THEORIA. Revista de Teoría, Historia y Fundamentos de la Ciencia
ISSN: 0495-4548
theoria@ehu.es
Universidad del País Vasco/Euskal Herriko Unibertsitatea
España

Bodanza, Gustavo Adrián
La argumentación abstracta en Inteligencia Artificial: problemas de interpretación y adecuación de las semánticas para la toma de decisiones
THEORIA. Revista de Teoría, Historia y Fundamentos de la Ciencia, vol. 30, núm. 3, 2015, pp. 395-414
Universidad del País Vasco/Euskal Herriko Unibertsitatea
Donostia, España

La argumentación abstracta en Inteligencia Artificial: problemas de interpretación y adecuación de las semánticas para la toma de decisiones*

Gustavo Adrián Bodanza

Received: 06/11/2014
Final Version: 29/03/2015

BIBLID 0495-4548(2015)30:3p.395-414
DOI: 10.1387/theoria.13150

RESUMEN: El modelo de *marcos argumentativos abstractos* es actualmente la herramienta más utilizada para caracterizar la justificación de argumentos derrotables en Inteligencia Artificial. Las justificaciones se determinan en base a los ataques entre argumentos y se formalizan a través de semánticas de extensiones. Aquí sostenemos que, o bien algunos marcos argumentativos carecen de sentido bajo ciertas concepciones de ataque específicas, o bien las semánticas más usadas en la literatura, basadas en el concepto de defensa conocido como *admisibilidad*, no resultan adecuadas para justificar, en particular, argumentos para la toma de decisiones.

Palabras clave: argumentación abstracta, Inteligencia Artificial, relaciones de ataque, toma de decisiones.

ABSTRACT: The *abstract argumentation frameworks* model is currently the most used tool for characterizing the justification of defeasible arguments in Artificial Intelligence. Justifications are determined on a given attack relation among arguments and are formalized as extension semantics. In this work we argue that, contrariwise to the assumptions in that model, either some argumentation frameworks are meaningless under certain concrete definitions of the attack relation, or some of the most used extension semantics in the literature, based on the defense notion of admissibility, are not suitable in particular for the justification of arguments for decision making.

Keywords: abstract argumentation, Artificial Intelligence, attack relations, decision making.

1. Introducción

El problema de razonar y tomar decisiones a través de la argumentación ha estado presente en el campo de la Inteligencia Artificial desde hace casi tres décadas. La motivación es la representación del conocimiento y del razonamiento de sentido común, de modo de poder extraer inferencias plausibles a partir de información incompleta y potencialmente contradictoria. Para enfrentar el problema, en los comienzos los investigadores se orientaron en la búsqueda de lógicas no monótonas, i.e. formalismos en los que las inferencias pueden variar ante el cambio de la información. Un enfoque común fue el de utilizar reglas derrotables (*default*) que expresan la inferencia tentativa de una conclusión en razón de información aceptada. Por ejemplo, en la lógica default (Reiter, 1980) los teoremas de

* Agradecemos las sugerencias de dos evaluadores anónimos que han mejorado notoriamente este artículo. El trabajo se realizó en el marco del proyecto PICT2013-1489 de la Agencia Nacional de Promociones Científicas y Técnologicas (Argentina).

una teoría, expresada en un lenguaje lógico clásico, pueden «extenderse» aplicando las reglas derrotables; si distintas reglas llevan a conclusiones contradictorias, éstas aparecerán en distintas *extensiones* de la teoría. Cada *extensión* está formada por el conjunto de las infinitas consecuencias lógicas de la teoría ampliada por las conclusiones obtenidas mediante la aplicación de las reglas derrotables. Esto plantea un primer problema: una extensión no es computable en tiempo finito. Y un segundo problema es que, en caso de haber varias extensiones, el sistema no determina cuál de ellas contiene las inferencias correctas. Lin y Shoham (1989) hablan por primera vez de *sistemas argumentativos* (*argument systems*). Proponen la idea de un argumento como un árbol cuya raíz es la conclusión y las hojas son átomos que representan hechos, y las conexiones se establecen a través de las reglas. Los autores demuestran que los formalismos más destacados de razonamiento de la época, a saber, la lógica derrotable (*default logic* —Reiter, 1980—), la lógica no monótona (*non-monotonic logic* —McDermott y Doyle, 1980—), la lógica circunscriptiva (*circunscription logic* —McCarthy, 1980—), etc. son casos especiales de sistemas argumentativos. La principal ventaja de este enfoque resultó ser el hecho de que, para determinar si una inferencia es validada por el sistema, no es necesario apelar a extensiones, sino computar sólo las conclusiones soportadas por los argumentos. Esta ventaja, que no es de orden lógico sino computacional, resultó clave para impulsar el enfoque argumentativo del razonamiento no monótono (recordemos que la discusión se daba en el campo de la Inteligencia Artificial antes que en el de la lógica).[1]

El manejo de argumentos contradictorios hizo necesaria la introducción de criterios de preferencia que llevaran a la selección de las conclusiones soportadas por los mejores argumentos. Algunos sistemas de razonamiento no monótono, como las redes de herencia semántica (*semantic inheritance networks*; e.g. Horty (1994)) implementaban el criterio de especificidad para determinar que una propiedad A usualmente presente en una clase B no puede ser «heredada» por un individuo de una subclase B' de B si usualmente A no está presente en esa subclase. Poole (1985) usa por primera vez este criterio para la comparación entre argumentos contradictorios. Otros autores también consideraron distintos criterios de comparación que podían aplicarse alternativamente, por ejemplo, el uso de evidencia preferida, como en el caso de Loui (1987).

Pero la comparación y derrota entre argumentos no es suficiente para justificar una conclusión. Desde la epistemología, Pollock (1974) venía desarrollando una teoría del razonamiento derrotable (*defeasible reasoning*) en la que presentaba la siguiente intuición: un argumento *A* está justificado (*warranted*) si, en caso de ser atacado por un argumento *B*, hay un argumento justificado *C* que defiende a *A* atacando a *B*. Simari y Loui (1992) presentaron un sistema argumentativo que combina una relación de derrota por especificidad al estilo de Poole con un mecanismo de justificación al estilo de Pollock. A estos sistemas siguieron los de Prakken y Sartor (1996), Verheij (1995), Vreeswijk (1997), etc., afianzando la línea de investigación.[2]

Brevemente, entonces, los sistemas argumentativos modelan la extracción de inferencias plausibles a partir de información incompleta, construyendo argumentos, comparán-

[1] Para un mayor detalle remitimos a Reed y Norman (2004).

[2] Chesñevar, Maguitman y Loui (2000) y Prakken y Vreeswijk (2002) recopilan y describen los trabajos principales.

dolos, y determinando en base a esto las conclusiones justificadas. Veamos informalmente cómo opera un sistema argumentativo a través de algunos ejemplos típicos:

Ejemplo 1

Supongamos que tenemos la información de que todos los pingüinos son aves, que las aves usualmente vuelan, que los pingüinos no vuelan y que cierto individuo, Pipo, es un pingüino. Entonces pueden construirse los siguientes argumentos derrotables:

A: Pipo es un ave; luego, podemos inferir tentativamente que Pipo vuela, ya que las aves usualmente vuelan.
B: Pipo es un pingüino; luego, podemos inferir tentativamente que Pipo no vuela, ya que los pingüinos no vuelan.

Asumiendo que el argumento *B* ataca al argumento *A* y que *B* es preferido por especificidad, *B* debe resultar aceptado (puesto que no tiene atacantes) y *A* rechazado. Ahora bien, supongamos que la información cambia de modo que la evidencia de que Pipo es un pingüino es revisada dando lugar a la construcción del siguiente argumento:

C: La información de que Pipo es un pingüino se basa en una observación dudosa, por lo que no podemos afirmar que Pipo es un pingüino.

Asumiendo que ahora tenemos un nuevo ataque de *C* hacia *B,* es de esperar que *C* resulte justificado (por no tener atacantes), *B* rechazado y *A* justificado (ya que *C* «defiende» a *A,* atacando a su atacante, *B*).

A los conjuntos de argumentos justificados se los llama *extensiones* (*extensions*)[3] y se los caracteriza aplicando esta idea de defensa según distintos criterios más o menos fuertes. Según cierto criterio, denominado «escéptico» en la literatura, un sistema argumentativo debería producir a lo sumo un conjunto de argumentos elegidos; según otro, denominado «crédulo», se pueden aceptar varios conjuntos alternativos. Así, distintos criterios dan lugar a distintas *semánticas de extensiones* (*extension semantics*). La interpretación que puede darse a éstos es que una extensión escéptica contiene sólo los argumentos indudablemente defendibles en el marco dado, mientras las extensiones crédulas contienen argumentos que son defendibles, pero no indudablemente (porque tienen atacantes), aunque tampoco son indudablemente indefendibles (porque a sus vez sus atacantes son dudosamente defendibles). Los casos que más claramente dividen las aguas son aquellos donde los ataques se dan cíclicamente, quedando todos los argumentos atacados. Por ejemplo,

Ejemplo 2

A: Juan es humanista; luego, podemos inferir tentativamente que Juan es pacifista, ya que los humanistas tienden a ser pacifistas.
B: Juan es nacionalista; luego, podemos inferir tentativamente que Juan no es pacifista, ya que los nacionalistas tienden a no ser pacifistas.

Si consideramos que estos argumentos se atacan el uno al otro sin poder determinar una diferencia de fuerza o preferencia entre ellos, entonces ninguno de ellos resultará justificado según un criterio escéptico (ya que ninguno es indudablemente defendible), o bien ambos

[3] No confundir con las extensiones de las lógicas derrotables de Reiter, que son conjuntos de fórmulas.

podrían quedar justificados alternativamente según un criterio crédulo (ya que ambos se pueden defender a sí mismos, aunque no son indudablemente defendibles).

En cualquier caso, está ampliamente aceptado que la elección de los mejores argumentos depende solamente de cómo éstos interactúan a través de los ataques, más allá de cómo se establezcan tales ataques. Así, para estudiar el problema de la justificación se pueden dejar de lado otras cuestiones, tales como cuál es la estructura interna de los argumentos o en qué sentido un argumento es mejor que otro. A este enfoque se lo conoce como *argumentación abstracta*. El modelo más extendido de argumentación abstracta es el de Dung (1995), quien define un *marco argumentativo* simplemente como un conjunto de argumentos sobre los que se da una relación de ataque. En base a la idea de defensa mencionada antes, Dung construye distintas semánticas de extensiones. Dada su simpleza y poder de análisis, este modelo marcó un hito en la investigación sobre la justificación de argumentos en el campo de la Inteligencia Artificial.

Ahora bien, puesto que en este modelo cualquier relación binaria puede, en principio, representar una relación de ataque entre argumentos, nos preguntamos si, efectivamente, todo marco argumentativo formalmente posible puede representar una situación argumentativa significativa o si, más bien, algunas configuraciones de marcos argumentativos carecen de sentido al no poder interpretarse bajo ninguna noción de ataque. Aquí nos detendremos en ciertos casos especiales que son problemáticos en relación a nuestra pregunta: ¿Qué ocurre si tenemos que A ataca a B, B ataca a C y C ataca a A? ¿Qué argumentos podrían quedar justificados en esta situación —si es que queda alguno—? Estos casos son reconocidamente problemáticos en el área. Dung caracteriza a los argumentos involucrados en ciclos de ataques de longitud impar como *controversiales*, y todas sus semánticas los excluyen de la justificación. Entiende que, directa o indirectamente, todos los argumentos en el ciclo se atacan, en definitiva, a sí mismos —todos serían paradójicos, de algún modo—. En este trabajo nos encargaremos de discutir tal interpretación. Apelaremos tanto a algunas nociones formales específicas —no abstractas— de ataque, utilizadas por algunos sistemas argumentativos, como a contraejemplos de sentido común. Por ejemplo, sostendremos que tales situaciones de ataques cíclicos suelen ocurrir en contextos de toma de decisiones, donde la elección arbitraria de cualquiera de los argumentos resulta más adecuada desde el punto de vista práctico que el rechazo de todos los argumentos involucrados. Advirtiendo que los argumentos no siempre son paradójicos, podemos pensar en ciertas semánticas que permitan su justificación. En este sentido, mencionaremos algunas semánticas de extensiones que logran capturar la idea y argumentaremos que éstas resultan más apropiadas que aquellas que simplemente excluyen a esos argumentos.

El trabajo se organiza como sigue. En la sección 2 introducimos los conceptos básicos de la argumentación abstracta en términos del modelo de marcos argumentativos de Dung. En la sección 3 definimos los argumentos controversiales y discutimos su carácter aparentemente paradójico. En la sección 4 argumentamos que algunos marcos argumentativos, en especial algunos que cuentan con argumentos controversiales, no parecen tener sentido bajo ciertas interpretaciones habituales de la relación de ataque, particularmente aquellas definidas como una combinación de conflicto y preferencia. En la sección 5 mostramos que si, en cambio, el ataque se define —siguiendo a Pollock— como derrota por bloqueo o socavamiento, entonces cualquier marco argumentativo puede representar una situación con sentido, pero entonces las semánticas de extensiones más comunes no parecen apropiadas para capturar la justificación de argumentos para la toma de decisiones. En la sección 6 presentamos un contraejemplo que muestra la necesidad de apartarse de las semánticas «estándar» y argumentamos que abandonar el criterio de admisibilidad es una alternativa apropiada

para ese fin. En la sección 7 discutimos nuestra propuesta en relación a algunos enfoques lógicos (no clásicos) de la argumentación abstracta. Finalmente, en la sección 7 presentamos nuestras conclusiones.

2. La argumentación abstracta

El problema de la justificación de argumentos puede investigarse atendiendo simplemente de los argumentos con que se cuenta y la interacción entre ellos una vez establecidos los ataques. En los sistemas argumentativos no abstractos, tales como los de Simari y Loui (1992) o Prakken y Sartor (1996), podrían reconocerse, a grandes rasgos, al menos tres «módulos» distintos: 1) el que establece un lenguaje y define la construcción de argumentos, 2) el que define las condiciones bajo las cuales un argumento derrota a otro, y 3) el que establece qué argumentos resultan justificados, i.e. las extensiones del sistema, en base a las interacciones determinadas en el segundo módulo. El problema de la justificación es, pues, objeto del módulo 3, y puede estudiarse independientemente de los módulos 1 y 2. Se conoce como *argumentación abstracta* al estudio enfocado exclusivamente en ese tercer módulo, obviando las cuestiones relativas a la definición de los argumentos y los ataques entre ellos que son objeto de los dos primeros módulos. Dung (1995) presentó el modelo hoy más extensamente utilizado de argumentación abstracta. El modelo consiste simplemente en definir un *marco argumentativo* como un par $AF = \langle AR, \mathit{ataca}\rangle$, donde AR es un conjunto de entidades llamadas argumentos, y *ataca* es una relación binaria sobre AR, es decir, $\mathit{ataca} \subseteq AR \times AR$ (notación: para denotar $(A, B) \in \mathit{ataca}$ escribiremos «A *ataca* B»). Nótese que las nociones de «argumento» y «ataque» son primitivas en el modelo. La relación *ataca* no se supone que cumpla ninguna propiedad formal en especial. (En adelante «AF» denotará un marco argumentativo arbitrario pero fijo.)

El modelado de extensiones gira en torno a dos nociones básicas de defensa: la de *aceptabilidad* de un argumento con respecto a un conjunto de argumentos y la de *admisibilidad* de un conjunto de argumentos. La idea detrás de la aceptabilidad es tomar aquellos argumentos que, si son atacados, pueden ser defendidos mediante otros ataques («contraataques»). Un argumento A es *aceptable* con respecto a un conjunto de argumentos $S \subseteq AR$ si y sólo si para todo argumento $B \in AR$ tal que B *ataca* A, existe un argumento $C \in S$ tal que C *ataca* B. Formalmente:

$$\forall B\,((B \in AR \wedge B \text{ ataca } A) \rightarrow \exists C\,(C \in S \wedge C \text{ ataca } B)).$$

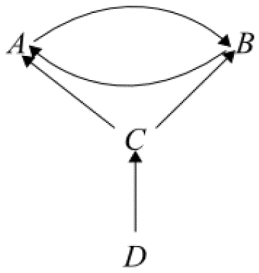
Un conjunto de argumentos $S \subseteq AR$ es *admisible* si, y sólo si, para cualesquiera argumentos A y B de S, no es el caso que A *ataca* B (i.e. S está *libre de conflictos*) y todo argumento perteneciente a S es aceptable con respecto a S.

De las distintas semánticas de extensiones que Dung propone, mencionaremos las extensiones preferidas (*preferred extensions*) y la extensión fundada (*grounded extension*), prototípicas de las semánticas crédulas y escépticas, respectivamente. Las extensiones preferidas de un marco argumentativo pueden ser múltiples, capturando la idea de que distintos conjuntos de argumentos pueden defenderse entre sí: S es una *extensión preferida* de AF si y sólo si S es un conjunto máximamente (con respecto a $\subseteq$) admisible de argumentos de AF. Por otro lado, la semántica escéptica se obtiene tomando el menor punto fijo de un operador F que, aplicado a un conjunto de argumentos, devuelve todos los argumentos acepta-

bles con respecto a ese conjunto. Formalmente, la *extensión fundada* de *AF* es el menor (con respecto a $\subseteq$) conjunto S tal que $F(S) = S$, donde F: Pot(AR) $\rightarrow$ Pot(AR)[4] (i.e. a cada conjunto de argumentos le hace corresponder otro conjunto de argumentos) y, para cualquier conjunto S de argumentos, $F(S) = \{A \in AR$: A es aceptable con respecto a $S\}$.

Ejemplo 3

Sea $AF_0 = \langle\{A, B, C, D\}, \{(A, B), (B, A), (C, A), (C, B), (D, C)\}\rangle$ *(figura 1). Entonces tenemos dos extensiones preferidas,* $\{A, D\}$ *y* $\{B, D\}$*, mientras la extensión fundada es* $\{D\}$*. Así, D resulta justificado en ambas semánticas, mientras A y B están sólo crédulamente justificados, de acuerdo a las extensiones preferidas.*

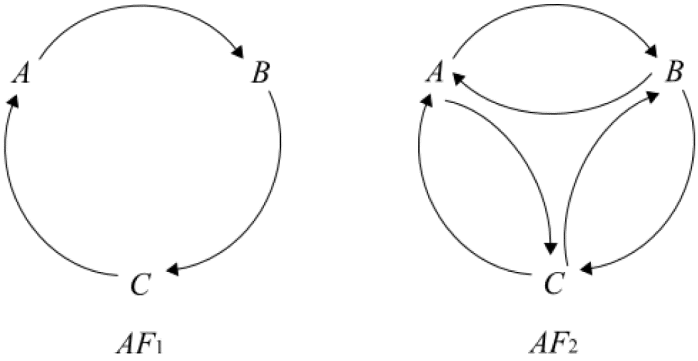


AF_0

Figura 1

Ejemplo 4

Sea $AF_1 = \langle\{A, B, C\}, \{(A, B), (B, C), (C, A)\}\rangle$*. Entonces tenemos una única extensión preferida,* $\emptyset$*, que coincide con la extensión fundada. Ningún argumento está justificado, ni escéptica ni crédulamente (figura 2, izq.).*



AF_1 AF_2

Figura 2

[4] Pot(AR) es el *conjunto potencia* de AR, i.e. la clase de todos los subconjuntos de AR.

A pesar de las diferencias entre ambas semánticas, nótese que ambas sancionan como extensiones conjuntos admisibles.

3. Los argumentos controversiales

¿Qué ocurre en el marco argumentativo del ejemplo 4? ¿Por qué allí la única extensión sancionada en cualquiera de sus semánticas es el conjunto vacío? En términos de Dung, todos los argumentos en este caso son *controversiales*: argumentos que atacan y defienden a la vez, directa o indirectamente, a algún argumento. Un argumento *A ataca indirectamente* a un argumento *B* si existe una secuencia $X_1, ..., X_{2n}$ tal que X_i *ataca* X_{i+1} $(1 \leq i < 2n)$, donde $X_1 = A$ y $X_{2n} = B$ (es decir, se da un número impar de ataques entre *A* y *B*). Por su parte, un argumento *A defiende indirectamente* a otro argumento *B* si existe una secuencia $X_1, ..., X_{2n+1}$ tal que X_i *ataca* X_{i+1} $(1 \leq i \leq 2n)$, donde $X_1 = A$ y $X_{2n+1} = B$ (es decir, se da un número par de ataques entre *A* y *B*). Un argumento *A* es *controversial* con respecto a un argumento *C* si *A* ataca y defiende indirectamente a *C.* Según estas definiciones, todo argumento que participa de un ciclo de ataques de longitud impar es controversial, resultando que cada uno de ellos ataca y defiende indirectamente, a la vez, a todos los argumentos que forman parte del ciclo incluido él mismo (i.e. entre dos argumentos cualesquiera del ciclo se puede encontrar una secuencia de ataques de longitud par y otra de longitud impar). Sin embargo, no necesariamente todo argumento controversial se ataca indirectamente a sí mismo. Por ejemplo, en el marco $AF = \langle\{A, B, C\}, \{(A, B), (B, C), (A, C)\}\rangle$, *A* es controversial con respecto a *C*, pero no se ataca indirectamente a sí mismo. Los argumentos controversiales que se atacan indirectamente a sí mismos sólo ocurren en ciclos de ataques impares. En estos casos los argumentos parecerían ser, de algún modo, paradójicos. Según esta interpretación, entonces, parece razonable que cualquier semántica de extensiones, sea escéptica o crédula, excluya a tales argumentos.

Ahora bien, cabe preguntarse si existe, o es al menos imaginable, alguna situación de la vida real donde la argumentación caiga en este tipo de ciclos. Y en caso afirmativo, si siempre resultarán paradójicos los argumentos involucrados. En primer lugar, si no existiera tal posibilidad, entonces no resultaría adecuado tratar una relación de ataque, en lo formal, como una relación binaria *arbitraria* entre argumentos: deberían excluirse al menos esos casos, pues de otro modo se generarían marcos argumentativos espurios. Por otra parte, si dichos ciclos se dieran en alguna situación argumentativa real tal que los argumentos no lucieran paradójicos, entonces podría pensarse en alguna semántica que permita justificar esos argumentos, del mismo modo que las extensiones preferidas permiten justificar argumentos involucrados en ciclos de longitud par. En lo que sigue mostraremos que, bajo ciertas interpretaciones concretas de la relación de ataque, los ciclos impares unidireccionales no pueden ocurrir, por lo que, según esas interpretaciones, no cualquier relación binaria arbitraria puede representar una relación de ataque. Por otro lado, sostendremos que existen otras interpretaciones de la relación de ataque que pueden dar lugar a ciclos impares unidireccionales, pero en estos casos resultaría apropiado adoptar otras semánticas que permitan la justificación de los argumentos involucrados.

4. La relación de ataque como conflicto y preferencia

Como hemos dicho, los marcos argumentativos abstraen todo posible significado de la relación de ataque entre argumentos al definirla, simplemente, como una relación binaria, sin más. Sin embargo, al momento de utilizar el formalismo para modelar una situación argumentativa en particular, se vuelve imprescindible preguntarse sobre mayores especificaciones que permitan una correcta aplicación. En este sentido, algunos sistemas no abstractos —expresivamente más ricos— suponen que en el ataque confluyen dos relaciones distintas: *conflicto* (o *desacuerdo —disagreement—*) y *preferencia*. La idea es sencilla: si dos argumentos están en conflicto, entonces debe prevalecer el preferido. Distintas intuiciones se han seguido para formalizar estas relaciones. Veámoslas brevemente:

— *Relación de conflicto*: dos argumentos están en conflicto si, ya sea por cuestiones lógicas, pragmáticas o de otra índole, esos argumentos no pueden ser aceptados a la vez. Por ejemplo, en una decisión médica pueden aparecer distintos argumentos a favor o en contra de aplicar determinadas terapias; mientras los expertos involucrados pueden ponderar de distinto modo las consecuencias de las terapias en cuestión, puede darse el caso de que dos argumentos que promueven la aplicación de terapias distintas, respectivamente, sean considerados en conflicto si la aplicación conjunta de ambas terapias pudiera tener consecuencias indeseadas que no tendría la aplicación de sólo una de ellas. El conflicto entre los argumentos, en tal caso, no se reconocería en cuestiones lógicas (por ejemplo, por tener conclusiones contradictorias) sino en cuestiones pragmáticas (consecuencias indeseadas de las decisiones que soportan si se aceptaran a la vez).

— *Relación de preferencia*: la preferencia entre argumentos puede estar establecida también por distintos factores y puede basarse en distintos criterios (evidencia preferida o fiabilidad de las fuentes que aportan los datos en los que se basa un argumento, la fuerza conclusiva, etc.). En cualquier caso, supone un ordenamiento de los argumentos que, en lo formal, cumple al menos la propiedad transitiva (si bien, como veremos, hay excepciones a esto).

En una primera aproximación, entonces podríamos decir que *A ataca a B* supone que 1) *A* y *B* están en conflicto y 2) que *B* no es preferido a *A*. Así la noción de ataque va cobrando sentido, aunque aún se pueden ir puliendo vaguedades. Antes de analizar distintas posibilidades con ese fin, ilustremos esta idea a través de algunos de los ejemplos vistos.

Ejemplo 5 (ejemplo 1 revisitado)

Considérense los argumentos del ejemplo 1. Por un lado, A y B están en desacuerdo por cuestiones lógicas (sus conclusiones se contradicen). Por otra parte, sin otra información disponible y teniendo en cuenta que B se basa en información más específica que A, podemos asumir que B es estrictamente preferido a A (i.e. B es al menos tan preferido como A pero A no es al menos tan preferido como B). En consecuencia, puede considerarse que B ataca a A. Por otro lado, C está en conflicto con B (la conclusión de C rechaza cierta evidencia en la que se apoya B) y, suponiendo que C se base en información confiable, C resulta preferido a B, dando lugar al ataque de C sobre B. Así, podemos representar la situación mediante el marco argumentativo $\langle\{A, B, C\}, \{(B, A), (C, B)\}\rangle$.

Ejemplo 6 (ejemplo 2 revisitado)

Considérense los argumentos del ejemplo 2. A y B están en desacuerdo por cuestiones lógicas (sus conclusiones se contradicen). Por otra parte, sin otra información disponible y asumiendo que ambos argumentos tienen la misma fuerza conclusiva, pueden tomarse como indiferentes en cuanto a preferencias (i.e., A es al menos tan preferido como B y B es al menos tan preferido como A). En consecuencia, puede entenderse que A y B se atacan mutuamente, dando lugar a su representación mediante el marco argumentativo $\langle\{A, B\}, \{(A, B), (B, A)\}\rangle$.

A continuación, analizaremos algunas variantes de esta concepción de ataque encontradas en la literatura y veremos que, cada una, da lugar a alguna de las siguientes conclusiones problemáticas:

— *Problema 1*: La relación de ataque no puede ser arbitraria: no cualquier conjunto de pares ordenados de argumentos representa la relación de ataque. Según como se la defina, la relación de ataque podría excluir algunos pares ordenados. Luego, algunos marcos argumentativos (como los define Dung) no podrán construirse o no tendrán sentido.
— *Problema 2*: Cualquier relación binaria podría ser una relación de ataque, pero en tal caso las semánticas de extensiones más usuales no reflejarán, o serán incompatibles, con las intuiciones que subyacen a la noción de ataque definida.

La primera definición a analizar es la siguiente:

(1) *A ataca B* si, y sólo si, *A* está en conflicto con *B* y *A* es al menos tan preferido como *B*.

Aquí el conflicto se entiende como una relación entre dos argumentos tales que sus conclusiones, en conjunción con las fórmulas que constituyen la base de conocimiento del sistema, implican contradicción. Esta relación es simétrica: que *A* esté en conflicto con *B* implica que *B* está en conflicto con *A*. La preferencia, por su parte, es reflexiva: todo argumento es al menos tan preferido como sí mismo; y transitiva: si *A* es al menos tan preferido como *B* y *B* es al menos tan preferido como *C*, entonces *A* es al menos tan preferido como *C*.

Kaci, van der Torre y Weydert (2006) han mostrado que esta definición da lugar a marcos argumentativos *estrictamente acíclicos* (*strictly acyclic*), lo que significa que para cualquier ciclo en la relación de ataque (*A ataca B*, *B ataca*... *ataca X*, *X ataca A*) habrá otro ciclo en el sentido inverso (*A ataca X*, *X ataca*... *ataca B*, *B ataca A*).

Observemos el marco argumentativo $AF_1 = \langle\{A, B, C\}, \{(A, B), (B, C), (C, A)\}\rangle$ (figura 2, izq.). Según la definición (1) y los resultados hallados por nuestros autores, deben darse también los ataques en sentido inverso, dando lugar al marco argumentativo $AF_2 = \langle\{A, B, C\}, \{(A, B), (B, A), (B, C), (C, B), (C, A), (A, C)\}\rangle$ (figura 2, der.). El marco argumentativo AF_1 simplemente no tendrá sentido, puesto que la relación graficada no es una relación de ataque según la definición (1). Por lo tanto, caeremos en el Problema 1.

Ahora bien, intentemos buscarle algún sentido al marco argumentativo AF_1, entendiendo la relación de ataque según la definición (1). Asumamos que ese sentido es que AF_1 debe debe ser interpretado como el marco argumentativo AF_2 (quizá como una versión «simplificada» de AF_2). En este caso, es de esperar que los argumentos justificados en AF_1 según una semántica de extensiones determinada sean los mismos argumentos justificados

según la misma semántica en AF_2. Sin embargo, este no es el caso para algunas semánticas. La semántica de extensiones preferidas de Dung, por ejemplo, sanciona para AF_1 una sola extensión: el conjunto vacío ($\emptyset$); mientras para AF_2 las extensiones son tres: $\{A\}$, $\{B\}$ y $\{C\}$. Luego, si AF_1 es un marco argumentativo válido que debe ser interpretado como AF_2, entonces la semántica preferida no permite justificar los argumentos esperados. El dilema se completa: o bien AF_1 no puede ser visto ni usado como una simplificación de AF_2 o bien algunas semánticas de extensiones no podrán capturar en AF_1 los mismos argumentos que justifica en AF_2.

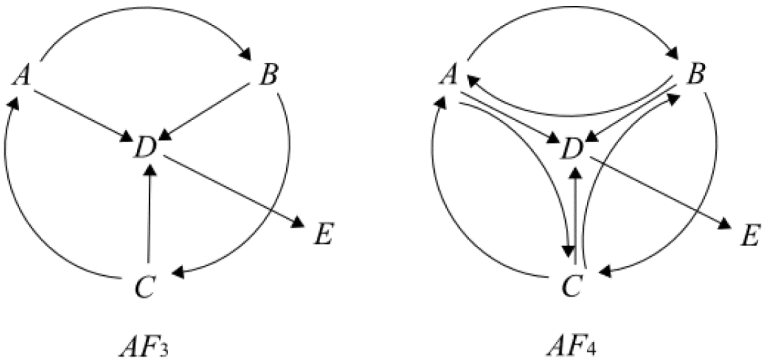


Figura 3

Por otra parte, a partir del ejemplo visto podría pensarse que el Problema 2 está limitado a las semánticas crédulas —como la de extensiones preferidas— ya que si pensamos en una semántica escéptica —como la de extensiones fundadas— tanto en AF_1 como en AF_2 se obtendrá el mismo conjunto de argumentos justificados, a saber, $\emptyset$. Sin embargo, la discriminación de las semánticas crédulas no resuelve el problema, puesto que también algunas semánticas escépticas pueden dar tratamientos distintos a AF_1 y AF_2. Por ejemplo, una semántica escéptica alternativa a la de extensiones fundadas es la que propone tomar como única extensión la intersección de todas las extensiones preferidas. Pero esta semántica también puede arrojar distintas extensiones. Veamos la figura 3, por ejemplo, donde tenemos que la única extensión preferida de AF_3 es $\emptyset$; en AF_4, en cambio, obtendríamos $\{E\}$, que es la intersección de las extensiones preferidas $\{A, E\}$, $\{B, E\}$ y $\{C, E\}$. Así, pues, no podríamos tomar a AF_3 como una representación simplificada de AF_4 que mantenga un comportamiento correcto respecto de la semántica escéptica mencionada.

Una idea de ataque más fuerte es la que exige una preferencia estricta:

(2) *A ataca B* si, y sólo si, *A* está en conflicto con *B* y *A* es estrictamente preferido a *B*.

Por ejemplo, Martínez, García y Simari. (2006, 2007, 2008) introducen la relación de derrota propia (*proper defeat*) del mismo modo que definimos el ataque en (2), distinguiéndola de la derrota bloqueante (*blocking defeat*), que se da cuando los argumentos están en conflicto pero ninguno es preferido estrictamente al otro. Nuevamente aquí la relación de conflicto es simétrica, mientras la preferencia estricta es una relación irreflexiva, asimétrica y transitiva.

Aquí caemos en el mismo dilema que con la definición (1). El marco argumentativo más simple que no puede ser interpretado según esta definición es $\langle\{A\}, \{(A, A)\}\rangle$ (un único argumento que se ataca a sí mismo); aún entendiendo que *A* puede estar en conflicto consigo mismo, como en el caso de una paradoja, es imposible que *A* sea estrictamente preferido a *A*. Pero no sólo estos casos de argumentos paradójicos no pueden ser interpretados. Tomemos nuevamente como caso el marco AF_1 (figura 2). Por la definición (2), se sigue que (i) *A*, *B* y *C* están en conflicto dos a dos y que (ii) *B* no es estrictamente preferido a *A*, ni *C* es estrictamente preferido a *B*, ni *A* es estrictamente preferido a *C*. Claramente, no puede darse el caso de que *A* sea estrictamente preferido a *B*, *B* estrictamente preferido a *C* y *C* estrictamente preferido a *A*, puesto que esto entraría en contradicción con la propiedad transitiva. Entonces, de (ii) se sigue que los argumentos o bien son indiferentes entre sí, o bien son incomparables entre sí. Pero ambos casos, en conjunción con (i), implican que deben darse también los ataques en sentido inverso, pues se trataría de derrotas bloqueantes. En consecuencia, ninguna relación de ataque construida según la definición (2) tendrá ciclos unidireccionales. Por lo tanto, un marco argumentativo con tales ciclos en la relación de ataque carecerá de sentido bajo estas condiciones. Si, por otra parte, le damos sentido a AF_1 interpretándolo como una simplificación de AF_2 —tal como planteamos para la definición de ataque (1)— entonces algunas semánticas, como la de extensiones preferidas, no podrán capturar en AF_1 los mismos argumentos justificados que en AF_2.

Dimopoulos, Moraitis y Amgoud (2008), por ejemplo, ofrecen otra posibilidad:

(3) *A ataca B* si, y sólo si, *A* está en conflicto con *B* y *B* no es al menos tan preferido como *A*.

Aquí también encontraremos marcos argumentativos sin interpretación. El contraejemplo más simple es, otra vez, el de un argumento que se auto-ataca, ya que violaría la reflexividad de la relación *al menos tan preferido como* (i.e., si *A* se ataca a sí mismo, entonces *A* no es al menos tan preferido como *A*). En general, ningún ciclo de ataque unidireccional de longitud impar puede ocurrir. Por ejemplo, AF_1 (figura 2): de la definición (3) se infiere que si *A* ataca a *B* pero *B* no ataca a *A*, entonces *A* es estrictamente preferido a *B*; si *B* ataca a *C* pero *C* no ataca a *B*, entonces *B* es estrictamente preferido a *C*; y si *C* ataca a *A* pero *A* no ataca a *C*, entonces *C* es estrictamente preferido a *A*. Nuevamente, esto viola la transitividad de la preferencia estricta.

Podríamos pensar como vía de escape que las condiciones para el ataque planteadas en (1), (2) y (3) deberían ser suficientes pero no necesarias, tal vez porque caracterizan ciertas formas de ataque, pero no todas las posibles. Pero lo cierto es que algunos sistemas agotan todas las nociones de ataque en definiciones como éstas, ya sea ignorando otras formas de ataque (por ejemplo, en los trabajos mencionados de Martínez, García y Simari se plantean las condiciones de (2) sólo como suficientes, pero no se brindan otras posibilidades de ataque más que las de derrotadores propios y bloqueantes) o estableciendo que las condiciones son además necesarias (como en el caso de Dimopoulos, Moraitis y Amgoud (2008)). Dejando tales sistemas de lado, vamos a pasar a ver qué ocurre en sistemas cuyas nociones de ataque sí brindan una instanciación posible para cada marco argumentativo abstracto.

5. Ataque como derrota por bloqueo o por socavamiento

Según la taxonomía defendida por Pollock a lo largo de su trayectoria (desde Pollock, 1970, 1974, en adelante), hay sólo dos tipos de ataque entre argumentos: por bloqueo (*rebbuting defeaters*) o por socavamiento (*undercutting defeaters*). Si P es una razón derrotable para Q, (i) R es un *derrotador por bloqueo* si y sólo si R es una razón para negar Q y (ii) R es un *socavador* si y sólo si R es una razón para negar que P justifica (*guarantees*) Q. Cualquier tipo de ataque especial —como, por caso, el ataque por especificidad— es, según el autor, un subtipo de alguno de los dos mencionados —el caso de ataque por especificidad es un subtipo de socavador— (Pollock, 2001: 235). Las derrotas por bloqueo son simétricas, ya que la conclusión de una razón derrotable niega la conclusión de la otra razón derrotable. Las derrotas por socavamiento, en cambio, no son simétricas. Según estas nociones es posible cualquier configuración de la relación de ataque de un marco argumentativo. En particular, los ataques en ciclos unidireccionales de longitud impar se deben interpretar bajo estas nociones como socavamientos (el hecho de que los ataques sigan una sola dirección impide que se los interprete como bloqueantes). Incluso se da la posibilidad de auto-socavadores, como en el caso de la *paradoja* de la paradoja de la lotería (*lottery paradox* paradox; Pollock, 2001: 241-242). Como vimos, las semánticas de Dung, tanto la de extensiones fundadas como la de extensiones preferidas, rechazan todos los argumentos del ciclo (supuesto que no haya interferencias provocadas por ataques externos al ciclo). Según la propia semántica de Pollock, ese comportamiento para estos casos también resulta adecuado. Nuestro autor ha defendido siempre una semántica de asignaciones de estado *derrotado* / *no-derrotado*, que en estos casos no puede asignar *no-derrotado* a ningún argumento (aunque tampoco puede asignar *derrotado*). En suma, según la taxonomía de derrotadores de Pollock, cualquier marco argumentativo posible tiene sentido, aunque en su semántica los argumentos involucrados en ataques cíclicos unidireccionales de longitud impar no pueden ser justificados, del mismo modo que en las semánticas de Dung. Tenemos el mismo resultado en el sistema de Prakken y Sartor (1997), ya que combina una noción de derrota (*defeat*), basada en las intuiciones de Pollock que distinguen derrotadores bloqueantes y socavadores, con un criterio de justificación basado en el operador de punto fijo de Dung.

Sin embargo sostendremos que, al menos en la argumentación para la toma de decisiones y la acción, el rechazo de todos los argumentos parece extremadamente escéptico y resultaría más justificable una elección arbitraria entre ellos. En este sentido, mencionaremos algunas semánticas que, a diferencia de las de Dung y la de Pollock, capturan la idea que sostenemos.

6. Semánticas de extensiones no admisibles para la justificación de decisiones

Consideremos el siguiente ejemplo:

Ejemplo 7

> Supongamos que con mi familia acabamos de mudarnos a otra ciudad y con mi esposa debemos elegir una escuela para nuestra hija. En una preselección encontramos tres alternativas, llamémoslas e1, e2 y e3, a las que hemos evaluado de acuerdo a tres criterios distintos: proximidad,

costo de la matrícula y ambiente social. Obviamente, sólo podemos elegir exactamente una escuela. Supongamos también que no tenemos ninguna prioridad entre los criterios de evaluación, de modo que hemos acordado que una alternativa es preferida a otra si es mejor con respecto a la mayoría de ellos. Luego de una comparación uno-a-uno de las alternativas, nos encontramos discutiendo en base a los siguientes tres argumentos:

A: «Aunque e2 es mejor que e1 con respecto a la proximidad, e1 es mejor que e2 con respecto a la matrícula y el ambiente; entonces deberíamos elegir e1»;

B: «Aunque e3 es mejor que e2 con respecto al ambiente, e2 es mejor que e3 con respecto a la proximidad y la matrícula; entonces deberíamos elegir e2»;

C: «Aunque e1 es mejor que e3 con respecto a la matrícula, e3 es mejor que e1 con respecto al ambiente y la proximidad; entonces deberíamos elegir e3».

Nótese que A es un socavador para B, B es un socavador para C y C es un socavador para A. Por ejemplo, veamos el caso del ataque de A sobre B: además de soportar la elección de *e1*, las premisas de A expresan razones para creer que las premisas de B no alcanzan para justificar la elección *e2*. La situación se puede modelar mediante el marco argumentativo $\langle\{A, B, C\}, \{(A, B), (B, C), (C, A)\}\rangle$ (cuyo gráfico es otra vez el de AF_1 en la figura 2). Ahora bien, de acuerdo tanto a la semántica de extensiones fundadas como a la de extensiones preferidas, ninguno de los conjuntos de argumentos $\{A\}$, $\{B\}$ y $\{C\}$ contienen argumentos justificados: $\varnothing$, el conjunto vacío, es la única extensión de acuerdo a ambas semánticas. Esto nos lleva a pensar que, al menos en contextos de toma de decisiones como las de este ejemplo, aquellas semánticas no resultan apropiadas ya que parece más justificable la aceptación de cualquiera de los argumentos antes que ninguno (después de todo, siguiendo a esas semánticas nuestra hija se quedaría sin escuela).

Podría pensarse que el problema radica en la diferencia entre razonamiento acerca de hechos y razonamiento práctico, y que las semánticas «estándar» se adecuan al primero y no al segundo. En nuestro ejemplo, no podemos establecer una preponderancia entre las alternativas en cuanto a consideraciones de tipo fáctico, y las semánticas «estándar» parecerían, *prima facie,* reflejar correctamente esa apreciación: no hay una razón fáctica para escoger una escuela sobre otra. Pero dado el fin de la argumentación se vuelve imperioso hacer una elección aunque sea arbitraria, y para ello necesitaríamos otras semánticas.

Podría pensarse también que, a los efectos prácticos, si la elección sólo puede ser arbitraria entonces da lo mismo si una semántica sanciona tres conjuntos de alternativas igualmente aceptables o ninguna.

Para responder a estas posibles objeciones, cambiemos un poco el ejemplo 7. Digamos que tenemos que elegir entre dos escuelas, e4 y e5, que son indiferentes en cuanto a la matrícula, pero e4 está más próxima a nuestro hogar mientras e5 es preferible en cuanto al ambiente social. Los argumentos relevantes para la elección depararán un marco argumentativo como el del ejemplo 6: $\langle\{A, B\}, \{(A, B), (B, A)\}\rangle$. Recordemos que en este marco la semántica preferida sanciona —de acuerdo con nuestra intuición— dos extensiones, $\{A\}$ y $\{B\}$. Entonces, ¿por qué en el caso de las tres escuelas, a diferencia de éste, sería correcto tener sólo la extensión vacía? Evidentemente, en un caso u otro la semántica debería comportarse de modo similar. Es decir, uno esperaría que ocurra lo mismo con cualquier número de alternativas que surjan de cualquier situación de ataques entre argumentos.

En cuanto a la dicotomía razonamiento fáctico/razonamiento práctico, tampoco parece haber una justificación aceptable para las semánticas «estándar»: no parece haber una

razón fáctica con la cual justificar que en un caso se puede optar indistintamente por cualquier argumento y en otro caso por ningún argumento.

Por último, notemos que la cuestión tampoco puede zanjarse a través de la dicotomía credulidad/escepticismo planteada para las semánticas de extensiones: como hemos visto, tanto la semántica de extensiones preferidas como la fundada rechazan a todos los argumentos en marcos como AF_1 (cf. Prakken y Vreeswijk, 2002: 252).

El problema parece radicar más bien en las condiciones que las semánticas «estándar» exigen para la defensa de un argumento. Éstas (incluyendo a las de Dung y las de Pollock) sancionan sólo extensiones que son conjuntos *admisibles* de argumentos (recordemos: un conjunto de argumentos es admisible si está libre de conflictos y todos sus argumentos son aceptables en el propio conjunto). Así es que en el caso de la elección entre e4 y e5 los conjuntos $\{A\}$ y $\{B\}$ son admisibles, pero en el caso de la elección entre e1, e2 y e3, $\{A\}$, $\{B\}$ y $\{C\}$ no lo son. Estas consideraciones han llevado a cuestionar la noción de admisibilidad, proponiendo en cambio semánticas basadas en criterios más laxos. Entre ellas podemos mencionar, por ejemplo, la semántica *CF2* de Baroni, Giacomin y Guida (2005), la semántica *stage* de Verheij (1996), la semántica *stage2* de Dvořák y Gaggle (2014) y las semánticas de extensiones tolerantes (*tolerant extensions*) de Bodanza y Thomé (2009). La semántica *stage* toma como extensiones todos los subconjuntos libres de conflicto S tales que $S \cup S^+$ es máximo (con respecto a $\subseteq$), donde $S^+ = \{A: \exists B\ (B \in S \wedge B\ ataca\ A)\}$. Las definiciones de las semánticas *stage2* y *CF2* son mucho más complejas, por lo que remitimos a los trabajos mencionados. Nos enfocaremos en la semántica tolerante, cuya definición es más intuitiva y resultará suficiente para ilustrar nuestro punto.

La semántica tolerante está basada en una relación de *fuerza argumentativa* (*cogency*) entre conjuntos de argumentos que permite elecciones más amplias que la noción de admisibilidad, en el sentido de que todo subconjunto admisible está contenido en alguna extensión, pero éstas no constituyen necesariamente conjuntos admisibles. La idea es escoger un subconjunto de argumentos como una posible *estrategia argumentativa* de un agente tal que cada argumento de la estrategia puede ser defendido contra los argumentos de otras posibles estrategias de otros agentes. La diferencia está en que mientras la noción de admisibilidad supone la defensa coherente de un argumento contra *cualquier argumento* atacante, la noción de fuerza argumentativa supone la defensa coherente de un argumento contra otra *estrategia* coherente opositora (por «coherente» entendemos aquí que no contiene argumentos que se atacan entre sí). Si entre los argumentos disponibles en una disputa entre usted y yo, su estrategia es elegir un subconjunto S y la mía es elegir un subconjunto S', usted sólo debe preocuparse de que su estrategia sea defendible nada más que de los ataques que yo pueda inferirle con mis argumentos de S'; o sea, usted no debería preocuparse por defenderse de argumentos que yo no voy a usar. La idea de fuerza argumentativa aquí es como otro tipo de «admisibilidad», de carácter relativo, no absoluto, y ciertamente más débil.

La idea está expresada formalmente del siguiente modo. Dado un marco argumentativo $\langle AR, ataca\rangle$ y dos subconjuntos $S,T \subseteq AR$, S es *al menos tan fuerte argumentativamente como* T, en símbolos, $S \geq_{cog} T$, si y sólo si S es admisible en el marco argumentativo $<AR, ataca|_{S\cup T}>$, donde $ataca|_{S\cup T} = \{(A, B): A,B \in S\cup T$ y $(A, B) \in ataca\}$ (o sea, S es admisible en el marco restringido a los ataques entre argumentos de $S \cup T$). S es *estrictamente más fuerte argumentativamente* que T, en símbolos, $S >_{cog} T$, si y sólo si $S \geq_{cog} T$ y no $T \geq_{cog} S$. Es importante tener en cuenta que las relaciones $\geq_{cog}$ y $>_{cog}$ no son transitivas y, por lo tanto,

admiten ciclos. La semántica tolerante, en efecto, está definida por conjuntos cíclicamente fuertes: Un conjunto de argumentos T es *cíclicamente fuerte* si y sólo si

$$\forall S \subseteq AR\,(S >_{cog} T \rightarrow \exists S_1 ... S_n \subseteq AR\,((S_1 = T \wedge S_n = S) \wedge S_i >_{cog} S_{i+1}))\,(1 \leq i < n)$$

(i.e. si hay una estrategia argumentativa estrictamente más fuerte que T entonces siempre se encontrará una cadena de estrategias, cada una más fuerte que la anterior pero que en algún eslabón repite a T). Luego, una *extensión tolerante* es un conjunto que satisface máximamente (con respecto a $\subseteq$) la propiedad de fuerza cíclica. Intuitivamente, esta semántica rescata aquellos argumentos que, al volver a ellos una y otra vez, nunca resultan definitivamente derrotados. Volviendo al problema del ejemplo 7 (y a su representación a través del marco argumentativo AF_1 de la figura 2), allí tenemos tres extensiones tolerantes: $\{A\}$, $\{B\}$ y $\{C\}$. De acuerdo a esta semántica, entonces, la elección de cualquiera de las tres escuelas está justificada por un argumento defendible con una estrategia cíclicamente fuerte. El lector puede corroborar que esos conjuntos también son extensiones de la semántica *stage*. ¿Y qué ocurre en casos como el de elegir entre e4 y e5, donde dos argumentos A y B se atacan mutuamente? Como es de esperar, las extensiones tolerantes serán $\{A\}$ y $\{B\}$, que a su vez son conjuntos admisibles (toda conjunto admisible está incluido en alguna extensión tolerante, pero no viceversa). Las semánticas *CF2*, *stage* y *stage2* arrojan el mismo resultado en estos casos.

En consecuencia, viendo que estas semánticas parecen comportarse de un modo más intuitivo que las semánticas «estándar», al menos con respecto a la argumentación para la toma de decisiones, pensamos que la clave de su adecuación está justamente en usar criterios de defensa más laxos que el de admisibilidad.

7. Discusión

Hemos mencionado algunos sistemas en relación a los problemas que nos han ocupado sin el ánimo de ser exhaustivos. Sin embargo, comentaremos un sistema más, el de Vreeswijk (1997) —a quien, después de todo, se debe el nombre *abstract argumentation*—. Lo particular de este sistema es que, aún siendo abstracto, establece un orden de fuerza conclusiva (*conclusive force*) entre argumentos. Un *sistema argumentativo abstracto* se define como un triplo $(L, R, \leq)$, donde L es un lenguaje proposicional con $\perp$ como elemento distinguido, R es un conjunto de reglas de inferencia, y $\leq$ es una relación (de fuerza conclusiva) reflexiva y transitiva entre argumentos. Por *argumento* entiende Vreeswijk, brevemente, cualquier elemento de L o encadenamiento de reglas de inferencia (estrictas o derrotables). La noción de ataque (*defeat*) es definida como una relación entre un conjunto de argumentos S y un argumento A. S ataca a A si S es incompatible con A (i.e., $S \cup \{A\}$ permite construir un argumento con conclusión $\perp$) y A no socava (*undermines*) a S (i.e., A no tiene mayor fuerza conclusiva que ningún argumento de S). Esta noción generaliza la definida en (3) (en el caso particular de ataque proveniente de un argumento solo a otro se entiende como un ataque de un conjunto unitario de argumentos hacia un argumento). Consecuentemente, puesto que la relación de fuerza conclusiva es reflexiva y transitiva, surgirán problemas similares a los señalados para (3). En particular, no pueden darse ciclos impares en la relación de ataque. De hecho, a tales situaciones Vreeswijk las califica de «anómalas» —lo hace en un comentario sobre el sistema de Loui (1987)— (Vreeswijk, 1997: 227-229).

Hemos visto que estas «anomalías» son tales para las semánticas basadas en admisibilidad. De modo que la propiedad de admisibilidad puede usarse como criterio para clasificar semánticas de extensiones. Baroni y Giacomin (2007), por ejemplo, han planteado una serie de «principios» —incluido el de admisibilidad— en base a los cuales evaluar las semánticas. De este modo, la evaluación no queda librada a la vaga discusión acerca de si determinados ejemplos se deben resolver o no de tal o cual manera.

También se puede recurrir a caracterizaciones axiomáticas utilizando, por ejemplo, un lenguaje lógico modal.[5] En primer lugar, porque los marcos argumentativos de Dung comparten su forma con los marcos de Kripke. Y en segundo lugar, porque ya se ha trabajado en la expresión de las semánticas basadas en admisibilidad a través de tales lógicas. Por ejemplo, Grossi (2011) define un *modelo argumentativo* como una estructura (en términos afines a los usados aquí) $M = (AF, I)$ donde $AF = \langle AR, ataca \rangle$ es un marco argumentativo e I es una función tal que $I = P \to 2^{AR}$, donde P es un conjunto de átomos proposicionales. El hecho de que un argumento $A \in AR$ pertenece al conjunto $I(p)$ es expresado por $M,A \vDash p$, que puede leerse «A pertenece al conjunto $I(p)$» o «A tiene la propiedad p» (en el modelo M). Utiliza dos operadores modales $\langle\rightarrow\rangle$ y $\langle\leftarrow\rangle$, tales que $\langle\rightarrow\rangle p$ define la clase de argumentos que atacan a algún argumento de $I(p)$, mientras $\langle\leftarrow\rangle p$ define a la clase de argumentos que son atacados por argumentos de $I(p)$. Así, $M,A \vDash \langle\rightarrow\rangle p$ se interpreta como «A ataca a algún argumento de $I(p)$» y $M,A \vDash \langle\leftarrow\rangle p$ como «A es atacado por algún argumento de $I(p)$». Esto permite expresar varios de los conceptos que hemos tratado como, por ejemplo, que A defiende a un argumento de $I(p)$: $M,A \vDash \langle\rightarrow\rangle\langle\rightarrow\rangle p$, o que el conjunto $I(p)$ está libre de conflictos: $M \vDash p \Rightarrow \neg\langle\rightarrow\rangle p$ (donde $\Rightarrow$ es el condicional material), o sea, ningún argumento de $I(p)$ ataca a algún argumento de $I(p)$. Del mismo modo, podría caracterizarse las propiedades de las extensiones de una semántica determinada. Así, la admisibilidad queda plasmada en el axioma:

(Adm) $$M \vDash p \Rightarrow (\neg\langle\rightarrow\rangle p \wedge \neg\langle\leftarrow\rangle\neg\langle\leftarrow\rangle p)$$

La componente izquierda de la conjunción expresa que $I(p)$ está libre de conflictos, y la derecha, la aceptabilidad con respecto a $I(p)$. Como hemos visto, una semántica que justifica argumentos envueltos en ciclos de ataques de longitud impar no sanciona extensiones admisibles. Supongamos que A *ataca* B, B *ataca* C y C *ataca* A. Sea $I(p) = \{A\}$. Entonces está claro que $M,A \vDash \langle\leftarrow\rangle\neg\langle\leftarrow\rangle p$ (i.e. A es atacado por un argumento que no es atacado desde $I(p)$, a saber, C). Las extensiones de una semántica tal, pues, no cumplirán con el axioma (Adm).

Utilizando el lenguaje de las lógicas dinámicas, Doutre, Herzig y Perrussel (2014) caracterizan no sólo semánticas de extensiones sino también posibles actualizaciones de un marco argumentativo por introducción o eliminación de argumentos o ataques. Mediante letras proposicionales proponen expresar que un argumento ataca a otro o que un argumento pertenece a determinado conjunto. Éstas reciben valuaciones (verdadero o falso) que son interpretadas como programas, y los programas a su vez pueden modificarse por composición, iteración y *test*, que luego dan lugar a nuevas valuaciones. Los operadores modales aquí se usan para expresar que una fórmula p es verdadera luego de cada ejecución del

[5] Este comentario ha sido sugerido por un evaluador anónimo.

programa P, en símbolos $[P]p$, o luego de alguna ejecución del programa P, en símbolos, $\langle P\rangle p$. Cuentan con un operador «de conversión» $^{--}$, tal que $[P^{--}]p$ (resp., $\langle P^{--}\rangle p$) expresa que p era verdadera antes de cada (resp., alguna) ejecución de P. De este modo pueden representarse modificaciones sobre un marco argumentativo como, por ejemplo, el forzar la aceptación de un argumento en una extensión removiendo mínimamente algunos ataques. Una inmediata aplicación de este enfoque la vemos en lo que podríamos llamar una «lógica de anuncios públicos de argumentos», de utilidad para atacar cuestiones acerca de cómo actualizar un marco argumentativo una vez que un argumento ha quedado justificado luego de un anuncio.

También con una preocupación por la dinámica argumentativa, podemos mencionar los trabajos de Booth, Kaci, Rienstra y van der Torre (2013) y Moguillansky, Rotstein, Falappa, García y Simari (2013), entre otros, cuyos enfoques formales, en cambio, se hayan ligados a la lógica del cambio de creencias (Alchourrón, Gärdenfors y Makinson, 1985).

Mencionaremos finalmente la caracterización de la aceptabilidad de argumentos a través de juegos dialécticos —ciertamente, un enfoque más corriente en la comunidad de investigadores de sistemas argumentativos que los ligados a las lógicas antes mencionadas—. Estos formalismos quizá sean los que mejor reflejan, aunque de un modo ideal, la práctica argumentativa humana, representando la defensa de un argumento por parte de un proponente frente a un oponente (una sucinta presentación puede encontrarse en Modgil y Caminada, 2009). Para el caso especial que hemos tratado acerca de los argumentos involucrados en ciclos de ataques, Bodanza y Tohmé (2014) definen un protocolo de juego específico para capturar las semánticas tolerantes. Intuitivamente, entienden la defensa de un argumento como parte de una «teoría» del proponente, representada por un conjunto de argumentos, a la que el oponente intenta derrotar con otra teoría superadora. Tratándose de una semántica más laxa que las basadas en admisibilidad, el oponente está obligado a mostrar que su teoría es «mejor», en el sentido de que satisface criterios más restrictivos que los impuestos sobre el proponente (por ejemplo, mostrando que su teoría, a diferencia de la del proponente, es admisible). Si el oponente no logra su cometido, entonces la teoría del proponente resulta aceptable. El protocolo se muestra especialmente adecuado cuando se trata de toma de decisiones prácticas, ya que el proponente tendrá estrategias ganadoras para defender cualquier elemento de un conjunto de alternativas indiferentes entre sí (recordar el ejemplo 7).

8. Conclusión

En este trabajo hemos planteado algunos problemas para los marcos argumentativos abstractos y las semánticas que hemos llamado «estándar», basadas en la noción de admisibilidad. En primer lugar, vimos que algunas nociones concretas de ataque, aquellas que lo entienden como una combinación de conflicto y preferencia, hacen imposibles ciertas configuraciones de marcos argumentativos, por lo que son incompatibles con la idea de Dung acerca de que cualquier relación binaria arbitraria puede representar una relación de ataque entre argumentos. En segundo lugar vimos que otras nociones, como aquellas definidas por Pollock como derrotadores bloqueantes y socavadores, si bien permiten interpretar cualquier relación binaria como una relación de ataque, en algunos de los marcos argumentativos las semánticas «estándar» no capturan correctamente ciertas intuiciones

acerca de qué argumentos resultan justificables. Específicamente, hemos visto el caso en que los ataques se dan en ciclos de longitud impar donde, según nuestra intuición, no es correcto rechazar a todos los argumentos involucrados, al menos cuando se trata de argumentación para la toma de decisiones. En nuestra opinión, el problema se debe a que aquellas semánticas sancionan como extensiones sólo conjuntos admisibles de argumentos, cuando la admisibilidad —tal como la define Dung— parece reflejar un criterio de defensa demasiado restrictivo.

Por otra parte, hemos argumentado que la diferencia de comportamiento entre las semánticas «estándar» y las semánticas no admisibles, particularmente la semánticas de extensiones tolerantes, *CF2*, *stage* y *stage2*, no puede ser explicada por la diferencia entre comportamiento crédulo y escéptico. La semántica preferida y la semántica fundada, prototipos canónicos de comportamiento crédulo y escéptico, respectivamente, rechazan por igual a todos los argumentos involucrados en ciclos de ataque de longitud impar. Tampoco se puede afirmar que la razón de que las semánticas «estándar», a diferencia de las no admisibles, no capturen el comportamiento deseado en casos de razonamiento para la toma de decisiones es que procuran un comportamiento adecuado al razonamiento fáctico. En tal caso, una semántica como la preferida no debería hacer diferencias entre ciclos de ataque de longitud par y de longitud impar (contrariamente a lo que vimos con los marcos argumentativos de los ejemplos 6 y 7).

Finalmente, hemos discutido distintos enfoques posibles para atacar los problemas vistos, relacionados con algunas lógicas no clásicas (modales, dinámicas, epistémicas) y con juegos dialécticos.

REFERENCIAS

Alchourrón, Carlos E., Peter Gärdenfors, and David Makinson. 1985. «On the Logic of Theory Change: Partial Meet Contraction and Revision Functions.» *Journal of Symbolic Logic* 50 (2): 510-30.

Baroni, Pietro, and Massimiliano Giacomin. 2007. «On Principle-Based Evaluation of Extension-Based Argumentation Semantics.» *Artificial Intelligence* 171 (10-15): 675-700. doi:http://dx.doi.org/10.1016/j.artint.2007.04.004.

Baroni, Pietro, Massimiliano Giacomin, and Giovanni Guida. 2005. «SCC-recursiveness: a general schema for argumentation semantics.» *Artificial Intelligence* 168 (1-2): 162-210.

Bodanza, Gustavo, and Fernando Tohmé. 2009. «Two Approaches to the Problems of Self-Attacking Arguments and General Odd-Length Cycles of Attack.» *Journal of Applied Logic* 7 (4): 403-20.

Bodanza, Gustavo, and Fernando Tohmé. 2014. «Beyond Admissibility: Accepting Cycles in Argumentation with Game Protocols for Cogency Criteria.» *Journal of Logic and Computation*. doi:10.1093/logcom/exu004. To appear.

Booth, Richard, Souhila Kaci, Tjitze Rienstra, and Leendert van der Torre. 2013. «A Logical Theory about Dynamics in Abstract Argumentation.» In *Scalable Uncertainty Management*, edited by Weiru Liu, V.S. Subrahmanian, and Jef Wijsen, 8078:148-61. Lecture Notes in Computer Science. Springer Berlin Heidelberg. http://dx.doi.org/10.1007/978-3-642-40381-1_12.

Dimopoulos, Yannis, Paulos Moraitis, and Leila Amgoud. 2008. «Characterizing the Outcomes of Argumentation-Based Integrative Negotiation.» In *Proceedings of the IEEE/WIC/ACM International Conference on Intelligent Agent Technology (IAT'08)*, edited by Li, Y., Pasi, G., Zhang, C., Cercone, N., and Cao, L., 456-460. Los Alamitos, CA: CPS.

Doutre, Sylvie, Andreas Herzig, and Laurent Perrussel. 2014. «A Dynamic Logic Framework for Abstract Argumentation.» In *Principles of Knowledge Representation and Reasoning: Proceedings of the Fourteenth International Conference, KR 2014, Vienna, Austria, July 20-24, 2014*. http://www.aaai.org/ocs/index.php/KR/KR14/paper/view/7993.

Dung, Phan Minh. 1995. «On the Acceptability of Arguments and Its Fundamental Role in Nonmonotonic Reasoning, Logic Programming and N-Person Games.» *Artificial Intelligence* 77 (2): 321-57.

Dvořák, Wolfgang, and Sarah Gaggl. 2014. «Stage Semantics and the SCC-Recursive Schema for Argumentation Semantics.» *Journal of Logic and Computation*. doi:10.1093/logcom/exu006.

Grossi, Davide. 2011. «Argumentation in the View of Modal Logic.» In *Argumentation in Multi-Agent Systems*, edited by Peter McBurney, Iyad Rahwan, and Simon Parsons, 6614:190-208. Lecture Notes in Computer Science. Springer Berlin Heidelberg. http://dx.doi.org/10.1007/978-3-642-21940-5_12.

Horty, John. 1994. «Some Direct Theories of Nonmonotonic Inheritance.» In *Handbook of Logic in Artificial Intelligence and Logic Programming*, edited by Dov Gabbay, Christopher Hogger, and Alan Robinson, John, 3:111-87. New York: Oxford University Press.

Kaci, Souhila, Leon van der Torre, and Emil Weydert. 2006. «Acyclic Argumentation: Attack = Conflict + Preference.» In *Proceedings of the 17th European Conference on Artificial Intelligence*, edited by Gerhard Brewka, Silvia Coradeschi, Ana Perini, and Paolo Traverso, 725-26. Amsterdam: IOS Press.

Lin, Fangzhen, and Yoav Shoham. 1989. «Argument Systems: A Uniform Basis for Nonmonotonic Reasoning.» In *Proceedings of the 1st International Conference on Knowledge Representation and Reasoning*, 245-55. Morgan Kaufmann Publishers.

Loui, Ronald. 1987. «Defeat among Arguments: A System of Defeasible Inference.» *Computational Intelligence* 3 (1): 100-106.

Martínez, Diego, Alejandro García, and Guillermo Simari. 2006. «On Acceptability in Abstract Argumentation Frameworks with an Extended Defeat Relation.» In *Computational Models of Argument. Proceedings of COMMA 2006*, edited by Paul Dunne and Trevor Bench-capon, 273-78. Amsterdam: IOS Press.

—. 2007. «On Defense Strenght of Blocking Defeaters in Admissible Sets.» In *Knowledge Science, Engineering and Management, Second International Conference, KSEM 2007*, edited by Zili Zhang and Jörg Siekman, 4798:140-52. Lecture Notes in Computer Science. Berlin: Springer-Verlag.

—. 2008. «Strong and Weak Forms of Abstract Argument Defense.» In *Computational Models of Argument. Proceedings of COMMA 2008*, edited by Pierre Besnard, Silvie Doutre, and Anthony Hunter, 172:216-27. Frontiers in Artificial Intelligence. Amsterdam: IOS Press.

McCarthy, John. 1980. «Circumscription. A Form of Non-Monotonic Reasoning.» *Artificial Intelligence* 13: 27-39, 171-72.

McDermott, Drew, and John Doyle. 1980. «Non-Monotonic Logic I.» *Artificial Intelligence* 13: 41-72.

Modgil, Sanjay, and Martin Caminada. 2009. «Proof Theories and Algorithms for Abstract Argumentation Frameworks.» In *Argumentation in Artificial Intelligence*, edited by Iyad Rahwan and Guillermo Simari, 105-129. Springer.

Moguillansky, Martín O., Nicolás D. Rotstein, Marcelo A. Falappa, Alejandro Javier García, and Guillermo Ricardo Simari. 2013. «Dynamics of Knowledge in \emphDeLP through Argument Theory Change.» *TPLP* 13 (6): 893-957. doi:10.1017/S1471068411000603.

Pollock, John. 1970. «The Structure of Epistemic Justification.» *American Philosophical Quarterly* 4: 62-78.

—. 1974. *Knowledge and Justification*. Princeton: Princeton University Press.

—. 1987. «Defeasible Reasoning.» *Cognitive Science* 11: 481-518.

—. 2001. «Defeasible Reasoning with Variable Degrees of Justification.» *Artificial Intelligence* 133: 233-82.

Poole, David. 1985. «On the Comparison of Theories: Preferring the Most Specific Explanation.» In *Proceedings of the Ninth International Joint Conference on Artificial Intelligence*, 144-47. Los Angeles.

Prakken, Henry, and Giovanni Sartor. 1996. «A Dialectical Model of Assessing Conflicting Arguments in Legal Reasoning.» *Artificial Intelligence and Law* 4: 331-68.

Prakken, Henry, and Gerhard Vreeswijk. 2002. «Logics for Defeasible Reasoning.» In *Handbook of Philosophical Logic*, 2nd ed., 4:217-316. Dordrecht: Kluwer Academic Publishers.

Reed, Chris, and Norman, Timothy. 2004. «A Roadmap of Research in Argument and Computation.» In *Argumentation Machines*, edited by Reed, Chris and Norman, Timothy, 9:1-13. Argumentation Library, XXV. Kluwer Academic Publishers.

Reiter, Raymond. 1980. «A Logic for Default Reasoning.» *Artificial Intelligence* 13 (1-2): 81-132.

Simari, Guillermo, and Ronald Loui. 1992. «A Mathematical Treatment of Defeasible Reasoning and Its Implementation.» *Artificial Intelligence* 53 (2): 125-57.

Verheij, Bart. 1995. «Arguments and Defeat in Argument-Based Nonmonotonic Reasoning.» In *Progress in Artificial Intelligence. 7th Portuguese Conference on Artificial Intelligence (EPIA '95)*, edited by Carlos Pinto Ferreira and Nuno Mamede, 990:213-24. Lecture Notes in Artificial Intelligence. Berlin: Springer-Verlag.

Verheij, Bart. 1996. «Two Approaches to Dialectical Argumentation: Admissible Sets and Argumentation Stages.» In *Proceedings of the 8th Dutch Conference on Artificial Intelligence (NAIC 1996)*, edited by Linda van der Gaag and John-Jules Meyer. University of Utrecht.

Vreeswijk, Gerhard. 1997. «Abstract Argumentation Systems.» *Artificial Intelligence* 90 (1): 225-79.

Gustavo Adrián Bodanza es profesor asociado en la Universidad Nacional del Sur e investigador independiente en el Consejo Nacional de Investigaciones Científicas y Tecnológicas (CONICET), Argentina. Trabaja principalmente en el desarrollo de modelos teóricos formales de argumentación y razonamiento.

Address: Departamento de Humanidades, 12 de Octubre 1098 e Instituto de Investigaciones Económicas y Sociales del Sur, San Andrés 800, Altos de Palihue, Bahía Blanca (B8000CTX), Argentina. E-mail: bodanza@gmail.com