IV Jornadas de Investigación en Humanidades

Homenaje a Laura Laiseca

29, 30 y 31 de agosto de 2011

Departamento de Humanidades

Universidad Nacional del Sur

ACTAS

ACTAS

IV Jornadas de Investigación en Humanidades
Homenaje a Laura Laiseca

Bahía Blanca, 29, 30 y 31 de agosto de 2011

Departamento de Humanidades
Universidad Nacional del Sur

La búsqueda de un basamento argumentativo de las preferencias tendiente a la construcción de un modelo formal de democracia deliberativa

Gustavo Bodanza
UNS - CONICET
bodanza@gmail.com

1.

La Teoría de la Elección Social (TES) trata sobre la posibilidad de determinar preferencias sociales en base a las preferencias de los individuos, sin tener en cuenta el porqué de tales preferencias. Pero desde el punto de vista de la procuración del bien común, la teoría de la democracia deliberativa propone la inclusión de toda la diversidad de intereses y opiniones de las partes afectadas por las decisiones sociales y su sometimiento a deliberación. Así, las partes deberían tener la posibilidad de argumentar para la justificación de sus preferencias a fin de mostrar los beneficios para el conjunto de la sociedad, lo que haría revisables a todas las opiniones a la luz de los argumentos mejor fundados.

Un estudio combinado de la elección social con fundamentos argumentativos se enfrenta a la dificultad de construir modelos apropiados que en el presente no están desarrollados o se encuentran en etapas de desarrollo apenas incipiente (algunos de estos últimos se conocen con el nombre de sistemas argumentativos multi-agentes). En Tohmé, Bodanza y Simari (2008) se dio un primer paso en este sentido, y luego en Bodanza y Auday (2009) se analizó la posibilidad de agregar criterios argumentativos individuales en algunos mecanismos de elección social usuales (mayoría absoluta, mecanismos basados en cuotas, etc.) con resultados pesimistas.

En este trabajo intentaremos dar un (modesto) paso más en la búsqueda de un modelo apropiado de democracia deliberativa tratando de basar las preferencias individuales en argumentos que los propios individuos sean capaces de defender. La idea es que cada individuo sea capaz de justificar sus preferencias de modo que si los argumentos que las soportan son atacados, puedan contraatacar con otros argumentos. De este modo cada individuo contará con estrategias persuasivas buscando argumentos que ataquen los argumentos de aquellos individuos que sostienen distintas preferencias.

En cuanto al aspecto técnico de la construcción del modelo, se utilizará como base la noción de *marco argumentativo* de Dung (1995), combinado con elementos comunes en los modelos de elección social, tales como individuos, alternativas y preferencias sobre alternativas.

2.

Plantearemos la idea a través de un ejemplo motivador, previamente utilizado en Thomé *et al.* (2008). Imaginemos un equipo médico con tres miembros (1, 2 y 3)

deliberando acerca de qué terapia aplicar a un paciente (alternativas p, q y r). La deliberación se enfoca en tres argumentos principales:

a: "Los síntomas x, y, z son signos de la enfermedad e_1, luego hay que aplicar la terapia p"
b: "Los síntomas x, w, z son signos de la enfermedad e_2, luego hay que aplicar la terapia q"
c: "Los síntomas x, z son signos de la enfermedad e_3, luego hay que aplicar la terapia r"

Consideremos en primer lugar que cada individuo puede tener su propio criterio para evaluar los argumentos. Por ejemplo, la opinión del médico 1 podría ser la siguiente:

- Considera compatibles las terapias p y r, pero a estas incompatibles con q.
- Prefiere el argumento b sobre el argumento c por basarse en evidencia más específica.
- Prefiere el argumento a sobre el argumento b por considerar el síntoma y más relevante que el síntoma w.

Aceptaremos aquí la siguiente noción de ataque entre argumentos, usualmente presente en la literatura (*e.g.*, Simari & Loui [1992], Kaci, van der Torre & Weydert [2005], etc.): un argumento a ataca a un argumento b si a y b están en conflicto y a es preferido a b. Teniendo en cuenta la incompatibilidad de las terapias que soportan los argumentos (según la opinión del individuo 1), podemos considerar que hay conflicto entre los argumentos a y b y entre los argumentos b y c. Por otra parte, la preferencia del individuo 1 de a sobre b y de b sobre c, que denotamos con '$a >_1 b >_1 c$', permiten representar la opinión del individuo 1 de la siguiente manera:

$$a \rightarrow_1 b \rightarrow_1 c$$

donde '$\rightarrow_1$' simboliza la relación de ataque según el individuo 1. Las opiniones de los individuos 2 y 3 podrían representarse de un modo similar. Supongamos que tales opiniones son:

$$c \rightarrow_2 b \rightarrow_2 a$$
$$b \rightarrow_3 a, b \rightarrow_3 c$$

las cuales a su vez pueden estar basadas, respectivamente, en las preferencias

$$c >_2 b >_2 a$$
$$b >_3 a, b >_3 c$$

Ahora bien, con lo visto hasta aquí podemos pensar en un modo, de apariencia razonable en principio, de fundamentar las soluciones grupales en la agregación de las preferencias sobre argumentos. Si aplicáramos el mecanismo de mayoría absoluta, por ejemplo, obtendríamos el siguiente orden de preferencia social:

$b > a, b > c$

ya que hay dos de tres individuos (2 y 3) que prefieren a *b* sobre *a* y dos de tres individuos (1 y 3) que prefieren a *b* sobre *c*. En consecuencia, el criterio grupal de ataque entre argumentos sería:

$b \rightarrow a, b \rightarrow c$

coincidiendo con la opinión del individuo 3. De acuerdo a este criterio el grupo optaría por aplicar la terapia *q* ya que, según este mismo criterio, ningún argumento logra atacar al argumento que la soporta, a saber, *b*.

En resumen, este tipo de procedimientos podría evaluarse como modo de obtener una decisión social deliberada, es decir, no votando directamente sobre las alternativas, sino votando sobre los criterios de evaluación de los argumentos que soportan cada alternativa.

3.

Pero también podemos pensar el problema desde otro punto de vista. Tal como la TES supone, una decisión social puede obtenerse agregando las preferencias individuales sobre las alternativas (en lugar de agregar las preferencias sobre argumentos, como hemos hecho arriba). Sin embargo, las preferencias individuales se suponen dadas en la TES. Nos preguntamos si no es posible derivar, de algún modo formalmente preciso, las preferencias sobre las alternativas a partir de los criterios de evaluación de los argumentos que soportan a las alternativas. Si esto es posible, entonces se podrá encontrar un modo claro de determinar cómo las preferencias de los individuos pueden ser deliberativamente modificadas. O sea, si los individuos son capaces de evaluar los argumentos de acuerdo a criterios distintos, la deliberación acerca de los valores implícitos en sus criterios podrá modificar las preferencias sobre las alternativas al modificar las preferencias sobre los argumentos. A continuación proponemos un modo de hacerlo.

En primer lugar, veamos un modo concreto de elegir los "mejores" argumentos a partir de una relación de ataque. Las nociones de 'aceptabilidad' y de 'extensión fundada' (*grounded*) de Dung (1995) nos proporcionan una herramienta para ello:[1]

Aceptabilidad: un argumento *a* es aceptable con respecto a un conjunto de argumentos *S* si y solo si, para todo argumento *b* tal que $b \rightarrow a$, existe un argumento *c* en *S* tal que $c \rightarrow b$.

En otras palabras, es aceptable con respecto a *S* cualquier argumento que pueda "defenderse" con *S*. Esta definición permite construir la "función característica" del modelo de marcos argumentativos de Dung:

Función característica: $F(S)=\{a: a$ es aceptable con respecto a $S\}$

[1] Por supuesto, el criterio de elección dado por estas nociones puede ser reemplazado por otros (por ej., extensiones preferidas, estables, etc.). De todos modos, escogemos ese por su simpleza teniendo en cuenta el carácter exploratorio de este trabajo.

O sea, esta función opera sobre un conjunto de argumentos *S* obteniendo el conjunto de todos los argumentos que son aceptables con respecto a ese conjunto *S.* Veamos cómo obtenemos el conjunto de argumentos elegidos usando esta función. Supongamos que el marco argumentativo sobre el que vamos a elegir es el que representa el criterio de nuestro individuo 1:

$a \rightarrow_1 b \rightarrow_1 c$

Comencemos por aplicar la función *F* sobre el conjunto vacío. Esto nos dará el conjunto de todos los argumentos que son aceptables con respecto al conjunto vacío, los cuales pueden ser únicamente aquellos que no tienen atacantes (ya que no necesitan "defensores"), en este caso, solo *a*:

$F(\varnothing) = \{a\}$

Luego en una segunda aplicación de la función, es decir, aplicándola sobre el resultado de la aplicación anterior, obtendremos todos los argumentos que o bien no necesitan defensa o bien son defendidos por *a*:

$F^2(\varnothing) = F(\{a\}) = \{a, c\}$

Si ahora aplicamos la función sobre el último resultado obtendremos:

$F^3(\varnothing) = F(\{a, c\}) = \{a, c\}$

En este momento hemos llegado a un punto fijo de la función. Pero más que eso, es el menor punto fijo de *F*.[2] Esto es lo que define a la *extensión fundada*, que resulta en el (único) conjunto de argumentos elegidos.

La secuencia de niveles en los que se va aplicando la función *F* nos da una idea de cómo las alternativas pueden ordenarse de acuerdo a un criterio de defendibilidad de los argumentos que las soportan. En el ejemplo que venimos viendo concerniente a la opinión del individuo 1, las alternativas resultarían ordenadas del siguiente modo:

- la terapia *p* es al menos tan preferida como la terapia *r* porque en todos los niveles en los que aparece el argumento *c* que soporta a *r*, aparece también el argumento *a* que soporta a *p*.
- las terapias *p* y *r* son estrictamente preferidas a *q* porque el argumento *b* que soporta a *q* no aparece en ningún nivel.

Siguiendo esta idea, definimos formalmente las preferencias sobre las alternativas del siguiente modo:

- $x \succeq y$ si y solo si $\forall B \ \{\text{Sop}(B,y) \Rightarrow \forall i \ [B \in F^i(\varnothing) \Rightarrow \exists A \ (\text{Sop}(A,x) \ \& \ A \in F^i(\varnothing))]\}$
- $x \approx y$ si y solo si $x \succeq y \ \& \ y \succeq x$
- $x \succ y$ si y solo si $x \succeq y \ \& \ \neg \ x \approx y$

[2] Un punto fijo es cualquier elemento *S* tal que $F(S) = S$. Tecnicismo: la función *F* es creciente, por lo cual la clase de todos los puntos fijos forma un lattice, garantizando la existencia de un elemento mínimo.

Es decir, una alternativa *x* es al menos tan preferida como una alternativa *y* si y solo si, para todo argumento que soporta a *y* y para todo nivel *i* en el que aparezcan esos argumentos, existe algún argumento que soporta a *x* que aparece en todos esos niveles *i.* De acuerdo a esta definición, las preferencias de cada individuo sobre las alternativas quedarían así (indicamos con un subíndice en '$\succ$' el individuo en cuestión):

$p \succ_1 r \succ_1 q$
$r \succ_2 p \succ_2 q$
$q \succ_3 p, q \succ_3 r$

Veamos ahora qué ocurre si agregamos estas preferencias con el mismo mecanismo que aplicamos en el otro procedimiento, es decir, mayoría absoluta. Las preferencias grupales sobre las alternativas estarían dadas por:

$p \succ q, r \succ q$

Nótese que, en este caso, la opinión grupal no coincide con la del individuo 3: ahora las alternativas ganadoras son *p* y *r*, mientras *q* resulta descartada.

4.

La diferencia de resultados en los dos procedimientos vistos remite a la cuestión acerca de si hay ciertas condiciones razonables que harían que ambos coincidan. En Bodanza y Auday (2009) se investigó esto con cierta profundidad, hallando que la coincidencia puede darse solo bajo condiciones muy estrictas (por ejemplo, si no hay más de tres argumentos o no hay más de tres individuos, y si hay acuerdo unánime sobre alguno de los argumentos, etc.). Dada la imposibilidad general de la coincidencia, discutiremos un poco la conveniencia de seguir uno u otro procedimiento.

Un punto clave aquí es que no solo se trata de hallar un procedimiento justo (por ejemplo, en el sentido de cumplir con los conocidos postulados de Arrow en TES –cf. Arrow [1963]) sino también de procurar que la deliberación sea incorporada al modelo. Pero la deliberación, cuya finalidad será encontrar el mejor fundamento para la elección, podría dar lugar, como hemos visto, a dos procedimientos de elección distintos. En nuestra opinión, deliberar sobre los valores utilizados en la ponderación de los argumentos que soportan las distintas alternativas resulta apropiado. Pero una vez culminada la deliberación, no nos parece apropiado que la elección se realice agregando los criterios de ataque entre argumentos. Una vez que los individuos han revisado sus opiniones y reformulado sus preferencias sobre las alternativas, estas preferencias pueden agregarse del modo corrientemente considerado por la TES.

Esta opinión marca un cambio de parecer respecto del enfoque anteriormente seguido en el trabajo en colaboración con Tohmé y Simari (Tohmé *et al.*, 2008). La razón que hemos encontrado aquí es que el modo de fundamentar las preferencias sobre las alternativas en base al nivel de defendibilidad de los argumentos que las soportan, sea correcto o no, sugiere, al menos, la posibilidad de encontrar un fundamento racional para las preferencias individuales. Esto marca una diferencia teórica clave con la TES,

donde las preferencias individuales se suponen dadas y son inmodificables dentro de la teoría. El modelo al que tiende el presente trabajo permitiría distinguir dos planos:

1. *Plano deliberativo*: se ponderan los argumentos a favor y en contra de las alternativas de acuerdo a las valoraciones expresadas en los ataques de unos argumentos a otros. Los valores manifestados pueden ser modificados en este plano, y con ello podrán verse modificadas las preferencias de algunos individuos respecto de las alternativas.
2. *Plano definitorio*: en este plano ya no hay lugar para la deliberación. Las preferencias de los individuos ya han sido fijadas y solo queda la agregación para determinar cual es la preferencia social.

Confundir estos dos planos puede llevar a resultados de apariencia paradójica como el que hemos visto. Por un lado, agregar valoraciones (en el modelo, relaciones de ataque) supone una uniformidad motivacional holística acerca de las alternativas que resulta implausible. Por otro, deliberar sobre qué alternativa es preferible sin cuestionar las valoraciones subyacentes en los argumentos que las soportan se reduciría a no más que una compulsa de egos. Sin embargo, queda claro que, según el modelo, el plano deliberativo determina la configuración del plano definitorio: la deliberación puede modificar la valoración sobre los argumentos, que puede modificar las preferencias individuales sobre las alternativas, que pueden modificar las preferencias sociales.

No hemos mencionado cómo modelar la deliberación. Existen en la literatura modelos de argumentación basada en valores (Bench-Capon, 2003a, 2003b) que se aplicarían perfectamente aquí. La idea es que distintas audiencias (cf. Perelman y Olbrecht-Tyteca, 1989) privilegian distintos valores que los argumentos promueven, otorgándoles distintas fuerzas. Favorece nuestro proyecto el hecho de que el modelo desarrollado por Bench-Capon, como el propuesto, se basa en una extensión de los marcos argumentativos de Dung.

Bibliografía

Bench-Capon, Trevor J.M. (2003a), "Try to see it my way: modeling persuasion in legal discourse", en: *Artificial Intelligence and Law* 11. 4, pp. 271-287.

Bench-Capon, Trevor J.M. (2003b) "Persuasion in practical argument using value-based argumentation frameworks", en: *Journal of Logic and Computation* 13. 3, pp. 429-448.

Bodanza, Gustavo & Auday, Marcelo (2009), "Social argument justification: some mechanisms and conditions for their coincidence", en: *Proc. ECSQARU 2009*, Verona, Italia, pp. 95–106.

Dung, Phan M. (1995), "On the acceptability of arguments and its fundamental role in non-monotonic reasoning, logic programming and n-person games", en: *Artificial Intelligence* 77. 2, pp. 321-358.

Kaci, Souhila, van der Torre, Leendert & Weydert, Emil, (2006), "Acyclic argumentation: Attack=conflict+preference", en: *Proceedings of the 17th European Conference on Artificial Intelligence (ECAI'06)*, 2006, pp. 725-726.

Perelman, Chaim & Olbrechts-Tyteca, Lucie (1989), *Tratado de la argumentación. La nueva retórica,* Madrid, Gredos.

Simari, Guillermo R. & Loui, Ronald P. (1992), "A mathematical treatment of defeasible reasoning and its implementation", en: *Artificial Intelligence* 53. 2-3, pp.125-157.

Tohmé, Fernando, Bodanza, Gustavo & Simari, Guillermo R. (2008), "Aggregation of attack relations: a social-choice theoretical analysis of defeasibility criteria", en: *Proc. FoIKS '08*, Pisa, Italia, pp. 8–23.