EPISTEMOLOGÍA E HISTORIA DE LA CIENCIA

SELECCIÓN DE TRABAJOS DE LAS XVI JORNADAS

VOLUMEN 12 (2006)

José Ahumada
Marzio Pantalone
Víctor Rodríguez
Editores

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

La adecuación de las decisiones argumentativas desde la teoría de juegos

*Gustavo A. Bodanza**

Introducción

Una de las formas destacadas en que los agentes racionales tratan de establecer la adecuación de las conclusiones es a través de la argumentación. La idea es que las opiniones más apropiadas deberían ser soportadas por los argumentos más convincentes. Cuando la argumentación es vista como un fundamento para sistemas de razonamiento, el objetivo no es determinar quién gana una discusión, sino qué conclusiones son soportadas por los mejores argumentos. Además, para el propósito de determinar las conclusiones apropiadas, no importa si la discusión tiene una naturaleza polémica o cooperativa. El primer caso es lo que usualmente llamamos 'debate', y en él algunos agentes ganan y otros pierden. En una discusión cooperativa, en cambio, los agentes colaboran a fin de encontrar la mejor respuesta a algún problema suscitado. Pero aún si los fines de quienes discuten no entran en conflicto, deben buscarse y ponderarse argumentos en pro y en contra para ver cuáles son los más apropiados. Parece natural, entonces, recurrir a la idea de un juego en el que dos jugadores ficticios hacen sus movidas, esto es, seleccionan los argumentos que van a utilizar en una discusión. Un buen ejemplo que muestra el tipo de "juego" al que nos referimos es el de dos o más científicos (o equipos de científicos) que intentan poner a prueba una hipótesis, buscando datos que la confirmen o la refuten, y que luego del proceso de investigación compilan todos sus argumentos y discuten los resultados.

Si bien los enfoques juego-teoréticos de la lógica son bien conocidos (p.e., Hintikka (1973), Hintikka y Sandu (1997), Lorenzen y Lorenz (1978), Bentham (2002)), nuestro objetivo aquí es algo diferente. Por un lado, buscamos una representación juego-teorética de la argumentación abstracta, siguiendo principalmente la teoría de Dung (1995), una de las obras más influyentes de los últimos años en el campo de los sistemas argumentativos. En esta veta, veremos a una discusión como un proceso para resolver un problema. La misma idea ha sido desarrollada previamente desde un punto de vista dialógico (por ejemplo, en Vreeswijk y Prakken (2000)), pero nuestro enfoque puede ser abstraído de cualquier estructura de diálogo. Desde este punto de vista, una nueva caracterización de la adecuación de una justificación argumentativa surge en términos de consideraciones estratégicas en juegos de argumentación. En estos juegos, vamos a considerar que todos los argumentos que pueden ser usados están disponibles para dos jugadores ficticios. Sus estrategias, en la forma estratégica (normal) del juego que representa la discusión, están conformadas por subconjuntos de argumentos del conjunto de todos los argumentos conocidos.

El punto principal en teoría de juegos involucra las *soluciones* de un juego, es decir, los resultados más plausibles del mismo. En un juego de argumentación estratégica, mostraremos que las soluciones pueden variar dependiendo de lo que usualmente es interpretado en sistemas de argumentación rebatible como *actitudes epistémicas*, en una escala que se mueve entre el

* Universidad Nacional del Sur – CONICET

Epistemología e Historia de la Ciencia, Volumen 12 (2006)

escepticismo y la credulidad, lo que tiene que ver con la capacidad de justificar menos o más argumentos. Las soluciones consistirán en los argumentos más adecuados de acuerdo a las especificaciones requeridas por cada actitud.

Nuestro principal objetivo será mostrar que cualquier solución de tipo crédulo debe atenerse a los equilibrios de Nash de un juego. Argumentaremos, entonces, que esta noción fundamental en teoría de juegos es una buena herramienta para marcar los límites usualmente considerados en sistemas argumentativos sobre la credulidad de un agente.

Marcos y juegos argumentativos

Siguiendo a Dung (*op. cit.*), los entes más básicos de nuestra ontología, o sea, las piezas del juego que buscamos, serán *argumentos*. Informalmente, entendemos un argumento como un objeto del lenguaje que expresa razones que sostienen una conclusión. Los argumentos pueden estar conectados por una relación de *ataque*. Para nuestros fines, ésta será una relación arbitraria (i.e., que no cumple necesariamente ninguna propiedad relevante). En la teoría de la argumentación abstracta de Dung, un *marco argumentativo (argumentation framework)* es un par $AF = \langle AR, \Rightarrow \rangle$, donde AR es un conjunto cualquiera de argumentos y $\Rightarrow$ es una relación binaria entre elementos de AR que representa los ataques que se dan entre los argumentos. El propósito es determinar qué argumentos resultan justificados en un marco argumentativo dado teniendo en cuenta los ataques que se producen.

Consideremos ahora un marco argumentativo genérico AF. Definimos un *juego argumentativo asociado a AF* como un triplo $JA_{AF} = \langle N, E, P. N \rangle = \{1,2\}$ es el conjunto de *jugadores* (jugador 1 y jugador 2) $E = 2^{AR}$ es el conjunto de *estrategias*, el mismo para cada jugador. Los jugadores harán su movida eligiendo una estrategia posible (un subconjunto de argumentos de AR), y obtendrán una utilidad de acuerdo a si la estrategia elegida puede ser defendida o no de los argumentos que forman la estrategia del otro jugador. La utilidad está definida por la *función de pago P*, que definimos:

1) P: $E \times E \rightarrow \{0,1\} \times \{0,1\}$, donde para cada par $(A,B) \in E \times E$, A es la estrategia elegida por 1 y B es la estrategia elegida por 2, mientras para cada par $(x,y) \in \{0,1\} \times \{0,1\}$, x es la utilidad obtenida por 1 e y es la utilidad obtenida por 2. Obviando este ordenamiento, por comodidad notacional denotaremos con '$P_i(A,B)$' el pago que recibe el jugador i (i=1,2) por jugar A contra la estrategia B del otro jugador, $-i$, mientras '$P_{-i}(A,B)$' denotará el pago que recibe el jugador $-i$ por jugar B contra la estrategia A del otro jugador, i.

2)

$$P_i(A,B) = \begin{cases} 0, & \text{si A tiene argumentos que se atacan entre sí, o existe un argumento } a \in A \text{ y existe un argumento } b \in B \text{ tales que } b \Rightarrow a \text{ y para todo } c \in A, \neg(c \Rightarrow b) \text{ (o sea, A no puede defender su argumento } a \text{ del ataque de } b \in B), \\ 1, & \text{en cualquier otro caso.} \end{cases}$$

Nótese que de lo anterior se sigue que $P(A,B)=(x,y)$ sii $P(B,A)=(y,x)$, lo que significa que los pagos dependen solamente de las estrategias jugadas y no de qué jugador las elige.

Ejemplo 1

El juego asociado al marco argumentativo AF = $\langle\{a, b\}, \{(a,b), (b,a)\}\rangle$ queda representado en la matriz de la Figura 1. Las filas corresponden a las estrategias del jugador 1 y las columnas a las del jugador 2. En cada celda, el número de la izquierda corresponde a la utilidad de 1 y el de la derecha a la de 2.

		2			
		$\emptyset$	$\{a\}$	$\{b\}$	$\{a,b\}$
1	$\emptyset$	1,1	1,1	1,1	1,0
	$\{a\}$	1,1	1,1	1,1	1,0
	$\{b\}$	1,1	1,1	1,1	1,0
	$\{a,b\}$	0,1	0,1	0,1	0,0

Figura 1

El concepto más importante en teoría de juegos es el de *solución*. Una solución es un *perfil del juego* (i.e. una tupla ordenada de estrategias, una para cada jugador) que indica las jugadas más plausibles de los jugadores. Puede haber distintas nociones de solución y cada una puede dar múltiples resultados, dependiendo de cómo se espera que los jugadores hagan sus movidas. En general, los jugadores decidirán su juego procurando obtener la mayor utilidad posible. Pero muchas veces ese criterio es insuficiente. Por ejemplo, en el juego de arriba este criterio nos indica que ninguno de los jugadores elegirá la estrategia $\{a,b\}$, ya que claramente ofrece una utilidad en todos los casos menor que cualquiera de las otras. Pero saber esto no nos alcanza para determinar cuál estrategia elegirán entre las otras tres.

Tratándose en especial de juegos argumentativos, supondremos que los jugadores elegirán siempre la estrategia que contenga la mayor cantidad de argumentos posible entre aquellas que consideren buenas candidatas. Con esto representaremos el hecho de que un agente racional no tiene inconvenientes en aceptar un argumento si no cuenta con buenas razones para rechazarlo. Llamaremos *principio de maximización argumentativa* (PMA) a este supuesto. Ahora bien, el criterio según el cuál cada jugador preselecciona las estrategias candidatas puede variar. Por ejemplo, un criterio sencillo sería tomar aquellas que aseguran la mayor utilidad posible. En el ejemplo anterior, éstas serían $\emptyset$, $\{a\}$ y $\{b\}$, ya que siempre pagan 1, no importa qué movida realice el otro jugador. Luego, siguiendo el PMA, cada jugador optará indistintamente entre $\{a\}$ y $\{b\}$, ya que éstas son las máximas (c.r a $\subseteq$) entre las candidatas. Consecuentemente, el juego tendrá cuatro soluciones posibles. $(\{a\},\{a\})$, $(\{a\},\{b\})$, $(\{b\},\{a\})$ y $(\{b\},\{b\})$. O sea, ambos jugadores estarán de acuerdo respecto de $\{a\}$, o ambos estarán de acuerdo respecto de $\{b\}$, o estarán en desacuerdo eligiendo de modo diferente. En sistemas argumentativos, a las decisiones indistintas entre alternativas incompatibles se las interpreta como "crédulas"[1]. En contraposición, una decisión "escéptica" consistiría en rechazar tanto $\{a\}$ como $\{b\}$, puesto que sus argumentos se atacan mutuamente sin que ninguno prevalezca sobre el otro. El criterio escéptico se justifica entendiendo que el jugador rechaza cualquier argumento "dudoso". En nuestro ejemplo, entonces, la solución escéptica será $(\emptyset, \emptyset)$. Nótese que la decisión escéptica

no está en desacuerdo con el PMA, sino que al elegir la mayor estrategia que no contiene argumentos "dudosos", en este caso, elige la estrategia vacía.

Estas mismas actitudes (credulidad, escepticismo) han sido modeladas en distintos sistemas argumentativos, y nuestro enfoque juego-teorético en realidad no agrega nada nuevo al respecto. En el aspecto en el que sí introduce una novedad, es en que permite marcar un límite preciso para la racionalidad de las decisiones crédulas a través de una noción fundamental en teoría de juegos: la de *equilibrio de Nash*. En lo que sigue veremos algunos ejemplos que problematizan las decisiones crédulas. Luego presentaremos la noción de equilibrio de Nash, y mostraremos cómo ésta permite una demarcación entre las soluciones crédulas adecuadas y las no adecuadas.

El equilibrio de Nash como criterio de adecuación de la credulidad

Uno de los problemas más discutidos en la literatura sobre argumentación rebatible es el de los argumentos que se autoatacan (cf. Pollock(1991)). La principal dificultad radica en que no es claro de qué modo debe considerarse la interacción de estos argumentos con otros mediante ataques, ya que ni siquiera es claro qué condiciones determinan que un argumento se ataque a sí mismo. En general se piensa en argumentos paradójicos, tales como "al decir 'miento' estoy mintiendo, luego, al decir 'miento' digo la verdad." Puesto que la conclusión contradice la premisa, podemos considerar que este argumento se autoataca. Ahora bien, hay cierto acuerdo respecto de que los argumentos autoatacados son injunstificables (sobre todo teniendo en cuenta que lo contrario plantearía la dificultad metodológica de abandonar los criterios habituales de justificación, que suponen que todo conjunto de argumentos justificados debe estar libre de ataques internos). Esto es ciertamente discutible, pero para seguir con nuestro plan supondremos que es correcto. Bajo esta hipótesis es lícito concluir (aunque hay objeciones, como la de Jakobovits y Vermeir (1999)) que si un argumento autoatacado a ataca a su vez a un argumento no autoatacado b, entonces (*ceteris paribus*) b debería quedar justificado, precisamente porque su atacante a no puede ser justificado al autoatacarse. Formalmente:

Ejemplo 2

Sea $AF = \langle\{a, b\}, \{(a,a), (a,b)\}\rangle$. El juego JA_{AF} tiene la forma descripta en la Figura 2

		2			
		Ø	{a}	{b}	{a,b}
1	Ø	1,1	1,0	1,1	1,0
	{a}	0,1	0,0	0,0	0,0
	{b}	1,1	0,0	1,1	0,0
	{a,b}	0,1	0,0	0,0	0,0

Figura 2

Este ejemplo nos muestra que tanto bajo una actitud eséptica como bajo una crédula –tal como las consideramos previamente– queda sancionado el perfil (Ø, Ø) como única solución aceptable. Esto es contrario a lo que esperábamos, i.e. que la solución fuera ($\{b\}$,$\{b\}$). Si $\{b\}$

es la estrategia sugerida, ésta debería ser elegida al menos desde el punto de vista crédulo. Sin embargo no ocurre[2]. La cuestión, entonces, es definir un criterio de credulidad alternativo, lo suficientemente amplio como para incluir el resultado esperado, pero que no valide otras soluciones no adecuadas. Ahora bien, ¿cómo podemos precisar hasta qué punto una solución crédula es adecuada?

La teoría de juegos ha dado un criterio de solución conocido como *equilibrio de Nash*, que puede ser de utilidad para nuestro fin. Un equilibrio de Nash es un perfil de juego tal que ningún jugador podría mejorar su utilidad cambiando la estrategia elegida si los otros jugadores no cambian las suyas. Formalmente, (A,B) es un equilibrio de Nash en nuestros juegos argumentativos sii para toda estrategia C, $P_i(C,B) \leq P_i(A,B)$ y $P_{-i}(A,C) \leq P_{-i}(A,B)$. En estos juegos, cualquier perfil que arroje el resultado (1,1) es un equilibrio de Nash, puesto que ningún jugador podría obtener un pago mayor que 1. Claramente, cualquier solución escéptica o crédula, tal como las definimos antes, es un equilibrio de Nash. Para nosotros, esta noción marca un criterio de racionalidad mínimo que toda solución adecuada debería cumplir. En efecto, lo que nos indica es bajo qué condiciones (o sea, bajo qué elecciones de estrategias posibles) un jugador no encontraría incentivos para cambiar su decisión. Si al definir un tipo de solución "más crédulo" que el propuesto –a fin de resolver problemas como el planteado en el– halláramos soluciones que no fueran equilibrios de Nash, entonces no sería adecuado, pues estaríamos validando perfiles espurios teniendo en cuenta que los jugadores podrían desviar sus elecciones razonablemente. En consecuencia, cualquier nuevo criterio de credulidad que introduzcamos debería arrojar como soluciones sólo equilibrios de Nash. En otras palabras, aunque éste no sea un requisito suficiente, sí parece ser necesario.

Algunos autores, como Baroni *et al.* (2005) proponen una solución al problema de los argumentos autoatacados que resuelve bien el problema del Ejemplo 2, o sea, valida la elección de la estrategia $\{b\}$. Sin embargo, no cumple con nuestro requisito en todos los casos. La propuesta, llamada por los autores *semántica CF2*, tiene una formulación demasiado complicada y extensa como para exponer aquí, pero de todos modos podemos analizar el siguiente ejemplo para observar que el resultado de su aplicación da soluciones que no son equilibrios de Nash.

Ejemplo 3

Sea $AF = \langle \{a, b, c\}, \{(a,b), (b,c), (c,a)\} \rangle$. El juego asociado a este marco se muestra parcialmente en la Figura 3

		2				
		Ø	$\{a\}$	$\{b\}$	$\{c\}$	...
1	Ø	1,1	1,1	1,1	1,1	...
	$\{a\}$	1,1	1,1	1,0	0,1	...
	$\{b\}$	1,1	0,1	1,1	1,0	...
	$\{c\}$	1,1	0,1	1,0	1,1	...
	.	.	.	.	.	
	.	.	.		.	
	.	.	.	.	.	

Figura 3

En este marco argumentativo la semántica *CF2* sanciona las extensiones $\{a\}$, $\{b\}$ y $\{c\}$, lo que sugiere la elección de esos subconjuntos como estrategias alternativas válidas para el tipo de credulidad que se busca modelar (Pollock(1994) apoya una decisión de este tipo). Sin embargo, puede verse que si las respectivas elecciones de los jugadores no coinciden entre sí (o sea, si eligen los perfiles $(\{a\},\{b\})$, $(\{a\},\{c\})$, etc.), entonces el resultado del juego no va a ser un equilibrio de Nash. Otras críticas a semánticas como éstas son habituales. Algunos autores, como Dung, sostienen que no es recomendable elegir entre $\{a\}$, $\{b\}$ y $\{c\}$ puesto que los argumentos que contienen forman un ciclo impar de ataques, y entienden que así cada argumento se ataca a sí mismos de un modo indirecto (es decir, dado el ciclo $a \Rightarrow b \Rightarrow c \Rightarrow a$, a se ataca a sí mismo al defender a su atacante c del ataque de b; lo mismo ocurre con los otros argumentos).

Vamos a proponer ahora un criterio de credulidad según el cuál siempre se obtendrán equilibrios de Nash. Diremos que A es una *estrategia vindicable* si y sólo si A es un subconjunto máximo (de acuerdo con el PMA) de argumentos tal que para toda estrategia B, $P_i(A,B) \geq P_{-i}(A,B)$ Por su parte, un perfil (A,B) será una *solución vindicable* si y sólo si A y B son estrategias vindicables. Considerando los pagos, lo que una estrategia vindicable asegura es obtener al menos la misma utilidad que el otro jugador; desde el punto de vista argumentativo, lo que asegura es que el otro jugador no pueda refutar. En el
Ejemplo 2 $(\{b\},\{b\})$ es la única solución vindicable y es también un equilibrio de Nash, mientras en el Ejemplo 3 $(\emptyset,\emptyset)$ es la única solución vindicable y es también un equilibrio de Nash. Por su parte, $\{a\}$, $\{b\}$ y $\{c\}$ no son vindicables en este ejemplo. Así, hemos obtenido un noción de solución adecuada que resuelve el problema de los argumentos que se autoatacan e inhibe la justificación de argumentos "demasiado" crédulos.

Es fácil demostrar que las soluciones vindicables son "más crédulas" que las soluciones crédulas definidas inicialmente (no lo hacemos aquí por falta de espacio) Por otra parte, el siguiente resultado nos asegura lo que más nos importa, a saber, que se mantienen dentro de los límites de la racionalidad impuesta por los equilibrios de Nash.

Teorema 1

Para todo marco argumentativo *AF*, si (A,B) es una solución vindicable de JA_{AF} entonces (A,B) es un equilibrio de Nash de JA_{AF}

Prueba. Sea (A,B) una solución vindicable. Entonces A y B son ambas estrategias vindicables, lo que implica que $P_i(A,B) = P_{-i}(A,B)$. Luego, o bien (i) $P(A,B) = (0,0)$ o bien (ii) $P(A,B) = (1,1)$. Supongamos que se da (i); entonces o bien (i.a) ocurren ataques internos en A (y/o en B) o bien (i.b) existe un argumento $a \in A$ (B) que es atacado por algún $b \in B$ (A), pero b no es atacado por ningún argumento de A (B). Si (i.a) es el caso, entonces $P(\emptyset,A)=(1,0)$ (ó $P(\emptyset,B)=(1,0)$), contradiciendo que A (B) es vindicable. Si (i.b) es el caso, entonces $P(\{b\},A)=(1,0)$ (ó $P(\{b\},B)=(1,0)$), lo que también contradice la vindicabilidad de A (B). Luego, (i) no es el caso, esto es, $P(A,B) \neq (0,0)$. Por lo tanto se da (ii), $P(A,B) = (1,1)$, lo que claramente indica que (A,B) es un equilibrio de Nash. *Q.E.D.*

Conclusión

Distintas nociones de credulidad y escepticismo se han planteado en la investigación de los sistemas argumentativos. Sin embargo, hasta el momento no se había hablado de cuáles son los límites de la credulidad, esto es, cómo determinar para cualquier marco argumentativo cuál es el máximo de argumentos que se pueden justificar razonablemente. Aquí hemos defendido la noción de equilibrio de Nash como herramienta formal apropiada para marcar ese límite, previa representación de los marcos argumentativos como *juegos* argumentativos. A su vez, hemos presentado la noción de solución vindicable, que resuelve el problema de los argumentos autoatacados dentro de los límites mencionados. Nos proponemos investigar en el futuro si otros tipos de equilibrio conocidos en teoría de juegos, como los de *racionalizabilidad*, pueden convalidar los tipos de credulidad sancionados como no-racionales por los equilibrios de Nash.

Notas

[1] La noción de credulidad dada aquí se corresponde con las *semánticas preferidas* de Dung. Demostramos esto en Bodanza y Tohmé (2005).

[2] Por supuesto la semántica preferida de Dung, mencionada en la nota anterior, descarta esta posibilidad.

Bibliografía

Baroni, P., M. Giacomin, G. Guida (2005) "SCC-recursiveness: a general schema for argumentation semantics", *Artificial Intelligence*, en prensa.

Bentham, J. van (2002) "Extensive Games as Process Models", *Journal of Logic, Language and Information*, 11: 289-313

Bodanza, G., F. Tohmé (2005) "A game-theoretic representation of epistemic attitudes in argumentation frameworks", sometido a *Journal of Logic and Computation*.

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

Hintikka, J. (1973) *Logic, Language Games and Information*, Clarendon Press, London.

Hintikka, J., G. Sandu (1997) "Game-Theoretical Semantics" En J. van Benthem, A. ter Meulen (eds.), *Handbook of Logic and Language*, Elsevier, Amsterdam.

Jakobovits, H., D. Vermeir (1999) "Robust Semantics for Argumentation Frameworks", *Journal of Logic and Computation*, 9(2): 215-261

Lorenzen, P., K. Lorenz (1978) *Dialogische Logik*, Wissenschaftliche Buchgesellschaft, Darmstadt.

Pollock, J. (1991) "Self-defeating Arguments", *Minds and Machines*, 1: 367-392.

Pollock, J. (1994) "Justification and defeat", *Artificial Intelligence*, 67: 377-407.

Vreeswijk, G., Prakken, H. (2000) "Credulous and Sceptical Argument Games for Preferred Semantics", *Proc. JELIA 2000*, 239-253. LNAI 1919