EPISTEMOLOGÍA E HISTORIA DE LA CIENCIA

SELECCIÓN DE TRABAJOS DE LAS XVIII JORNADAS

VOLUMEN 14 (2008)

Horacio Faas
Hernán Severgnini

Editores

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

Los postulados de racionalidad de Arrow y la argumentación social

Gustavo A. Bodanza[*]

Los individuos suelen sopesar los mismos argumentos con distintos criterios, por lo que las conclusiones finales de cada uno pueden variar. Este trabajo plantea la posibilidad de agregar distintos criterios individuales de derrota entre argumentos, de modo de hallar un criterio global que refleje el pensamiento de la sociedad. Una relación de derrota no puede reducirse a otra de preferencia entre argumentos, lo que pone en duda la aplicabilidad de los resultados conocidos en teoría de la elección social al problema de la agregación de criterios de derrota. Aquí mostraremos, sin embargo, que si los individuos establecen preferencias sobre los argumentos y, en base a éstas, luego determinan las derrotas, entonces las propiedades de Arrow se mantendrán en su agregación.

Motivación

La argumentación rebatible consiste en la construcción de argumentos a fin de aceptar o rechazar una tesis determinada; tales argumentos son aceptables *prima facie*, pero su status final es determinado teniendo en cuenta las derrotas que se producen entre todos los argumentos considerados. Solo aquellos cuya defensibilidad queda probada son los juzgados como "justificados".

La relación de derrota entre argumentos ha sido usualmente considerada como la conjunción de dos relaciones previas: una de *desacuerdo* y otra de *preferencia* entre argumentos. La relación de desacuerdo se verifica entre dos argumentos si éstos no se consideran justificables a la vez (por lo tanto es simétrica). La relación de preferencia, por su parte, indica que un argumento es mejor que el otro (en algún sentido determinado), lo que no excluye *per se* que ambos sean justificables a la vez. Veamos los siguientes ejemplos:

A. El fumar me produce placer y yo debo procurar hacer todo lo que me produce placer, por lo tanto, debo seguir fumando.

B: El fumar me produce daño en los pulmones y yo debo evitar hacer todo lo que me produce daño, por lo tanto, debo dejar de fumar.

C. El fumar me deja olor desagradable en la ropa y el pelo y yo debo evitar hacer cosas que tienen efectos que me desagradan, por lo tanto, debo dejar de fumar.

El argumento *A* está en desacuerdo con los argumentos *B* y *C* puesto que sus conclusiones son contradictorias entre sí y no resulta razonable sancionar como justificables a la vez a *A* y *B* o a *A* y *C*. Por otra parte, el argumento *B* puede considerarse (según un criterio determinado) preferido al argumento *C*, aunque bien podrían considerarse justificables ambos a la vez. Supongamos que yo considero que entre mis reglas de conducta deben tener prioridad aquellas que preservan mi salud por sobre las demás. Entonces consideraría al argumento *B* preferible al argumento *A* y al argumento *C*, y quizá me mostraría indiferente entre *A* y *C* (i.e., igualmente preferidos). Según este criterio y los descuerdos establecidos, tendremos que el argumento *B*

[*] Centro de Investigaciones de Lógica y Filosofía de la Ciencia, Universidad Nacional del Sur, CONICET

derrota al argumento A, mientras A y C se derrotan mutuamente; por otra parte, no hay derrota entre B y C puesto que no están en desacuerdo Supongamos en cambio que yo considero que respecto de mis conductas es prioritario el placer. Entonces consideraría al argumento A preferible al argumento B, y podría dirimir la preferencia entre A y C según una gradación entre el palcer/displacer que me produzcan el fumar, por un lado, y el olor en la ropa y el pelo, por otro; supongamos que del resultado de esto determino que A es preferido a C. En este caso tendré que A derrota a B y a C, mientras no hay derrota entre B y C (sea cual fuere mi preferencia entre éstos).

Supongamos ahora que se plantea el problema de tomar una decisión colectiva que sea lo más justa posible respecto de las preferencias de los individuos, tal como, por ejemplo, prohibir o permitir fumar en un espacio cerrado compartido. Supongamos además que se realiza un debate entre los individuos que comparten ese espacio, en el que se plantean los distintos argumentos y se discuten las derrotas entre los mismos. La cuestión que trataremos de responder aquí es si la agregación de los distintos criterios de *derrota* individuales se puede reducir a la agregación de las *preferencias* individuales respecto de los argumentos.

La cuestión sería trivial si una relación de derrota tuviera exactamente las mismas propiedades que una relación de preferencia. Sin embargo no es así, y podemos ver esto a través de nuestro ejemplo. Según el primer criterio, puesto que B es preferido a A y A es (al menos tan) preferido a (como) C, tendremos que B es preferido a C, dado que las relaciones de preferencia son siempre transitivas. Pero, como vimos, B derrota a A y A derrota a C, pero B *no* derrota a C, lo que muestra que la relación de derrota *no* es transitiva. Luego, no es claro que al agregar criterios de derrota individuales obtendremos los mismos resultados que al agregar los criterios de preferencia individuales sobre los que se construyen tales derrotas

En su famosa tesis doctoral, K. Arrow (1963) postuló una serie de requisitos que toda agregación de preferencias individuales debería cumplir para ser considerada racional, y demostró que no hay ninguna función de agregación posible que de lugar a una preferencia social transitiva y completa, y que cumpla con todos aquellos requisitos a la vez. Considerando que tales requisitos son también razonables para la agregación de criterios de derrota, en este trabajo mostraremos que si la agregación de preferencias individuales cumple las propiedades arrovianas entonces la agregación de los criterios de derrota resultantes de las primeras también las cumplirán. Sin embargo, puesto que las relaciones de derrota no son (en general) transitivas y completas, el resultado de imposibilidad no se mantiene.

Nociones preliminares: proto-marcos argumentativos

Para nuestros fines introducimos la noción de *proto-marco argumentativo*, al que definimos como un par $PMA=\langle AR, \otimes\rangle$, donde AR es un conjunto de argumentos, y $\otimes$ es una relación binaria simétrica entre elementos de AR, que representa el *desacuerdo* entre dos argumentos. Un proto-marco argumentativo representa una situación en la que se debate esgrimiendo los argumentos en AR, y a partir de la cual hay que decidir cuáles serán, entre los que están en desacuerdo, los argumentos justificados o ganadores. Para determinar tal justificación pueden tenerse en cuenta uno o varios criterios de preferencia entre los argumentos, siendo de nuestro interés este último caso.

Criterios de preferencia abstractos y derrotas

Análogamente al enfoque usual de las preferencias en la economía, consideramos las preferencias entre argumentos como órdenes débiles (reflexivos, transitivos y completos) arbitrarios. Esto nos permitirá investigar las posibilidades formales de agregación de las preferencias sin necesidad de considerar su adecuación material. Partiendo de que un proto-marco argumentativo nos dice qué argumentos están en desacuerdo entre sí, la función de cada preferencia individual será dar un veredicto acerca de cada par de argumentos en desacuerdo: uno derrota al otro, el segundo derrota al primero, o ambas cosas a la vez (esta última posibilidad no habilita a *justificar* ambos argumentos a la vez, en virtud de estar en desacuerdo). Para nuestros fines tomaremos un conjunto $N = \{1, 2, \ldots, n\}$ de individuos, interesándonos solo los casos en que $n \geq 2$, y a cada uno corresponderá una relación de preferencia individual $\succeq_i \subseteq AR \times AR$, las que asumimos como órdenes débiles. La derrota de un argumento A por otro argumento B relativa a un individuo $i \in N$ se define como sigue:

Definición 1

Dados dos argumentos $A, B \in AR$, decimos que *A derrota a B según el individuo i*, en símbolos, '$A \rightarrow_i B$', sssi $A \otimes B$ y $A \succeq_i B$.

De este modo, para cada individuo i queda conformado un *marco argumentativo* $AF_i = \langle AR, \rightarrow_i \rangle$, tal como define la noción Dung (1995).

Puesto que cada $\succeq_i$ es completa, de la definición anterior se sigue que si $A \otimes B$ entonces para todo individuo i se cumple $A \rightarrow_i B$ o $B \rightarrow_i A$ (o ambas). Entonces, cuando dos argumentos que están en desacuerdo resultan igualmente preferidos o indiferentes bajo una preferencia individual, tendremos que, según tal preferencia, esos argumentos se derrotan mutuamente.

Es razonable esperar que la relación de derrota global o agregada esté vinculada con una relación de preferencia global, agregada, del mismo modo que ocurre entre las preferencias y derrotas individuales. Si además la preferencia global es también completa, tendremos que si $A \otimes B$ entonces A derrota globalmente a B o viceversa, o ambas. También es de esperar que si dos argumentos *no* están en desacuerdo, entonces *no* debe resultar globalmente determinada una derrota entre ellos. Estas consideraciones nos llevan a la siguiente definición:

Definición 2

Sea $\succeq^k$ una relación de *preferencia global*, resultante de la agregación de las preferencias individuales $\succeq_i$ según un criterio de agregación k determinado (*e.g.*, por mayoría). Asumiremos que $\succeq^k$ es un orden débil. Dados dos argumentos A y B, decimos que *A derrota globalmente a B* según la preferencia global $\succeq^k$, en símbolos, '$A \rightarrow^k B$', sssi $A \otimes B$ y $A \succeq^k B$.

A partir de esta definición obtenemos un marco argumentativo global o social que simbolizamos $AF^k = \langle AR, \rightarrow^k \rangle$.

Observación

De ahora en más nos referiremos a relaciones de preferencia y derrota globales obtenidas por un criterio arbitrario pero fijo (*e.g.*, voto mayoritario), por lo que omitiremos el superíndice 'k' para simplificar la notación.

Condiciones arrovianas de agregación de preferencias

Las condiciones que Arrow consideró para la agregación de preferencias pueden enunciarse del siguiente modo:

P1. (Pareto) $\forall i \in N\ A \geq_i B \Rightarrow A \geq B$. Esta condición dice que si todos los individuos coinciden en que A es preferido a B, entonces A debe resultar preferido globalmente a B.

P2. (Asociación positiva de valores individuales) Para cualesquiera dos perfiles de preferencias individuales $(\geq_1, \ldots, \geq_n)$ y $(\geq'_1, \ldots, \geq'_n)$, si $\{i: A \geq_i B\} \subseteq \{i: A \geq'_i B\}$ y $A \geq B$, entonces $A \geq' B$. Es decir, si en un perfil determinado $(\geq_1, \ldots, \geq_n)$, el argumento A resulta globalmente preferido a B, entonces debe obtenerse el mismo resultado global en caso de que algunos individuos que previamente disentían respecto de ese par cambien su opinion hacia un acuerdo con ese resultado (sin cambiar sus preferencias respecto del resto de los pares de argumentos).

P3. (Soberanía) $\geq$ no debe ser impuesta, i.e., para cada par de argumentos A y B, existe un perfil $(\geq_1, \ldots, \geq_n)$ tal que no $A \geq B$. Es decir, una preferencia global no puede resultar igual para todo ordenamiento que hagan los individuos.

P4. (Independencia de las alternativas irrelevantes) Sean $\geq_1, \ldots, \geq_n$ y $\geq'_1, \ldots, \geq'_1$ dos perfiles de preferencias y sean $\geq$ y $\geq'$ las correspondientes preferencias globales resultantes de sus agregaciones. Si para todo agente i, $\geq_i|_{\{A,B\}} = \geq'_i|_{\{A,B\}}$ para cualquier subconjunto $\{A,B\} \subseteq AR$, entonces $\geq|_{\{A,B\}} = \geq'|_{\{A,B\}}$. O sea, si dos perfiles de preferencias individuales son iguales respecto de un par de argumentos determinado, entonces las preferencias globales resultantes de ambos perfiles no pueden variar respecto de ese par de argumentos (variaciones en las ordenaciones de otros pares no deben afectar al ordenamiento global de ese par).

P5. (No dictadura) No existe i tal que para todo perfil de preferencias individuales $(\geq_1, \ldots, \geq_n)$, si $A \geq_i B$ entonces $A \geq B$. Las preferencias de un individuo no pueden imponerse para cualquier variación en las preferencias de los otros individuos.

Estos requisitos que se plantean para la agregación de preferencias individuales también resultan razonables para la agregación de derrotas individuales. Vamos a plantearlas una a una para verlo mejor.

D1. (Pareto) $\forall i \in N\ (A \rightarrow_i B) \Rightarrow A \rightarrow B$. Si todos los individuos están de acuerdo en que A derrota a B (aunque sea por distintas razones), entonces A debe derrotar globalmente a B.

D2. (Asociación positiva de valores individuales) Para cualesquiera dos perfiles de derrotas individuales $(\rightarrow_1, \ldots, \rightarrow_n)$ y $(\rightarrow'_1, \ldots, \rightarrow'_n)$, si $\{i: A \rightarrow_i B\} \subseteq \{i: A \rightarrow'_i B\}$ y $A \rightarrow B$, entonces $A \rightarrow' B$. Es obvio que un mayor consenso en cuanto a la misma derrota global establecida entre dos argumentos no puede hacer que dicha derrota cambie lo previamente consensuado.

D3. (Soberanía) $\rightarrow$ no debe ser impuesta, i.e., para cada par de argumentos A y B, existe un perfil $(\rightarrow_1, \ldots, \rightarrow_n)$ tal que no $A \rightarrow B$. Los individuos deber ser capaces de modificar las derrotas resultantes al modificar sus propios criterios.

D4. (Independencia de las alternativas irrelevantes) Sean $\rightarrow_1, \ldots, \rightarrow_n$ y $\rightarrow'_1, \ldots, \rightarrow'_1$ dos perfiles de derrotas y sean $\rightarrow$ y $\rightarrow'$ las correspondientes derrotas resultantes de sus agregaciones. Si para todo agente i, $\rightarrow_i|_{\{A,B\}} = \rightarrow'_i|_{\{A,B\}}$ para cualquier subconjunto $\{A,B\} \subseteq AR$, entonces $\rightarrow|_{\{A,B\}} = \rightarrow'|_{\{A,B\}}$. La derrota global resultante sobre un par de argumentos A y B no puede variar a

menos que algún individuo modifique su criterio respecto de ese mismo par (esto no implica, por supuesto, que no se modifique el conjunto de argumentos "justificados").

D5. (No dictadura) No existe i tal que para todo perfil de derrotas individuales ($\rightarrow_1$, ..., $\rightarrow_n$), si $A \rightarrow_i B$ entonces $A \rightarrow B$. Las derrotas globales no puede decidirlas un único individuo en todos los casos.

Observando que las propiedades arrovianas son razonables para la agregación de criterios de derrota individuales, nos preguntamos si el hecho de que se cumplan P1-P5 es suficiente para que se cumplan también D1-D5. El siguiente resultado da una respuesta afirmativa.

Teorema

Si la preferencia global $\geq$ cumple las condiciones de Pareto, asociación positiva de valores individuales, soberanía y no dictadura (P1-P5), esta última restringida al dominio de los pares de argumentos en desacuerdo, entonces la derrota global $\rightarrow$ también las cumple (i.e. cumple D1-D5).

Prueba.

- Pareto. Supongamos que si $A \geq_i B$ para todo i, entonces $A \geq B$. Supongamos además que para todo i, $A \rightarrow_i B$. De esto último y la definición 1 se sigue que $A \otimes B$ y $A \geq_i B$ para todo i. Luego, de la hipótesis se sigue que $A \geq B$; por lo tanto, por la definición 2, tenemos que $A \rightarrow B$.

- Asociación positiva de valores individuales. Supongamos que (i) para cualesquiera dos perfiles de preferencias individuales ($\geq_1$, ..., $\geq_n$) y ($\geq'_1$, ..., $\geq'_n$), si $\{i: A \geq_i B\} \subseteq \{i: A \geq'_i B\}$ y $A \geq B$, entonces $A \geq' B$. Supongamos ahora que (ii) $\{i: A \rightarrow_i B\} \subseteq \{i: A \rightarrow'_i B\}$ y (iii) $A \rightarrow B$. Por la definición 1 sabemos que si $A \rightarrow_i B$ entonces $A \geq_i B$ y, del mismo modo, si $A \rightarrow'_i B$ entonces $A \geq'_i B$. Luego, de la hipótesis (ii) se sigue que (1) $\{i: A \geq_i B\} \subseteq \{i: A \geq'_i B\}$. Por otra parte, de (iii) y la definición 2 se sigue que (2) $A \otimes B$ y $A \geq B$. Entonces, de (1), (1) y (2) tenemos que $A \geq' B$, que junto con la definición 2 implica que $A \rightarrow' B$.

- Independencia de las alternativas irrelevantes. Sean $\geq_1$, ..., $\geq_n$ y $\geq'_1$, ..., $\geq'_n$ dos perfiles de preferencias individuales y sean $\geq$ y $\geq'$ las correspondientes preferencias globales resultantes de sus agregaciones, y supongamos que (i) para todo individuo i se cumple que si $\geq_i|_{\{A,B\}} = \geq'_i|_{\{A,B\}}$ para cualquier subconjunto $\{A,B\} \subseteq AR$, entonces $\geq|_{\{A,B\}} = \geq'|_{\{A,B\}}$. Supongamos ahora que $\rightarrow_i|_{\{A,B\}} = \rightarrow'_i|_{\{A,B\}}$. Entonces por la definición 1 se cumple $\geq_i|_{\{A,B\}} = \geq'_i|_{\{A,B\}}$ y $A \otimes B$. De esto y (i) obtenemos $\geq|_{\{A,B\}} = \geq'|_{\{A,B\}}$ y $A \otimes B$. Luego, por la definición 2 llegamos a $\rightarrow|_{\{A,B\}} = \rightarrow'|_{\{A,B\}}$.

- Soberanía. Supongamos que para cada par de argumentos A y B, existe un perfil ($\geq_1$, ..., $\geq_n$) tal que no $A \geq B$. Por el absurdo, supongamos que $A \rightarrow B$ para todo perfil de derrotas ($\rightarrow_1$, ..., $\rightarrow_n$). De la definición 2 se sigue que $A \geq B$, cualquiera sea el perfil de preferencias individuales que produzca el perfil de derrotas ($\rightarrow_1$, ..., $\rightarrow_n$). Contradicción.

- No dictadura. Supongamos que para todo individuo i hay un par de argumentos A y B tales que $A \geq_i B$ pero no $A \geq B$ para algún perfil ($\geq_1$, ..., $\geq n$). Suponiendo que $A \otimes B$, es claro que para todo individuo i, $A \rightarrow_i B$ pero no $A \rightarrow B$. ♂

Conclusión

Este resultado constituye un primer paso en la búsqueda las condiciones de racionalidad de agregación de criterios de derrota. Los requisitos arrovianos se muestran razonables y

plausiblemente cumplibles en vista de que no se requiere de una relación de derrota que sea completa y transitiva (esto último es más bien indeseable); la relajación de estas dos propiedades esquiva las contradicciones que llevan al teorema de imposibilidad de Arrow.

El paso a seguir consistirá en buscar condiciones de racionalidad propias de la elección social de argumentos. Por ejemplo, si un individuo establece las derrotas $A \rightarrow_i B \rightarrow_i C$, entonces es de esperar que los argumentos de su elección sean A y C (el primero porque no recibe derrotas, y el segundo por ser defendido por el primero); la misma elección es de esperar que ocurra para una derrota global equivalente. Pero la forma de elegir los argumentos socialmente "defendibles" puede seguir dos mecanismos distintos: 1) primero agregar las derrotas individuales para obtener la derrota global, y luego en base a ésta decidir la elección; o bien 2) establecer las elecciones individuales en base a las derrotas individuales, y luego agregar todas las elecciones individuales. Una pregunta que intentaremos responder es si es posible hallar dos mecanismos tales que sean conmutables.

La agregación de criterios de derrota ha empezado a investigarse en el campo de la argumentación rebatible. Un caso destacado es el de Coste-Marquis *et al.* (2005) Estos autores proponen un mecanismo específico de agregación, pero no se preguntan por condiciones generales más que el requisito de que los argumentos socialmente elegidos deben formar un conjunto libre de conflictos (i.e. sin derrotas internas). Nuestro trabajo, en cambio, apunta a hallar criterios generales que debería cumplir *cualquier* mecanismo específico de agregación, para luego ver si es lógicamente posible la existencia de un mecanismo que cumpla con todos ellos a la vez.

Referencias

Arrow, K. Social Choice and Individual Values, Wiley, NY, 1963

Coste-Marquis, S. Devred, C. Konieczny, S. Lagasquie-Schiex, M-C. Marquis, P. "Merging Argumentation Systems", *Proceedings of of 20th National Conference on Artificial Intelligence (AAAI'05)*, 614-619, 2005

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