Revista Iberoamericana de Argumentación
ἐπει δὲ ταύτην ἐπιστήμην ζητοῦμεν

Segunda Época
RIA 23 (2021): 63-83

Directores: Luis Vega y Hubert Marraud **Editora:** Paula Olmos
ISSN 2172-8801 / http://doi.org/10.15366/ria2021.23 / https://revistas.uam.es/ria

Interacción de argumentos y valores. Puentes entre la Inteligencia Artificial y la Psicología del Razonamiento

Interaction of arguments and values. Bridges between Artificial Intelligence and the Psychology of Reasoning

Gustavo Adrián Bodanza
Departamento de Humanidades / Instituto de Investigaciones Económicas y Sociales del Sur
Universidad Nacional del Sur / UNS-CONICET – Argentina
bodanza@gmail.com

Artículo recibido: 16-03-2021
Artículo aceptado: 15-09-2021

RESUMEN
Los modelos de argumentación propuestos desde la Inteligencia Artificial ofrecen simplicidad y precisión para analizar la aceptabilidad de un argumento en interacción con otros. Sin embargo, se presentan dudas a la hora de ponderar su corrección, ya que el carácter, más dialéctico que lógico, de la argumentación impide contar con una semántica formal con la cual relacionarla. Aquí comentaremos los modelos de argumentación basada en valores de Gabbay y de Bench-Capon. Gabbay, por caso, busca implementar la intuición de que enfrentar argumentos que promueven un mismo valor (religioso, político, jurídico, etc.) es más efectivo que hacerlo desde un valor distinto no compartido. Valiéndome de algunos ejemplos tomados de la literatura, mostraré la importancia de tender puentes entre los modelos y lo empírico que permitan contrastar dicha intuición. Argumentaré que hay problemas tanto conceptuales como representacionales que es necesario atacar, y señalaré algunas líneas de investigación experimental en tales direcciones.

PALABRAS CLAVE: argumentación abstracta, argumentación basada en valores, Inteligencia Artificial, modelos formales de argumentación, psicología del razonamiento.

ABSTRACT
The argumentation models proposed from Artificial Intelligence offer simplicity and precision to analyze the acceptability of an argument in interaction with others. However, there are doubts when considering their correctness, since the character, more dialectical than logical, of the argumentation prevents having a formal semantics with which to relate it. Here we will discuss the value-based argumentation models by Gabbay and Bench-Capon. Gabbay, for instance, seeks to implement the intuition that confronting arguments that promote the same value (religious, political, legal, etc.) is more effective than doing it from a different, unshared value. Using some examples taken from the literature, I will show the importance of building bridges between the models and the empirical that enable to contrast such intuition. I will argue that there are both conceptual and representational problems that need to be addressed, and I will point out some lines of experimental research in these directions.

KEYWORDS: abstract argumentation, value-based argumentation, Artificial Intellgence, formal models of argumentation, psychology of reasoning.

Servicio de Publicaciones de la Universidad Autónoma de Madrid

The tension between normative and descriptive considerations characterizes much of the study of judgment and choice.
D. Kahneman y A. Tversky (2000)

1. INTRODUCCIÓN

Dirimir una disputa argumentando desde distintas posiciones valorativas unas veces puede resultar efectivo y otras puede resultar estéril. Por ejemplo, es común ver en el ámbito de la justicia que un caso bien fundamentado en la evidencia se pierde por argumentos que señalan vicios de procedimiento. El recurso suele ser efectivo porque las normas procesales conllevan un valor superior al de la prueba. También observamos el problema en interminables debates acerca de la legalidad del aborto, con argumentos favorables desde los marcos de valores de los derechos de la mujer y la salud colectiva, y contrarios desde los marcos de la religión y la vida humana de los embriones. En este caso, las discusiones suelen resultar estériles puesto que no hay acuerdos sobre la superioridad de un marco de valores sobre otro. En vista de estos problemas, Dov Gabbay (2014) analiza algunas propuestas para resolver las disputas. Una se basa en la direccionalidad de los ataques, rastreando el camino que siguen los argumentos llevando los ataques de un marco de valores a otro. La solución consiste en determinar qué argumentos se imponen dentro de cada marco de valor y ver cómo éstos afectan a los argumentos del valor siguiente. Otra propuesta es considerar que los marcos de valores mismos interactúan de manera análoga a cómo lo hacen los argumentos, por lo que primero habría que dirimir qué valores se resguardan para luego elegir los mejores argumentos dentro de ellos. Estas propuestas se analizan a través del modelo de marcos argumentativos de Dung (1995), que ofrece un tratamiento abstracto de la interacción entre argumentos facilitando la comprensión de las soluciones.

Mi objetivo es señalar algunas dificultades que se suscitan a partir de la propuesta de Gabbay, para concluir que un tratamiento puramente formal del problema no parece suficiente. Vale decir que el enfoque al que nos referimos proviene de la Inteligencia Artificial, donde la impronta formal se debe en gran parte al interés por conocer la complejidad computacional de resolver una disputa. Para ello los modelos ofrecen simplicidad y precisión. Sin embargo, sostendré que los modelos pueden enriquecerse y devenir en teorías de la argumentación dialéctica mediante estudios empíricos que muestren el comportamiento de la gente real en el marco de una disputa argumentada. Me basaré en el análisis de algunos ejemplos y su tratamiento por parte

de los modelos de Gabbay (2014) y Bench-Capon (2002, 2003a). Por otro lado, voy a plantear algunos puntos concretos que deberían someterse a prueba, relacionados con los conceptos involucrados en los modelos y sus implicaciones para la representación de casos. Esto tiene el fin de señalar el camino para la puesta a prueba de hipótesis empíricas sobre cómo opera el sentido común de las personas en el marco de la argumentación. Esto, a su vez, brindaría la posibilidad de contrastar los modelos como teorías de la argumentación. Dado que el artículo está orientado a un público no especializado en sistemas argumentativos de Inteligencia Artificial, presentaré las ideas básicas de los modelos del modo más informal posible, teniendo en cuenta que no es necesario más para la comprensión del punto principal que me ocupa.

El trabajo se organiza como sigue. En la sección 2 comento el modelo de base, el de marcos argumentativos de Dung, para pasar, en la sección 3, a las ampliaciones de Bench-Capon y Gabbay que incorporan valores. En la sección 4 analizaré el comportamiento de los modelos en base a un ejemplo tomado de la literatura. En la sección 5 identifico dos tipos de problemas, conceptuales (a qué refieren concretamente los elementos del modelo y sus relaciones) y representacionales (derivados de los conceptuales). En la sección 6 argumento que una mirada empírica, básicamente a través de experimentos, permitirá capturar intuiciones para explicar la justificación humana de argumentos a la vez que enriquezcan los modelos con nuevas perspectivas. Ofreceré mis conclusiones en la sección 7.

2. EL MODELO DE MARCOS DE ARGUMENTACIÓN

Para comprender la propuesta de Gabbay es necesario introducirse en el modelo de argumentación que utiliza, el de marcos argumentativos de Dung (1995). Este modelo -extensamente utilizado en el campo de la Inteligencia Artificial- es altamente abstracto, ya que considera a los argumentos como entidades primitivas dadas, obviando su constitución y considerando solamente su interacción a través de ataques, que también se suponen dados. El fin es determinar qué argumentos logran imponerse en las confrontaciones. El modelo busca representar las intuiciones de que si un argumento *b* (p. ej.: Tweety vuela porque es un ave) es atacado por otro argumento *a* (p. ej.: Tweety no vuela porque es un pingüino) y no hay otros argumentos ni ataques, entonces *a* resulta aceptado (porque no tiene atacantes) y *b* rechazado; y si, a su vez, se introduce un tercer argumento *c* (p. ej.: No hay evidencia de que Tweety sea un pingüino, por lo que no se puede afirmar que lo sea) que ataca a *a*, entonces *c* debe resultar aceptado (puesto que no tiene atacantes), *a* rechazado (puesto que no puede contrarrestarse el

ataque de *c*), y *b* aceptado (puesto que resulta "defendido" exitosamente por *c*). Formalmente, un *marco argumentativo* es un par <*A*,*R*>, donde *A* es un conjunto de entidades primitivas llamadas 'argumentos' y *R* es una relación binaria sobre *A*, también primitiva, tal que (*x*, *y*) ∈ *R* es interpretado como *'x* ataca a *y*'. A los distintos criterios para determinar los argumentos aceptados o ganadores, se los conoce con el nombre de *semánticas de extensiones*. Éstas determinan, para un marco argumentativo dado, subconjuntos de argumentos que pueden ser aceptados a la vez –las *extensiones*. Si en un marco argumentativo tenemos el conjunto de argumentos {*a*, *b*, *c*} y los ataques (representados por flechas) *a*→*b*→*c*, entonces las semánticas de extensiones que capturan las intuiciones antes mencionadas determinarán la extensión {*a*, *c*} como el conjunto de argumentos ganadores. Esta idea es expresada por Dung a través de la noción de *admisibilidad.* Un conjunto *S*⊆*A* es *admisible* si, y sólo si, (i) para cualesquiera *x* e *y*, si (*x*, *y*)∈*R* entonces o bien x∉*S*, o bien y∉*S* (i.e., *S* está *libre de conflictos*), y (ii) para todo *x*, si (*x*, *y*)∈*R* para algún *y*∈*S*, entonces existe *z*∈*S tal que* (*z, x*)∈*R* (i.e., todo argumento de *S* es *aceptable* en con respecto a *S*). En ciertos casos puede darse que una semántica determine más de una extensión. A éstas se las llama semánticas *crédulas* –diferenciándolas de las *escépticas*, que determinan una única extensión. Si, por ejemplo, en un marco tenemos sólo dos argumentos *a* y *b* que se atacan mutuamente, entonces una semántica crédula podría determinar dos extensiones posibles, {*a*} y {*b*}, representando la intuición de que cualquiera de los dos argumentos puede ser aceptado indistintamente (pero no ambos). Esto es lo que ocurre con la *semántica preferida* (*preferred semantics*), que sanciona como extensiones conjuntos máximamente (con respecto a ⊆) admisibles. En cambio, una semántica escéptica podría determinar como única extensión ∅ (el conjunto vacío), representando la intuición de que no puede aceptarse ninguno de los argumentos, ya que no han superado concluyentemente los ataques recibidos. Por ejemplo, tomar la intersección de todas las extensiones de la semántica preferida nos daría una semántica escéptica.[1] Las semánticas de extensiones pueden cumplir distintas propiedades o condiciones de racionalidad (Baroni y Giacomin, 2007) como, por ejemplo, la de *restablecimiento*: si desde una extensión E se atacan a todos los atacantes de un argumento *x*, entonces *x* debe pertenecer a E.

[1] Otra semántica escéptica dada por Dung es la *grounded semantics*, que es el menor punto fijo de un operador que, aplicado a un conjunto de argumentos *S*, devuelve el conjunto de todos los argumentos aceptables con respecto a *S*.

3. LAS PROPUESTAS DE GABBAY Y BENCH-CAPON

Siguiendo a Gabbay (2014), argumentos provenientes de distintos marcos de valores deberían representarse en distintos marcos argumentativos, y los ataques que cruzan de un marco a otro darían lugar a una familia de marcos argumentativos interactuantes. Si no se tiene en cuenta esto y se combinan todos los argumentos en un mismo marco, entonces las semánticas de extensiones actuales difícilmente capturarán los argumentos mejor justificados. Gabbay ofrece algunos ejemplos que motivan el enfoque:

> [Un ejemplo] es cuando un sistema teológico ataca el argumento *x*, presentado por un sistema social secular. No es práctico decir "No acepto tu religión; no es relevante atacar las consideraciones seculares con un sistema de creencias fabricado". Es mucho más efectivo entrar en las consideraciones teológicas mismas y argumentar que hay un punto de vista teológico (una extensión) que no ataca a *x*. Igual y simétricamente, es inútil que el sistema religioso descarte las consideraciones sociales como irrelevantes porque van en contra de la voluntad de Dios. Es mucho mejor entrar en los argumentos sociales y adoptar una extensión religiosamente más tolerante. De hecho, tanto el sistema de argumentación religiosa como el sistema social secular pueden ser el resultado de la fusión de varios subsistemas de segundo nivel de facciones religiosas y partidos políticos, respectivamente. En tal caso, un bando necesita dirigirse a las facciones internas del otro bando con la esperanza de influir favorablemente en el resultado final de la fusión.
>
> Los políticos saben esto. Si son criticados por la Iglesia o por algunos economistas ganadores del premio Nobel, o por algunos expertos, ¡lo mejor es conseguir que otro experto u otro premio Nobel u otro obispo los apoye!
>
> Otro ejemplo es el del dominio legal. Supongamos que construimos un alegato contra un acusado. Llamemos a esto Marco 2. El caso contiene una pieza de evidencia p, obtenida mediante algunos procedimientos cuestionables. El Marco 1 son las consideraciones de la teoría jurídica de la prueba para comprobar si es posible admitir p. La acusación adoptará una extensión y la defensa adoptará una extensión diferente. (Gabbay, 2014: 132-133. Trad. mía)

De modo que Gabbay va a considerar distintos marcos según los valores que distinguen y clasifican a los argumentos.[2] En un primer enfoque, considera sólo los casos donde los valores pueden ser puestos en un orden de preferencia que no dé lugar a ciclos (estas preferencias podrían representar las de un auditorio particular (Perelman y Olbrechts-Tyteca, 1989)), y se mantiene abierto respecto de la semántica de extensiones utilizada en cada uno de los marcos, admitiendo incluso la simple selección de subconjuntos de argumentos libres de conflicto. El orden de preferencias acíclico sobre los valores da lugar a una semántica de contexto direccional, donde los argumentos se van eligiendo según el orden de los ataques entre argumentos de los distintos valores.

[2] La idea de considerar marcos argumentativos basados en valores fue desarrollada previamente por Bench-Capon (2002, 2003, etc.).

Formalmente, un *marco argumentativo direccional basado en valores* tiene la forma $(A, R, V, <, e)$, donde A y R son tales que (A, R) es un marco argumentativo, V es un conjunto de *valores*, $<$ es un orden acíclico sobre V, y e es una función que asigna a cada argumento de A un valor de V, de tal modo que si $(x, y) \in R$ entonces $e(y) \leqq e(x)$.[3] (Nótese que si hay ciclos de ataques entre argumentos de distintos valores entonces no podremos tener un marco argumentativo direccional).

Veamos un ejemplo abstracto. La representación de un marco argumentativo se puede hacer mediante un digrafo, i.e. un conjunto de nodos que representan a los argumentos y flechas que representan los ataques:

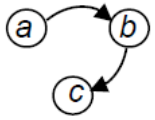

Figura1

Una semántica crédula basada simplemente en escoger conjuntos de argumentos máximamente libres de conflictos (digamos, puntos de vista coherentes) determinará las extensiones $\{a, c\}$ y $\{b\}$, permitiendo escoger cualquiera de ellas indistintamente. Ahora supongamos que clasificamos los argumentos de acuerdo a dos valores v_1 y v_2, de modo que a promueve el valor v_1 (i.e., $e(a)=v_1$) y b y c promueven el valor v_2 (i.e., $e(b)=e(c)=v_2$):

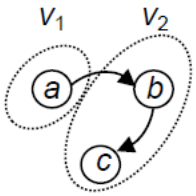

Figura 2

Supongamos además que v_1 es preferido a v_2 (i.e., $v_1 > v_2$). El enfoque direccional determina escoger primero una extensión en el submarco generado por v_1. Naturalmente, siguiendo el criterio de escoger conjuntos máximamente libres de conflicto, podemos escoger el conjunto que contiene al único argumento, $\{a\}$. Siguiendo la dirección de $>$, ahora eliminamos en v_2 todos los argumentos atacados por $\{a\}$ (o sea, eliminamos b). A continuación, elegimos la extensión del submarco generado por los argumentos sobrevivientes en v_2. Esto resulta en la elección del conjunto que contiene al único argumento sobreviviente, $\{c\}$. Finalmente, tomamos como extensión de todo el

[3] Definimos $\leqq$ y $>$ a partir de $<$ del modo obvio.

marco a la unión de todas las extensiones obtenidas en el procedimiento anterior, esto es, $\{a\} \cup \{c\} = \{a, c\}$.

En el caso de Bench-Capon, un *marco argumentativo basado en valores de auditorio específico* se define también como una tupla $(A, R, V, <, e)$, sólo que, a diferencia del enfoque direccional de Gabbay, $<$ representa las preferencias entre valores de un auditorio especifico. Bench-Capon define la noción de derrota entre argumentos del siguiente modo: *x derrota a y* si, y sólo si, x ataca a *y* y no es el caso que $e(y) > e(x)$. Supongamos que, en nuestro ejemplo anterior, para el auditorio en cuestión tenemos $v_2 > v_1$. Entonces el ataque de *a* pierde efectividad (pero no el de *b* sobre *c*, que se da dentro del mismo valor). La representación de esta situación es la siguiente:

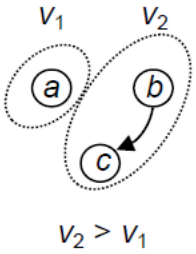

Figura 3

Pero esto no quiere decir que *a* deba ser rechazado; simplemente ya no presenta conflicto y, en consecuencia, la extensión será $\{a, b\}$.

El segundo enfoque de Gabbay es el de un haz de marcos argumentativos (*fibred argumentation networks*) en versión *jerárquica*. La intuición es que los valores se atacan entre sí de acuerdo a la relación de ataque entre sus argumentos. Es decir, pasamos a un marco de nivel superior o meta-marco. En este nivel, se determina una jerarquía entre los valores y, una vez hecho esto, se eligen los argumentos ganadores entre los que promueven los valores predominantes, de acuerdo a la semántica de extensiones elegida. Finalmente, la extensión del meta-marco es la unión de todas las extensiones de nivel inferior. En nuestro ejemplo tendremos el siguiente meta-marco:

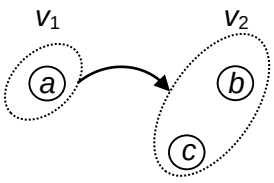

Figura 4

Tomemos por criterio de selección una semántica basada en admisibilidad (podría ser cualquier otra). Entonces v_1 es el valor escogido, y dentro del submarco determinado por este valor tomamos como extensión {*a*}. El valor v_2 resulta rechazado, por lo que ningún argumento se escoge en ese submarco. En consecuencia, la extensión de todo el meta-marco será simplemente {*a*}.

La pregunta de rigor es ¿cuál de los enfoques es correcto? Aquí no me ocuparé de responder esta pregunta en particular, sino de enfocar el problema de fondo: ¿Qué justificación (semántica) general daría una base sensata para escoger argumentos en la interacción de ataques y valores en juego? En primer lugar, es muy difícil responder a estas preguntas sin intuiciones claras que permitan comprender el comportamiento de una semántica con respecto a, al menos, ciertos ejemplos de referencia (*benchmark problems*). Esto es lo que han hecho los investigadores en Inteligencia Artificial devenidos en teóricos de la argumentación (ver, por ejemplo, Mailly y Maratea, 2019). En segundo lugar –y este punto aplica a cualquier teoría de la argumentación en general- no queda claro si el problema puede dirimirse en el plano puramente formal cuando la argumentación actúa en el terreno práctico. Esto ocurre particularmente cuando intervienen argumentos que promueven distintos valores. Soy consciente de que esta última observación amerita un inmenso debate, inabarcable dentro de los límites del presente trabajo. Sin embargo, sirve para marcar un aspecto de la argumentación que requiere enfoques metodológicos de base empírica. A continuación trataré estos puntos.

4. EL COMPORTAMIENTO DE LOS MODELOS

Para plantear mi argumento voy a analizar el comportamiento de los modelos de Gabbay y Bench-Capon tomando un ejemplo tratado por este último autor (introducido por Coleman (1992) y ampliado por Christie (2000)). Hal, un diabético, pierde su insulina en un accidente (ajeno a su responsabilidad). Antes de caer en coma corre hasta la casa de Carla, otra diabética, para pedirle insulina. Pero Carla no está en su casa. Hal, desesperado, decide entrar y usar su insulina. ¿Está jusificada la acción de Hal? ¿Carla tiene derecho a una compensación de parte de Hal? Los argumentos que surgen inmediatamente podrían clasificarse según el valor promovido: el de la vida, el del derecho a la propiedad privada. Veamos cómo modelar la situación. Primero, esbocemos algunos argumentos: *a*: La acción de Hal está justificada por una necesidad vital, *b*: Hal no debió tomar sin permiso la insulina. Infringió el derecho a la propiedad de Carla, *c*: Hal no infringe el derecho a la propiedad de Carla si asume compensarla, *d*:

Hal no tiene la obligación de compensar a Carla, porque si fuera demasiado pobre como para hacerlo igualmente estaría justificado en tomar la insulina para salvar su vida. Hasta aquí, los argumentos analizados por Bench-Capon (2002). La línea de ataques se muestra en la Figura 4, donde además se representan los valores promovidos por los argumentos: *a* y *d*, el valor de la vida; *b* y *c*, el valor del derecho a la propiedad privada.

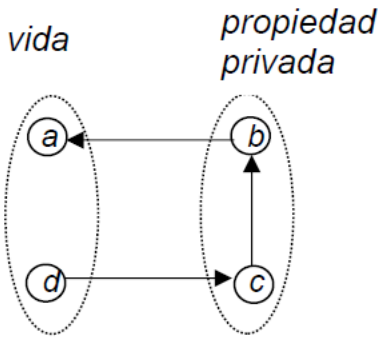

Figura 5

El efoque direccional puro de Gabbay no puede aplicarse, ya que tenemos un ciclo entre los valores. Pero si consideramos un orden entre los valores tal que, digamos, el valor de la vida es preferido al del derecho a la propiedad privada, entonces podemos aplicar el enfoque de haz de marcos argumentativos (*fibred argumentation frameworks*) en versión *direccional*, donde la dirección dada por la preferencia indica el orden de obtención de las extensiones. En ese caso elegimos una extensión en el marco de "vida" y luego pasamos a elegir una extensión en el marco de "propiedad privada". Entonces podríamos escoger {*a, d*} primero (ya que esos argumentos no están en conflicto entre sí, y luego podríamos elegir {*c*} en segundo lugar, obteniendo como resultado final la extensión {*a, d*, *c*}. Según el enfoque de Bench-Capon, con las mismas preferencias sobre los valores tenemos que *d* derrota a *c* (ya que el primero ataca al segundo y el valor del segundo no es preferido al del primero) y *c* derrota a *b* (por razón similar), pero el ataque de *b* a *a* queda sin efecto derrotador (ya que el valor de *a* es preferido por sobre el de *b*). En conclusión, los argumentos *a*, *d* y *b* son los ganadores, resultado que concuerda con el punto de vista que sostiene que la acción de Hal está justificada por resguardar su vida (*a*) y, aunque ha infringido el derecho a la propiedad de Carla (*b*), no debe compensarla (*d*). El modelo de Bench-Capon coincide ya que, suponiendo la superioridad del valor de la vida por sobre el de la propiedad privada, *d* derrota a *c* y *c* derrota a *b*, pero el ataque de *b* sobre *a* no tiene efecto derrotador.

Supongamos ahora que la preferencia entre valores se invierte, poniendo el derecho a la propiedad por encima del valor de la vida. Entonces los ataques efectivos o derrotas se dan de *c* a *b* y de *b* a *a*, mientras queda sin efecto el ataque de *d* a *c*. Ahora Hal infringe el derecho a la propiedad privada de Carla y debe compensarla (*c*). Pero ¿qué hay del argumento *d*? ¿Qué sentido tiene que no resulte descartado? El problema es que el modelo no contempla que el ataque presupone de algún modo el conflicto. La preferencia del valor del argumento atacado puede evitar el efecto derrotador del atacante, pero eso no quiere decir que elimine el conflicto subyacente. Suponer la superioridad de la propiedad privada sobre la vida puede evitar la derrota de *c* por *d*, pero eso no quiere decir que ambos argumentos sean aceptables a la vez. Evidentemente, sus conclusiones son contradictorias (Hal debe compensar a Carla - Hal no debe compensar a Carla) y eso hace que el modelo esté validando un resultado poco razonable. Quizá en la abstracción se gane mucho, pero también se pierda demasiado. O quizá el modelo no fue correctamente usado para representar el problema.

Desde el enfoque del haz de marcos argumentativos de Gabbay en versión *jerárquica*, el resultado luce mejor. Si la vida está por encima de la propiedad privada, entonces esa preferencia impone primero la elección del valor "vida" y se rechaza el valor "propiedad privada"; los argumentos *b* y *c* quedan descartados y sólo se aceptan *a* y *d*: Hal está justificado en tomar la insulina de Carla y no debe compensarla. Si la propiedad está por encima de la vida, entonces *a* y *d* se rechazan (por pertenecer al marco de valor inferior) y sólo se acepta *c* (*b* se rechaza por ser atacado por *c*): Hal no está justificado en tomar la insulina de Carla y debe compensarla.

Ahora bien, ¿esto muestra que el enfoque del haz de marcos en versión jerárquica funciona mejor que el de versión direccional? ¿Pueden salvarse los problemas observados en el segundo? Está claro que estoy ponderando dos enfoques en base a un solo ejemplo y eso hace que cualquier conclusión carezca de validez general, pero al menos el análisis servirá, en todo caso, para esclarecer el posible origen de los problemas y plantear algunas hipótesis para su solución.

5. PROBLEMAS CONCEPTUALES Y REPRESENTACIONALES

Voy a clasificar los problemas encontrados en a) conceptuales, y b) representacionales. Por problemas conceptuales me refiero a la determinación del significado de los términos de los modelos. Con respecto a los representacionales, me refiero a cómo identificar y hacer corresponder los elementos de un debate o situación argumentativa real con la ontología de los modelos, a fin de que éstos representen adecuadamente

dicha situación. La elucidación de los problemas conceptuales puede ayudar a resolver los representacionales. Estos problemas se manifiestan tanto en el enfoque de Gabbay como en el de Bench-Capon, pero en realidad son heredados en gran parte del modelo abstracto de Dung.

5.1. Problemas conceptuales

El problema conceptual que, en mi opinión, genera mayores dificultades es la noción de ataque. Al igual que en el modelo de Dung, aquí *ataque* (*attack*), al igual que *argumento*, es un término primitivo, lo que manifiesta en parte el carácter abstracto del sistema. En principio, no hay problema con ello, del mismo modo que no hay problema en que *conjunto* o la relación de *pertenencia* sean términos primitivos en teoría de conjuntos. Pero esto no quiere decir que las aplicaciones de los términos sean directas y unívocas en cualquier contexto. En los contextos de aplicación debe quedar claro qué condiciones mínimas deben cumplirse para considerar que un argumento ataca a otro, del mismo modo que se suponen determinadas condiciones para distinguir un conjunto de sus elementos cuando aplicamos teoría de conjuntos. Es decir, aún cuando se trata de un término teórico primitivo, hay un significado implícito de *ataque* que determina usos correctos e incorrectos en los contextos de aplicación.

Algunos análisis han contribuido para aclarar algunas de estas condiciones y sus implicaciones para las semánticas de extensiones (Bodanza, 2015). Una de las condiciones que parece ineludible es la de *conflicto*: si *x* ataca a *y* entonces *x* e *y* no pueden aceptarse a la vez. Es decir, el conflicto entre los argumentos parece una condición necesaria para que uno ataque al otro. El conflicto puede darse por razones lógicas (inconsistencia), pragmáticas (por ejemplo, por conducir a decisiones que, combinadas, producen resultados indeseados), o de alguna otra índole. Surge claramente que la relación de conflicto es simétrica. Sin embargo, la de ataque no lo es. Si pensamos en algunos ejemplos típicos de la literatura sobre razonamiento no monótono, veremos que es posible diferenciar cuándo los ataques se dan sólo en un sentido o en ambos. El argumento 'Tweety vuela porque es un ave y las aves normalmente vuelan' puede ser atacado por el argumento 'Tweety no vuela porque es un pingüino y los pingüinos son una clase específica de aves no voladoras'. Pero un ataque en sentido inverso no luce razonable, dada la mayor especificidad de la información contenida en 'Tweety es un pingüino' con respecto a la de 'Tweety es un ave'. Estas condiciones son tenidas en cuenta en algunos sistemas argumentativos no abstractos, donde los argumentos se construyen en base a reglas derrotables. Por ejemplo, en Simari y Loui (1992) y su secuela, García y Simari (2004), la derrota de *b*

por *a* supone (i) que *a* y *b* están en conflicto o desacuerdo (*disagreement*), entendiendo que esto se da cuando algunas proposiciones de *a* y de *b* se contradicen entre sí, y (ii) *a* se basa en información más específica que *b*. No podemos afirmar que el concepto de ataque utilizado en Dung *deba* entenderse en un sentido similar al de derrota en el sistema de Simari y Loui, ya que, además de la especificidad de la información, otros criterios de preferencia pueden tenerse en cuenta, tales como evidencia más o menos firme a favor de uno de los argumentos o, como hemos visto, una jerarquía entre los valores promovidos. Pero sí parece que, además del conflicto, algún criterio de preferencia queda implícito de cierto modo en el ataque. Esta preferencia puede ocurrir en versión asimétrica (estricta, como en el caso de "es más específico que" de Simari y Loui) o no (como en el caso de "no es más preferido que", equivalente a "tanto o más preferido que o incomparable a", de la derrota en Bench-Capon). En el primer caso, los ataques serán a su vez asimétricos, pero en el segundo, no necesariamente. El argumento 'Nixon es pacifista porque es cuáquero y los cuáqueros tienden a ser pacifistas' puede atacar a, y a su vez ser atacado por, el argumento 'Nixon no es pacifista porque es republicano y los republicanos tienden a no ser pacifistas'. Los argumentos están en conflicto, pero la ausencia de preferencias estrictas entre ellos habilita el ataque en ambos sentidos.

Así pues, asumiendo que el conflicto es una condición necesaria del ataque, que un ataque no se dé, se ignore, o no sea efectivo, no implica que desaparezca el conflicto subyacente. Supongamos que un argumento *a*, que promueve el valor del derecho a la propiedad, ataca un argumento *b*, que promueve el valor de preservación de la vida, y que en mi escala de valores el valor de la vida es superior al de la propiedad. Entonces yo puedo considerar que el ataque de *a* no tiene efecto derrotador sobre *b*. Sin embargo, esto no es suficiente para admitir, a la vez, los argumentos *a* y *b*, porque el conflicto puede persistir (por ejemplo, si los argumentos concluyen 'tengo justificación para infringir el derecho a la propiedad de otros en la circunstancia C' y 'no tengo justificación para infringir el derecho a la propiedad de otros en la circunstancia C', respectivamente). Si esto es así, un mínimo de racionalidad indica rechazar al menos uno de ellos.

Como veremos enseguida, obviar estas consideraciones conceptuales puede acarrear malas aplicaciones del modelo a la hora de representar una situación, con la consecuencia de producir resultados inaceptables.

5.2. Problemas representacionales

Dos trabajos de Bench-Capon en los que alude al caso de la insulina pueden servirnos para introducir el problema representacional. La representación vista antes es la ofrecida en Bench-Capon (2002, 2003a). En Bench-Capon (2003b), en cambio, por un lado el autor identifica al argumento *d* (Hal no debe compensar a Carla porque estaría justificado en tomar la insulina aún si fuera muy pobre) con el argumento *a* (Hal está justificado por una necesidad vital) y, por otro lado, representa los ataques formando un ciclo: $c \rightarrow b \rightarrow a \rightarrow c$. Aún aceptando que *a* y *d* puedan reducirse a un mismo argumento (lo que, naturalmente, despierta dudas), surge la pregunta de por qué, en tal caso, *a* ataca a *c* en una representación y no en la otra. No atribuyo estos problemas a defectos del modelo (ni a la capacidad de Bench-Capon para utilizarlo) sino a la dificultad general para comprender la argumentación real en los términos categoriales de los modelos. Los problemas representacionales se conectan directamente con los conceptuales. Sin un significado preciso de 'ataque' es imposible hablar de representaciones correctas o incorrectas. El concepto de 'argumento', a su vez, no presenta menos problemas. Bench-Capon (2003b) enuncia los argumentos como sigue:

> *a:* Hal puede tomar la insulina ya que de otro modo morirá
> *b*: Hal no debe tomar la insulina de Carla ya que es la propiedad de Carla
> *c:* Hal debe reponer la insulina de Carla una vez pasada la emergencia[4]

Aunque algunos aspectos del fraseo pueden considerarse irrelevantes y no dificultan la identificación de estos argumentos con la presentación de Bench-Capon (2003a), otros parecen claves. 'Puede tomar la insulina' y 'no debe tomar la insulina' son claramente contradictorias. Consideremos, además, el argumento *d* tal como es introducido en Bench-Capon (2003a). Refiriéndose a Christie (2000), el autor dice:

> "Él, entonces, introduce un cuarto argumento (*d*), que dice que si Hal fuera demasiado pobre para compensar a Carla, de todos modos tendría permitido tomar la insulina, ya que nadie debería morir por ser pobre. Además, él dice eso porque Hal no pagaría compensación si fuera demasiado pobre ni estaría obligado a hacerlo, aún si pudiera." (Bench-Capon, 2003a: 443. Trad. mía)

Entonces estamos frente a un argumento compuesto, ya que se esgrime en sostén de dos conclusiones: (1) que Hal tiene permitido tomar la insulina, y (2) que Hal no tiene obligación de compensar a Carla. Si consideramos la conclusión (1), entonces *d* está en conflicto (por contradicción lógica) con *b*, y si consideramos la conclusión (2), entonces *d* está en conflicto (por contradicción lógica) con *c*. Sin embargo, Bench-Capon entiende

[4] En el idioma original del texto: a: Hal can take the insulin as otherwise he will die; b: Hal must not take Carla's insulin because it is Carla's property; c: Hal must replace Carla's insulin once the emergency is over.

que *d* ataca a *c*, y no considera que *d* ataque a *b*. Es más, acepta que *b* y *d* resulten justificados a la vez.

Veamos qué resultado determinan los enfoques de Bench-Capon y Gabbay si aceptamos, de acuerdo a lo dicho, que *d* ataca a *b* y *c*, a la vez. La representación del escenario quedaría así:

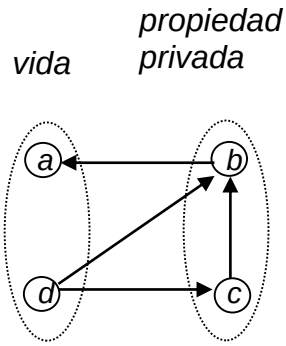

Figura 6

Entonces, si el valor de la vida es preferido sobre el de la propiedad, tanto Bench-Capon como Gabbay (en su versión jerárquica de haz de marcos argumentativos escogiendo conjuntos admisibles) determinarán la extensión {*a*, *d*}. En el caso de Bench-Capon, ahora el resultado parece más razonable que la aceptación de esos argumentos junto con *b*, como ocurría en la representación de la Figura 5. En el caso de Gabbay, en cambio, el resultado no varía. Si el valor preferido es el de la propiedad, entonces Bench-Capon sanciona el conjunto {*a*, *c*}, que parece más razonable que aceptar esos argumentos juntamente con *d*. En el enfoque de Gabbay se obtiene {*c*}, por lo que tampoco varía con respecto al marco anterior.

Tal vez mi interpretación haya llevado a una representación mejor. Pero el punto es que no hay un estándar objetivo con el cual comparar. Todo lo que tenemos son intuiciones, la de Bench-Capon, la de Gabbay, la mía. Pero, ¿qué tan comunes son estas intuiciones?

6. UNA MIRADA EMPÍRICA

En la sección anterior me referí a problemas conceptuales y representacionales del modelo de argumentación de Dung y las ampliaciones de Bench-Capon y Gabbay. Ahora quiero dar un giro al asunto planteando lo siguiente: ¿Puede resolverse de manera puramente analítica el problema de la justificación de argumentos? Si fuera un problema puramente lógico, quizá se resolvería con análisis veritativo funcionales, con semánticas formales, teoremas de representación o pruebas de completud y corrección.

Pero está claro que no lo es. La argumentación práctica es derrotable, las inferencias descansan en gran medida en el sentido común, y la aceptación o el rechazo se encuentran subjetivamente sesgados.

Representar el razonamiento de sentido común ha sido un objetivo de la Inteligencia Artificial desde sus inicios (McCarthy y Hayes, 1969) y diversos formalismos han pretendido capturar estas formas de razonar desde entonces, muchos sin siquiera una semántica clara (cf. Mc Dermott, 1987) y otros con semánticas claras, adecuadas desde un punto de vista lógico, pero cuya conexión con el sentido común se limita al cumplimiento de algunas propiedades formales.[5] En cambio, la pregunta acerca de si el comportamiento de esos sistemas tiene una contraparte verificable con el comportamiento real de las personas es algo poco explorado. La cuestión es que parece haber un salto de lo formal a lo empírico. Tratamos de avanzar lo más posible con nuestras herramientas lógicas y matemáticas, pero llega un punto en el que nos preguntamos si realmente es razonable lo que dicen los modelos. Después de todo, si hablamos de sentido común, ¿no deberíamos echar una mirada a lo que hace el sentido común? Puede objetarse, con cierta razón, que esto supondría abandonar toda pretensión normativa acerca de la argumentación correcta para pensar en una teoría descriptiva. Pero pienso más bien que tal vez debamos pararnos en un punto intermedio y buscar puentes entre lo formal y lo fáctico.

Este planteo no es original, pero el camino que marca ha sido poco explorado. En el caso de sistemas basados en argumentos podemos mencionar el trabajo de Rahwan et al. (2010),[6] que sirve para ilustrarnos qué tipos de cosas podemos descubrir. Allí ponen a prueba el "principio de restablecimiento" (*reinstatement*), según el cual un argumento derrotado es restablecido si todos sus atacantes son derrotados. Ellos realizaron experimentos que muestran que un argumento que resulta aceptable cuando no presenta derrotadores, también es aceptado luego de derrotado y restablecido, pero

[5] Las lógicas no-monótonas, por caso, el sistema P de Kraus, Lehmann y Magidor (1990) o el sistema de Schlechta (1997), definen semánticas de modelos preferenciales, es decir, modelos ordenados según su "normalidad". Por ejemplo, modelos en los que las aves vuelan son más normales que aquellos en los que no vuelan. Así, puede interpretarse que, en el razonamiento de sentido común, de A se infiere B si B es verdadero en todos los modelos más normales en los que A es verdadero. Esto permite validar algunas propiedades de las inferencias como, por ejemplo, la *monotonía cauta*: si de A se infiere B y de A se infiere C, entonces de A y B se infiere C; pero no otras, como la monotonía: si de A se infiere C, entonces de A y B se infiere C (C puede ser verdadero en los modelos más normales en los que A es verdadero, pero no en los modelos más normales en los que A y B son ambos verdaderos).

[6] Para el caso de lógicas no-monótonas y razonamiento por defecto (*default reasoning*) en general, los lazos con estudios experimentales y la psicología cognitiva han sido más explorados. Podemos mencionar, entre otros, los trabajos de Ford y Billington (2000), Ford (2005), Pelletier y Elio (2005), etc. Pero estos trabajos no refieren estrictamente a argumentos y su interacción.

en menor grado. Este resultado muestra concordancia con las semánticas de extensiones de Dung, pero a su vez señala que el modelo se ajustaría más al sentido común humano si incorporase distintos grados de confianza sobre la aceptación.

Mi propuesta es aplicar una estrategia similar para, al menos, enriquecer las intuiciones sobre qué modelos se ajustan mejor a lo que entendemos por justificación argumentada de sentido común cuando intervienen valores. Una idea es observar en qué medida una determinada representación de un escenario se ajusta a lo que las personas perciben. Por ejemplo, he dicho que en la Figura 4 deberíamos tener, además de la flecha de *d* a *c*, una flecha de *c* a *d*, pues creo que esos argumentos se atacan mutuamente. Pero, ¿qué tan común es esta intuición? Entonces podríamos realizar un experimento en el que les presentamos ambos argumentos a las personas y preguntamos si creen que presentan conflicto, si se atacan mutuamente, si uno vence al otro. Esto nos daría una idea para generar reglas de representación que hagan más útil al modelo. También podemos investigar variaciones en la percepción de los distintos elementos del modelo. Dejo aquí una lista (no exhaustiva) de elementos sobre los que indagar y algunas preguntas relevantes. El lector puede pensar, a modo de ejemplo, que el cuestionario versa sobre el escenario de Hal y Carla y los argumentos *a*, *b*, *c* y *d*.

- *Aceptabilidad*: ¿Cómo evalúa, en una escala de 1 a 3 (1=inaceptable, 2=medianamente aceptable, 3=muy aceptable), la aceptabilidad de cada argumento?[7]

Esta pregunta permitiría comparar una/las extensión/es sancionada/s por una semántica con el conjunto de argumentos que resulta justificado para las personas. Por ejemplo, si un alto porcentaje de entrevistados con preferencia por la vida antes que por la propiedad declara muy aceptables a los argumentos *a* y *d* pero inaceptable al argumento *c*, eso indicaría que la propuesta de haz de marcos en versión jerárquica de Gabbay se ajusta más que la propuesta de Bench-Capon a la intuición común de la gente.

- *Conflicto*: ¿Considera que los argumentos presentes en cada uno de los siguientes pares están en conflicto? (*a* con *b, a* con *c*, *a* con *d*,…). Respuestas para cada caso: 1= no están conflicto, 2=tal vez estén en conflicto, 3=están en conflicto.

Esto permitiría conocer distintas percepciones respecto al conflicto. Por ejemplo, si la gente se inclina mayormente por observar incompatibilidad entre *d* y *b*, eso indicará que

[7] Podrían utilizarse estas u otras escalas con más valores, p. ej. la de Likert, dependiendo de la finura que esperemos en las respuestas.

la aceptación conjunta de esos argumentos en una misma extensión no refleja una intuición común.

- *Derrota:* ¿Cuál de los argumentos presentes en cada uno de los siguientes pares considera que derrota al otro? Opciones en cada par: *x* derrota a *y*; *y* derrota a *x*; ninguno derrota al otro.

Esto permitiría conocer distintas percepciones con respecto a la derrota. En particular, según la hipótesis (basada en el concepto de Bench-Capon) de que la derrota presupone conflicto y no preferencia del argumento derrotado, deberá observarse la tendencia a señalar el conflicto y la preferencia adecuada en los casos en los que se observa la derrota.

- *Valores:* ¿Cómo evalúa la importancia [moral, ética, práctica, etc.] de los valores involucrados? v_1 superior a v_2, v_2 superior a v_1, igualmente importantes.

Esta pregunta permitiría observar si la percepción de la aceptabilidad o las derrotas están sesgadas por el valor preferido. En nuestro ejemplo, no deberíamos observar una tendencia a señalar la derrota de *a* por *b* cuando el valor preferido es el de la vida.

También podría consultarse sobre la asociación de los argumentos con los valores. Por ejemplo, Bench-Capon asume que *a* y *d* promueven el valor de la vida, mientras *b* y *c* promueven el de la propiedad. ¿Podría pensarse que *c* promueve ambos valores a la vez, entendiendo que al compensar se balancea la necesidad vital con el respeto de la propiedad ajena? Esta interpretación confrontaría directamente con las representaciones vistas.

- *Efectos de marco* (*framing effects*): También resultaría muy interesante investigar si la aceptación de un argumento está sujeta a efectos de marco.

Se trata de un sesgo cognitivo por el que las preferencias de las personas, ante un problema de decisión, cambian según cómo éste se presente (Tversky y Kahneman, 1981). Podemos comparar las respuestas realizando variaciones de escenario para un mismo marco argumentativo. Por ejemplo, una variación podría versar sobre la legitimidad del aborto, tomando los valores del derecho a la vida del nonato vs. el derecho a la determinación de la mujer sobre su propio cuerpo. Esto permitiría observar qué tan objetiva puede ser la percepción acerca de la justificación argumentativa. Otra variación podría implicar una inversión de los valores. Por ejemplo, si en un marco se presentan argumentos *a* y *d* promoviendo el valor v_1 y *b* y *c* promoviendo el valor v_2, en otro podemos presentar argumentos *a'* y *d'* promoviendo el valor v_2 y *b'* y *c'* promoviendo el valor v_1, sin variar la estructura de los ataques. De modo que cambios en la aceptación

de argumentos de un marco a otro mostrarían una independencia de la persuasión con respecto a la estructura, y una dependencia con respecto a los valores.

- En las respuestas intra-sujeto se podría observar cómo se perciben las conexiones (condición necesaria/suficiente) entre distintos conceptos tales como conflicto-derrota, derrota-valores, conflicto-valores, e incluso aceptabilidad-valores.
- Otros sesgos podrían investigarse que alteren la aceptabilidad de los argumentos. Menciono algunas posibilidades:
    - Sesgo de confirmación: preferir los argumentos que confirman o concuerdan con las propias creencias independientemente del procedimiento argumentativo (Esaiasson et al., 2019). Por ejemplo, una creencia firme en defensa del derecho a la propiedad privada podría llevar a aceptar el argumento de que Hal no debió tomar la insulina de Carla, sin importar qué argumentos se ofrezcan para justificar lo contrario (lo mismo podría decirse del argumento contrario con respecto a la convicción sobre el valor de la vida).
    - Adecuación a la norma: por ejemplo, preferir los argumentos "políticamente correctos". Este sesgo llevaría a adoptar argumentos no sinceros cuando los propios se oponen a los argumentos pautados por una norma común, por ejemplo, por temor al rechazo o al repudio (Lapinski y Rimal, 2005). En nuestro ejemplo, si la norma es ponderar la vida por sobre la propiedad este sesgo podría inclinar las respuestas a favor de *a* aún cuando el argumento preferido sea *b*.
    - Sesgo de benevolencia o indulgencia (por ejemplo, cuando la argumentación tiene el fin de condenar o absolver, el sesgo puede llevar a preferir los argumentos que promuevan la absolución (McCoun y Kerr, 1988)), etc. En nuestro ejemplo, ante la duda, este sesgo podría conducir a aceptar los argumentos *a* y *d* si, por caso, Hal corriera el riesgo de ser condenado por el delito de hurto.

7. CONCLUSIÓN

Los modelos de argumentación sobre los que hemos hablado provienen del campo de la Inteligencia Artificial, es decir, de un campo donde el objetivo es crear sistemas cuyo

comportamiento pueda caracterizarse como inteligente. En el caso concreto de los sistemas de argumentación, la búsqueda se emprende por el camino de la representación del conocimiento que es materia de la argumentación y la simulación de procesos de razonamiento y decisión humanos. Por ello, a modo de reflexión final quiero traer a colación las relaciones entre teoría, experiencia y simulación, como muy bien las explica Jorge Wagensberg (1994): "La simulación [en nuestro caso, los modelos de argumentación] puede describir, predecir y, por lo tanto, sustituir a la experiencia", pero "La experiencia puede contradecir una simulación y, por lo tanto, sugerir su revisión o demolición, o denunciar la no disponibilidad de una simulación y, por lo tanto, sugerir su elaboración" (pp. 101-102). La simulación juega el rol de una teoría con respecto a la experiencia, ya que la describe, y juega el rol de experiencia con respecto a la teoría, ya que la pone a prueba. Entender estas relaciones, en mi opinión, ayuda a aliviar la tensión de la que hablan Kahneman y Tversky en el epígrafe. Asimismo, pone en relieve el valor de un enfoque desde la psicología del razonamiento para elucidar los problemas que presenta la argumentación en Inteligencia Artificial.[8]

En este trabajo me he propuesto señalar algunas líneas de investigación que contribuyan a establecer puentes entre los aspectos formales de los modelos argumentativos y la experiencia, observando cómo las personas perciben la argumentación. Esas líneas se desprenden de dos tipos de problemas observados: conceptuales (¿qué es una ataque?, ¿cuándo ocurre el conflicto?, ¿cuándo un ataque deviene en derrota?, ¿depende la aceptación de un argumento de los valores que promueve?, ¿depende sólo de la estructura dada por la interacción de los argumentos?, etc.) y representacionales (si el modelo considera sólo la relación por ataque, ¿cómo representar el mero conflicto?, etc.). De esta manera, he pretendido subrayar algunos aspectos concretos del carácter interdisciplinario de la argumentación, con respecto a los cuales la psicología del razonamiento tiene mucho que ofrecer a los modelos formales.

[8] Como observa un revisor anónimo, la estrategia no difiere de la propuesta por la (así llamada) filosofía experimental (Knobe, 2016).

REFERENCIAS

Baroni, P., y Giacomin, M. (2007). "On principle-based evaluation of extension-based argumentation semantics". *Artificial Intelligence* 171(10), 675-700. https://doi.org/10.1016/j.artint.2007.04.004

Bench-Capon, T. J. M. (2002). "Agreeing to differ: modelling persuasive dialogue between parties with different values". *Informal Logic* 22/3, 231-245.

Bench-Capon, T. J. M. (2003a) "Persuasion in practical argument using value-based argumentation frameworks". *Journal of Logic and Computation* 13(3), 429–448.

Bench-Capon, T. J. M. (2003b). Try to See it My Way: Modelling Persuasion in Legal Discourse. Artificial Intelligence and Law, 11(4), 271-287. https://doi.org/10.1023/B:ARTI.0000045997.45038.8f

Bodanza, G. (2015). "La argumentación abstracta en Inteligencia Artificial: problemas de interpretación y adecuación de las semánticas para la toma de decisiones". *Theoria. An International Journal for Theory, History and Foundations of Science* 30/3, 395-414. doi:https://doi.org/10.1387/theoria.13150

Christie, G.C., (2000). *The Notion of an Ideal Audience in Legal Argument*, Kluwer Academic Publishers.

Coleman, J., (1992). *Risks and Wrongs*. Cambridge University Press.

Dung, P. M. (1995). "On the acceptability of arguments and its fundamental role in nonmonotonic reasoning, logic programming and n-person games". *Artificial Intelligence* 77/2, 321-357.

Esaiasson, P., Persson, M., Gilljam, M., y Lindholm, T. (2019). "Reconsidering the role of procedures for decision acceptance". *British Journal of Political Science* 49/1, 291-314. https://doi.org/10.1017/S0007123416000508

Ford, M. (2005). "Human Nonmonotonic Reasoning: The Importance of Seeing the Logical Strength of Arguments". *Synthese* 146/1/2, 71-92.

Ford, M., Billington, D. (2000). "Strategies in human nonmonotonic reasoning". *Computational Intelligence Journal* 16/3, 446-468.

Gabbay, D. (2014). "Systems of interacting argumentation networks". *The IfCoLog Journal of Logics and their Applications* 1/1, 131-176.

García, A. J., y Simari, G. R. (2004). "Defeasible Logic Programming: An Argumentative Approach". *Theory and Practice of Logic Programming (TPLP)* 4, 95-138.

Kahneman, D., y Tversky, A. (2000). "Choices, values, and frames". En D. Kahneman y A. Tversky (Eds.), Choices, Values, and Frames (pp. 1-16). Cambridge: Cambridge University Press. doi:10.1017/CBO9780511803475.002

Knobe, J. (2016). "Experimental philosophy is cognitive science". En Sytsma, J. y Buckwalter, W. (eds.) *A Companion to Experimental Philosophy*. Blackwell. https://campuspress.yale.edu/joshuaknobe/publications/#8

Kraus, S., Lehmann, D., y Magidor, M. (1990). "Nonmonotonic reasoning, preferential models and cumulative logics". *Artificial Intelligence* 44/1, 167-207. https://doi.org/10.1016/0004-3702(90)90101-5

Lapinski, M. K., y Rimal, R. N. (2005). "An explication of social norms". *Communication Theory* 15(2), 127-147. https://doi.org/10.1111/j.1468-2885.2005.tb00329.x

McCarthy, J., y Hayes, P. J. (1969). "Some philosophical problems from the standpoint of Artificial Intelligence. En B. Meltzer & D. Michie (Eds.), *Machine Intelligence* 4 (pp. 463–502). Edinburgh University Press.

MacCoun, R. J., y Kerr, N. L. (1988). "Asymmetric influence in mock jury deliberation: Jurors' bias for leniency". *Journal of Personality and Social Psychology 54*/1, 21-33. https://doi.org/10.1037/0022-3514.54.1.21

Mailly, J.-G., y Maratea, M. (2019). "Assessment of benchmarks for abstract argumentation". *Argument & Computation*, *10*(2), 107-112. https://doi.org/10.3233/AAC-192101

McDermott, D. (1987). "A critique of pure reason". *Computational Intelligence* 3/1, 151-160. https://doi.org/10.1111/j.1467-8640.1987.tb00183.x

Pelletier, F., & Elio, R. (2005). "The case for psychologism in default and inheritance reasoning". Synthese 146/1/2), 7-35.

Perelman, Ch. y Olbrechts-Tyteca, L. (1989). *Tratado de la argumentación. Nueva retórica.* Madrid: Gredos.

Rahwan, Y., Madakkatel, M., Bonnefon, J-F., Awan, R., Abdallah, S. (2010). “Behavioral experiments for assessing the abstract argumentation semantics of reinstatement”, *Cognitive Science* 34/8, 1483-1502.

Schlechta, K. (1997). *Nonmonotonic Logics: Basic Concepts, Results, and Techniques*. Springer-Verlag. https://doi.org/10.1007/BFb0021104

Simari, G. R., y Loui, R. P. (1992). “A mathematical treatment of defeasible reasoning and its implementation”. *Artificial Intelligence* 53, 125-157.

Tversky, A, y Kahneman, D. (1981). "The framing of decisions and the psychology of choice". *Science* 211/4481, 453–58.

Wagensberg, J. (1994). *Ideas sobre la complejidad del mundo* (3ra. edic.). Barcelona: Tusquets Editores.

AGRADECIMIENTOS: Agradezco los comentarios y críticas de los revisores anónimos de la revista, que me han llevado a mejorar sustancialmente el artículo. Este trabajo ha sido realizado en el marco de proyectos subvencionados por ANPCyT (PICT 2017-1702) y SECyT-UNS (PGI 24/I265), Argentina.

GUSTAVO ADRIÁN BODANZA: Doctor en Filosofía, Profesor Titular Ordinario de la Universidad Nacional del Sur e Investigador Independiente del Consejo Nacional de Investigaciones Científicas y Tecnológicas, Argentina. Ha realizado contribuciones sobre lógica y sistemas argumentativos en Inteligencia Artificial y sobre argumentación en la toma de decisiones colectivas. Ha publicado artículos en *Argument and Computation*, *International Journal of Psychology*, *Journal of Applied Logic*, *Journal of Cognitive Psychology*, *Journal of Logic and Computation*, *Journal of Logic*, *Language and Information*, entre otras revistas de prestigio.