atacan a B.

Caso a) $M=<\{A_1,A_2,B\}, \{(A_1,A_2),(A_2,A_1),(A_1,B)\}>$

Extensiones: Preferida=Completa=Estable=$\{\{A_1\},\{A_2,B\}\}$

Extensión Fija=$\varnothing$

[El caso b) es redundante: sería el mismo marco cambiando A_1 por A_2]

Caso c) $M=<\{A_1,A_2,B\}, \{(A_1,A_2),(A_2,A_1),(A_1,B),(A_2,B)\}>$

Extensiones: Preferida=Completa=Estable=$\{\{A_1\},\{A_2\}\}$

Extensión Fija=$\varnothing$

Habría además seis casos más que contemplan sólo ataques por socavación: d) A_1 ataca a A_2 y a B; e) A_2 ataca a A_1 y a B; f) A_1 ataca a A_2 y A_2 ataca a B; g) A_2 ataca a A_1 y A_1 ataca a B; h) A_1 ataca a A_2 y ambos a B; i) A_2 ataca a A_1 y ambos a B.

Caso d) $M=<\{A_1,A_2,B\}, \{(A_1,A_2),(A_1,B)\}>$

[Similar al caso e) intercambiando A_1 y A_2]

Extensiones: Preferida=Completa=Estable= Fija=$\{A_1\}$

Caso f) $M=<\{A_1,A_2,B\}, \{(A_1,A_2),(A_2,B)\}>$

[Similar al caso g) intercambiando A_1 y A_2]

Extensiones: Preferida=Completa=Estable= Fija=$\{A_1,B\}$

Caso h) $M=<\{A_1,A_2,B\}, \{(A_1,A_2),(A_1,B),(A_2,B)\}>$

[Similar al caso i) intercambiando A_1 y A_2]

Extensiones: Preferida=Completa=Estable= Fija=$\{A_1\}$

Si bien en el ejemplo 1 lo intuitivo era que resultara elegido B, al separarse el argumento A en dos partes el marco es más complejo, y los resultados intuitivamente esperable variarán. En el caso a) –y en el b)– los subargumentos de A se atacan mutuamente, y sólo uno de ellos ataca a B. Ahora B no entra en conflicto con uno de los subargumentos y entonces ese subargumento es

compatible con B. Por otra parte, el otro subargumento de A entra en conflicto con los demás argumentos del marco y también puede sostenerse. Estas intuiciones se corresponden con las dos extensiones: o bien el argumento B junto con su "defensor" (si A_1 ataca a B y A_2 ataca a A_1, puede decirse que A_2 defiende –indirectamente– a B) o bien el argumento que ataca tanto a B como al otro subargumento de A. Como se verá, en los demás casos las extensiones se corresponden con los resultados intuitivos.

En el caso c) B es atacado por ambos subargumentos y, por no atacar a su vez, no puede ser sostenido. Los subargumentos de A podrán sostenerse por separado. En el caso d) –y en el e)– hay un subargumento de A que ataca a los demás argumentos y no hay argumento que lo ataque, con lo cual intuitivamente es el único argumento sostenible. En el caso f) –y en el g)– se tiene que un subargumento de A ataca al segundo y éste a B. Entonces, B es "defendido" por un argumento con el cual no tiene conflicto, y como el atacante de B no tiene defensor, intuitivamente resultarán sostenibles sólo B y su defensor. Finalmente, en el caso h) –y en el i)– ambos subargumentos de A atacan a B como en el caso c). Como B entra en conflicto con los demás argumentos del marco y no ataca a ninguno, no puede ser sostenido. Sólo podrá sostenerse al subargumento que ataque a los demás argumentos.

Comentarios finales

Los marcos argumentativos son el enfoque más abstracto para modelar razonamiento no monótono. Con todo, surgen resultados antiintuitivos cuando los marcos incluyen argumentos que se atacan a sí mismos. El refinamiento que se ha propuesto propone que un argumento que se ataca a sí mismo puede desdoblarse en dos argumentos cuya relación de ataque puede o no ser simétrica. Se ha mostrado con un ejemplo cómo este refinamiento permite que las extensiones se acomoden a la intuición. En el ejemplo, los marcos reformulados tienen extensiones diferentes entre sí (no todos) y diferentes al ejemplo original. Si bien la intuición del ejemplo original no se satisface (B no resulta única extensión en ningún caso), esto era esperable ya que el refinamiento propuesto incorpora argumentos al marco. Las extensiones del ejemplo revisado se corresponden con la intuición.

Lo atractivo y original de la propuesta radica en su simpleza. No se introducen nociones que compliquen el enfoque ni se incrementa el formalismo. Sólo se analizan los posibles tipos de ataque y se aplican únicamente a los argumentos que se atacan a sí mismos, ya que los ataques que no sean simétricos serán (según el análisis) siempre por socavación, y en caso de ataques simétricos es indistinto el tipo de ataque. De todo ello resulta que se mejora la adecuación entre las extensiones y la intuición cuando los marcos tienen argumentos que se atacan a sí mismos y no altera en nada el resultado en los demás marcos. Quedan por analizar otros casos, y los resultados se pueden generalizar incorporando las nociones de marco bien fundado (*well-founded*), relativamente fundado (*relatively grounded*), coherente (*coherent*), controversial (*controversial*), controversial limitado (*limited controversial*), no controversial (*uncontroversial*), etc., nociones presentes en Dung (1995:331 y *ss*) que conectan las diferentes semánticas y analizan resultados sobre ciclos de ataque.

Referencias

Bondarenko, A., Dung, P. M., Kowalski, R. A., y Toni, F. (1997) *An abstract, argumentation-theoretic approach to default reasoning*, Artificial Intelligence, 93 (1–2):63–101

Dung, P. M. (1993). *On the Acceptability of Arguments and its Fundamental Role in Nonmonotonic Reasoning and Logic Programming* Proceedings of the 13th International Joint Conference on Artificial Intelligence, Morgan Kaufmann Publishers:Chambéry, págs. 852–857.

Dung, P. M. (1995). *On the Acceptability of Arguments and its Fundamental Role in Nonmonotonic Reasoning and Logic Programming and n-Person Games*, Artificial Intelligence, 77(2):321–357

Lin, F. y Shoham, Y. (1989). *Argument Systems: a uniform basis for nonmonotonic reasoning* En Levesque, H. J., Brachman, R. J., y Reiter, R. (eds.), Proceedings del 1st International Conference on Principles of Knowledge *Representation and Reasoning*, Morgan Kaufmann:Toronto, págs. 245–255

Loui, R. P. (1987). *Defeat Among Arguments: A System of Defeasible Inference*, Computational Intelligence, 3(2):100–106.

Nute, D. (1987). *Defeasible Reasoning*, Proceedings of the XX Annual Hawaii International Conference on System Sciences, págs. 470–477

Pollock, J. L. (1987). *Defeasible Reasoning*, Cognitive Science 11 (4):481–518.

Pollock, John L. (1990) *Nomic Probability and the Foundations of Induction*, Oxford University Press.New York.

Pollock, J. L. (1992) *How to reason defeasibly*, Artificial Intelligence 57(1):1–42.

Pollock, J. L. (1994). *Justification and Defeat*, Artificial Intelligence 67(2):377–407.

Poole, D. L. (1985) *On the Comparison of Theories: Preferring the Most Specific Explanation*, Proceedings of the 9th International Joint Conference on Artificial Intelligence, págs. 144–147

Poole, D. L. (1988) *A Logical Framework for Default Reasoning*, Artificial Intelligence, 36(1):27–47.

Prakken, H. (1993) *Logical Tools for Modelling Legal Argument*, Tesis Doctoral, Vrije Universiteit:Amsterdam.

Prakken, H. y Vreeswijk, G. (2.002) Logics for defeasible argumentation. En Gabbay, D. y Guenthner, F. (eds), Handbook of Philosophical Logic, Vol 4, Kluwer Academic Publishers:Dordrecht/Boston/London, 2da edición, págs. 219–318.

Reiter, R. (1980). *A Logic for Default Reasoning*, Artificial Intelligence, 13 (1–2):81–132.

Simari, G. R. y Loui, R. P. (1992). *A mathematical treatment of defeasible reasoning and its implementation*. Artificial Intelligence, 53(2–3):125–157