CUADERNOS FHyCS-UNJu, Nro. 50: 213-231, Año 2016

RESTABLECIMIENTO, DERROTA Y DESACTIVACIÓN: UNA DISCUSIÓN SOBRE LA ACEPTABILIDAD DE ARGUMENTOS

(REINSTATEMENT, DEFEAT AND DEFUSE: A DISCUSSION ABOUT ARGUMENTS ACCEPTABILITY)

Claudio Andrés ALESSIO*

RESUMEN

El restablecimiento es un principio de los *sistemas argumentativos* de acuerdo al cual, un argumento derrotado puede contar como justificado siempre y cuando sus derrotadores estén derrotados. No es relevante distinguir el tipo de derrotas para ello. Todas son consideradas equivalentes y adecuadas para este fin. En el presente trabajo, tal idea será cuestionada. Este cuestionamiento llevará inmediatamente a la siguiente pregunta ¿cómo debe ser re-definido el restablecimiento? Para responder al interrogante se seguirá una metodología usual en el área, primero se exponen una serie de ejemplos, unos que se adecúan a la noción estándar del restablecimiento, *ejemplos bien comportados,* y otros en los que se obtienen resultados no esperados y cuestionan su validez. Se comparan ambos tipos de ejemplos, se destacan similaridades y diferencias, se analizan posibles causas de tales diferencias y se propone una manera de interpretar los resultados. Como resultado se obtiene que en los *ejemplos bien comportados* hay un tipo de derrota que está ausente en los que no lo hacen. Específicamente, esta derrota se caracteriza porque la aceptación del derrotado lleva a creer que el derrotado se encuentra sustentado en algún supuesto falso, incorrecto o poco plausible. Este tipo especial de derrota será denominado *desactivación*. El nuevo significado del restablecimiento propuesto será entonces así: un argumento, a pesar de haber sido derrotado por otro, puede considerarse justificado siempre y cuando sus derrotadores estén desactivados.

Palabras Clave: razonamiento default, argumentación rebatible, restablecimiento.

ABSTRACT

Reinstatement is a principle in Argumentation Systems according to which defeated arguments should be regarded as justified if the arguments defeating them are themselves defeated. The kind of defeat involved in the process is irrelevant. All defeats are equivalent to any other. In the paper, this idea is questioned and motivates the following question: ¿How reinstatement should be re-defined? We

* Universidad Nacional de Santiago del Estero - Facultad de Humanidades, Ciencias Sociales y de la Salud - Av.Belgrano (S) 1992 – CP 4200 - Santiago del Estero - Argentina.
Correo Electrónico: manon@arnet.com.ar

answer the question following a usual method in the field: first, we show some well-behaved example with respect to standard reinstatement and others examples suggest that reinstatement cannot be taken as a general principle. Second, we compare both kinds of examples and identify similarities and differences into the examples. Later, we analyze possible causes of such differences and propose a way of interpreter this situation. As a result we have that in the well-behaved examples there is special kind of defeat that is not present in the others. This defeat can be characterized like this: the acceptation of the defeater argument leads to accept that defeated argument is based on false, improbable, or incorrect support. This special kind of defeat is named defuse in the paper. So, the new meaning of reinstatement proposed here is: an argument can be justified if all defeaters are in turn defused by a justified argument.

Key Words: *Default Reasoning, Defeasible Argumentation, Reinstatement.*

INTRODUCCIÓN

Los *sistemas argumentativos* (*argumentation systems*) en tanto formalismos para la representación del conocimiento y el razonamiento tienen como objetivo representar información tentativa y potencialmente contradictoria, en un lenguaje formal determinado. A partir de ella se construyen estructuras denominadas argumentos. Los argumentos son entidades que sustentan afirmaciones y que permiten extender la teoría de partida. Cuando los argumentos son construidos es posible que dos o más de ellos no puedan ser aceptados de manera conjunta. Por ello, en los sistemas argumentativos los argumentos son estructuras que derrotan a (o son derrotados por) otras entidades del mismo tipo. El objetivo final de un sistema argumentativo consiste en estipular cuál es el conjunto de conclusiones razonables aceptadas por el sistema en base a la comparación y evaluación de los argumentos en conflicto que sustentan tales conclusiones. Las conclusiones de aquellos argumentos que prevalezcan frente a sus rivales serán consideradas las aceptadas por el sistema (Prakken y Vreeswijk, 2001). De modo que una fase crucial en el proceso consiste en identificar una subclase de los argumentos construidos que pueden considerarse como los argumentos justificados.

Un principio fundamental en los formalismos argumentativos, del que depende el mecanismo de justificación, es el restablecimiento: Si un argumento *a* es derrotado por *b*, y en ese momento un agente debe decidir si cree en la conclusión de *a* o *b;* elegirá creer en la conclusión de *b* pues este no cuenta con un derrotador. Ahora bien, si luego sabe que *b* es derrotado por *c,* entonces, el agente puede creer tanto en la conclusión de *c,* como en la conclusión de *a,* puesto que *a* se ha visto restablecido (o defendido) por *c.* Más en general, un argumento derrotado podrá contar como justificado cuando todos sus derrotadores estén a su vez derrotados.

A pesar de que el restablecimiento es considerado un principio general en los sistemas argumentativos, es posible encontrar ejemplos que sugieren su invalidez. Esto puede ser ilustrado con la siguiente historia.

El director de la Agencia Internacional de Espionaje (AIE), el teniente *K*, tiene buenas razones para creer que el contra-agente John (de la Agencia Especial de Espionaje Rufa) ha tenido un destino fatal basado en la información suministrada por el agente especial Charles el día domingo. Este le dijo: *John se cayó de un avión*. Ahora supóngase que al día siguiente, *K* recibió información del superagente Robert: *"K, he visto a John arrojarse de un avión pero con un paracaídas puesto"*. En este momento, *K,* cree que John se ha salvado de su fatal destino. Dos días después, el director de la agencia, recibe un nuevo mensaje, pero ahora de parte de la Señorita Escarlata: "*John saltó de un avión con un paracaídas averiado puesto*". Con esta nueva pieza de información, *K* puede creer que John, efectivamente tuvo un destino fatal.

El anterior ejemplo, puede ser representado mediante los siguientes tres argumentos.

Ejemplo 1.1

a. John ha tenido un destino fatal puesto que quien cae de un avión tiene un destino fatal.

b. John no ha tenido un destino fatal puesto que quien se arroja de un avión con un paracaídas puesto, usualmente cae ileso.

c. John ha tenido un destino fatal puesto que se arrojó de un avión con un paracaídas averiado.

Para *K,* el único argumento razonable sería el basado en la información suministrada por la Señorita Escarlata (*c*). Vale la pena remarcar que a pesar de que *b* derrota *a*, y *c* a *b*, no es razonable la restauración del estado del argumento basado en lo dicho por Charles, i.e. en este ejemplo claramente no debería aplicarse el restablecimiento (obviamente se restaura el estado de la conclusión del argumento *a*, pero no su sustento)

Ahora bien ¿Por qué el restablecimiento parece un principio intuitivamente razonable pero no lo es tanto a la luz de ejemplos como el considerado? ¿Ejemplos como el propuesto evidencian que el restablecimiento no es un principio general y por lo tanto debe ser explicitado el alcance o las condiciones de aplicación? Esto último debe pensarse cuidadosamente, puesto que, como ya se dijo, en los sistemas basados en argumento, el restablecimiento es efectivamente un principio general y se supone que puede ser aplicado en cualquier caso donde se dé que al menos un argumento derrote a todos los rivales de otro.

En el ejemplo 1.1 se verifica las condiciones indicadas pero no parece razonable que se restablezca el estado original del argumento *a*. Por ello, en el presente artículo se pretenderá identificar una posible causa de la aparición del comportamiento del restablecimiento frente a casos como el ejemplo 1.1.

La metodología empleada será la usual en el área, primero se expondrán una serie de ejemplos, unos que se comportan adecuadamente cuando el restablecimiento es considerado y otros en los que se obtienen resultados no esperados. Se comparan ambos tipos de ejemplos, se destacan similaridades y diferencias. Luego, se analizan posibles causas de tales diferencias,

para proponer una manera de interpretar los resultados que explique la situación.

Como resultado se obtiene que en los ejemplos bien comportados se pueden detectar notables diferencias con respecto a los mal comportados. Especialmente, el análisis permite determinar que el restablecimiento es intuitivo y correcto cuando se consideran ejemplos en los que el restablecedor lleva a creer que el rival del defendido se encuentra sustentado en algún supuesto falso, incorrecto o poco plausible. Por su parte, en los ejemplos que sugieren la invalidez del restablecimiento, el que operaría como restablecedor, no efectúa el mismo tipo de derrota que la marcada recientemente. Parece provechoso atender a esta diferencia y sugerirla como una característica principal para la aplicación del restablecimiento.

Si todas las derrotas (aquellas que revelan que el rival está sustentado en algo falso, incorrecto o poco plausible y las que no) son consideradas en los sistemas argumentativos como equivalentes (como actualmente sucede) emergen ejemplos problemáticos. Si en cambio se contara con una distinción entre las derrotas que permiten restablecer por un lado, de las que no, los casos en los que no valdría el restablecimiento podrían ser convenientemente atendidos. Por ello la contribución del presente artículo consiste en la distinción conceptual entre ambas derrotas. Llamando a las primeras como desactivación y a las otras simplemente derrotas. Esta distinción conceptual viene acompañada de una definición de la desactivación y de una redefinición del restablecimiento.

SISTEMAS ARGUMENTATIVOS: NOCIONES BÁSICAS

Los sistemas argumentativos son formalismos en los que se construyen argumentos según ciertas reglas previamente fijadas partiendo de un conjunto de información expresado en un lenguaje formal determinado. En el conjunto de información inicial, en general, se encuentran sentencias o enunciados referidos a hechos y a reglas. Ejemplos de tales enunciados pueden ser los siguientes:

- Hechos: '*Bob es un bebé murciélago'*, '*Tweety es un pingüino', 'Tom robó una biblioteca'.*
- Reglas rebatibles. '*Por lo general los murciélagos vuelan', 'Por lo general los mamíferos no vuelan', 'Por lo general las aves vuelan'.*
- Reglas estrictas*: 'Todos los murciélagos son mamíferos', 'Todos los pingüinos son aves'.*
- Presunciones: '*Tweety parece ser un ave'*, *'Bob parece ser un murciélago'*, '*me parece que es Tom el que está robando allí'*.

En la teoría (*conjunto de información inicial*) también podría haber información dada por informantes, dicho de manera más simple, podrían haber sentencias que expresen dichos como: '*Juan dice que Tweety es un ave', 'Pedro dice que Bob es un bebé murciélago',* etc.

Estos sistemas fueron propuestos inicialmente en los trabajos de Pollock (1987) y Loui (1987). Tales trabajos han permitido caracterizar a los sistemas en base a dos dimensiones: una denominada lógica y otra dialéctica. Ambos

trabajos tienen algo en común, a pesar de pertenecer a terrenos disciplinares diversos, el primero, del ámbito de las ciencias de la computación, el segundo desde la filosofía. En uno y otro se ha pretendido diseñar o proponer una teoría que contribuye al diseño y creación de un *artilect* (agente racional), pero también ambos trabajos han contribuido al desarrollo de una teoría de la prueba rebatible o del razonamiento rebatible. Por tanto, los sistemas argumentativos son desde sus inicios de gran interés tanto para la filosofía como para la computación.

La primera dimensión, la lógica, permite definir el proceso de construcción de los argumentos, tal como brevemente se ha dicho. En ella se determina el lenguaje formal, que expresaran sentencias como las que han sido ejemplificadas más arriba, las reglas para la construcción de argumentos, las reglas de inferencia y una noción de argumento propiamente dicha. Un ejemplo de argumento modelado por estos sistemas, que ya se considera canónico, es el siguiente:

Ejemplo 2.1
Dado que por lo general las aves vuelan y que Tweety es un ave, es razonable concluir que Tweety vuela.

También, los argumentos que apelan a información suministrada por alguien como:

Ejemplo 2.2
Dado que por lo general las personas dicen la verdad y que Juan di ce que Luis no es digno de confianza se puede concluir que Luis no es digno de confianza.

Esta primera dimensión no necesariamente debe dar cuenta de todos los aspectos (*premisa, conclusión, preferencia, subargumento, etc.*), varios de ellos pueden permanecer no especificados, permitiendo de este modo concentrarse en la dimensión dialéctica de los sistemas.

La segunda dimensión, la dialéctica, define el tipo de *relaciones que pueden darse entre los argumentos*. Se denomina dialéctica simplemente por el hecho de tratarse de los aspectos interactivos entre argumentos. Usualmente las relaciones consideradas en los sistemas son *conflicto, preferencia* y *derrota*.

Además de los tipos de relaciones, en la dimensión dialéctica se considera lo que se denomina como *estado final* de un argumento. Luego de finalizado el proceso de comparación, donde se dan las relaciones de conflicto, preferencia y derrota, algunos argumentos prevalecerán por sobre otros, al igual que en una discusión o debate, y otros no. El prevalecer o no frente a los rivales es lo que se denomina *estado final* de un argumento. El estado final que un argumento puede adquirir son: *justificado,* o *rechazado*. Hay otros análisis que complejizan el estado final pero no es necesario hacerlo aquí.

Retomando la cuestión relativa a las relaciones que pueden darse entre argumentos, se procederá a brindar una breve explicación de las principales.

Una relación de *conflicto* entre pares de argumentos se da cuando una proposición (premisas o conclusión) de uno niega una proposición del otro (premisas o conclusión). Por ejemplo:

Ejemplo 2.3
a. Teniendo en cuenta que por lo general los cuáqueros son pacifistas y que Nixon es cuáquero se puede concluir que Nixon es pacifista.
b. Teniendo en cuenta que por lo general los republicanos no son pacifistas y que Nixon es republicano se puede concluir que Nixon no es pacifista.

En el ejemplo 2.3 puede verse que las conclusiones de ambos argumentos son contradictorias. Podría haber otras situaciones conflictivas, por ejemplo, debido a contradicción entre la conclusión de un argumento y una premisa en el otro.

A partir de un orden entre los argumentos previamente especificado se establecen las relaciones de *preferencia* que permite decidir, en caso de conflicto, qué argumento es mejor. Dependiendo del tipo de información modelada se empleará un criterio de preferencia específico, por ejemplo, prioridad entre leyes en el razonamiento legal; confiabilidad en el razonamiento plausible; especificidad en el razonamiento default, entre otros. Por ejemplo

Ejemplo 2.4
a. Teniendo en cuenta que por lo general los mamíferos no vuelan y que Stella Maris es un mamífero, es posible concluir que Stella Maris no vuela.
b. Teniendo en cuenta que por lo general los murciélagos vuelan sabiendo que Stella Maris es un murciélago, es posible concluir que Stella Maris vuela.

En el ejemplo 2.4, ambos argumentos están en conflicto porque sus conclusiones son contradictorias. Ahora bien, también es claro que el argumento *b* apela a información más específica y lo hace más razonable. En ese caso se dice que *b* es *preferido* a *a* por especificidad.

La relación de *derrota* es una relación dada entre dos argumentos en conflicto. Esta relación determina qué argumento prevalece en caso de conflicto. Ha sido definida como:

$derrota =_{df} conflicto + no\ preferencia.$

Esta noción no necesariamente da cuenta de todos los casos considerados tal como se discute en (Bodanza, 2015). En los ejemplos presentados aquí, la noción funciona adecuadamente. De modo que no será preciso atender a tal situación.

La derrota se aplica en el ejemplo 2.4 de la siguiente manera. Dado que *a* y *b* están en conflicto y *a* no es preferido a *b* entonces *b* derrota a *a.*

En el ejemplo 2.3, *a* derrota a *b* y *b* derrota a *a* porque ambos están en conflicto y no hay preferencia por uno u otro.

Obviamente que podría argüirse aquí que en el ejemplo 2.3 serían definibles preferencias en base a lo que se entiende por pacifista o una preferencia política, sin embargo, la idea del ejemplo, tal como ha sido entendida por la tradición en lógica default o argumentación rebatible, consiste en ilustrar la noción de conflicto irresoluble, y en el mejor de los conocimientos del autor, se desconoce una interpretación de este ejemplo en la literatura especializada que se haga ese tipo de interpretación, o donde se destaque una preferencia posible que no sea explícitamente considerada.

De la misma manera que en la dimensión lógica, algunos aspectos de la dimensión dialéctica, en particular las relaciones de conflicto y preferencia, pueden permanecer parcial o totalmente no especificadas. Esto es así porque para determinar el estado final de los argumentos y conocer cuáles son los argumentos justificados, solo es necesario saber cuáles son los argumentos y cuáles las relaciones de derrota que mantienen entre ellos. Pero también es requerido que se cuente con un criterio de evaluación de los argumentos, denominado por Dung (1995) *'semántica de extensiones'* (*extensions semantics*).

Esta idea fue propuesta originalmente por Dung (1995) y expresado en un formalismo llamado marcos argumentativos (*argumentation framework)*. Los marcos argumentativos determinan cuál o cuáles de todos los argumentos disponibles (en un marco argumentativo particular) están justificados en base a las semánticas de extensiones. Un criterio común que subyace a cualquiera de las semánticas clásicas es el de argumento aceptable y el de conjunto admisible. Intuitivamente, un argumento aceptable es un argumento que se entiende como un buen candidato para considerarlo, al finalizar el proceso, como aquel que prevalece frente a sus rivales. Precisando un poco más la idea, puede decirse que un argumento *a* será considerado aceptable cuando para cualquier argumento *b* que lo derrote, existe al menos un argumento *c* en un conjunto de argumentos *S* que lo defiende, es decir, *c* derrota a *b.* Esta noción permite advertir que un argumento será considerado justificado cuando sus rivales estén derrotados.

El conjunto admisible establece las condiciones mínimas de justificación de un conjunto de argumentos. Éstas son: que todos los miembros del conjunto sean aceptables, i.e. que se defiendan entre sí y que puedan estar todos juntos (que no se derroten entre ellos).

Una noción central, involucrada en la definición de aceptabilidad, y por tanto en los criterios de evaluación de argumentos, es la de *restablecimiento* (*reinstatement*), también llamada defensa (*defense*). En términos más o menos generales, el restablecimiento puede ser considerado como la propiedad que permite a un argumento recuperar su estado: Supóngase que se cuenta con el argumento *a*. Supóngase también que originalmente *a* cuenta con un estado epistémicamente positivo, ahora bien, si un argumento *b* derrota a *a*, *a* pierde ese estado, pero si un argumento *c* derrota a *b*, se dice que *c* restablece (o defiende) a *a* y *a* recupera su estado original.

El restablecimiento es un principio general de la argumentación rebatible que se comporta adecuadamente, sin embargo, es posible identificar ejemplos

que sugieren que el restablecimiento no podría ostentar tal carácter, como se verá a continuación.

RESTABLECIMIENTO DE ARGUMENTOS

El restablecimiento de argumentos es una noción central en la mayoría de los sistemas argumentativos, en especial aquellos que pueden ser definidos como instancias de un marco argumentativo (Dung, 1995). La idea subyacente en este principio es que un argumento puede considerarse aceptable, cuando todos los derrotadores disponibles en el marco en que está siendo considerado están a su vez derrotados. El siguiente ejemplo ilustra la idea:

Ejemplo 3.1
a. Tweety vuela porque es un ave.
b. Tweety es un pingüino dado que esto ha sido observado, luego Tweety no vuela.
c. La observación de que Tweety es un pingüino no es confiable dado que fue hecha durante una tormenta de nieve.

En el marco definido por estos argumentos, parece tener sentido el hecho de que Tweety vuele y considerar a los argumentos *a* y *c* como aceptables o *justificados*, dado que *c* carece de derrotadores y el argumento que podría llevar al rechazo de *a* ha sido derrotado por *c*. Sin embargo, otros ejemplos parecen sugerir que el restablecimiento no es un principio general.

Ejemplo 3.2
a. Tweety vuela porque es un ave y las aves por lo general vuelan.
b. Tweety no vuela porque es un pingüino y los pingüinos por lo general no vuelan.
c. Tweety vuela porque es un pingüino genéticamente modificado y los pingüinos genéticamente modificados por lo general tienen la capacidad de volar.

Horty (2001) emplea un ejemplo similar para cuestionar la validez del restablecimiento, cuestionamiento que parece plausible dado que si *a* y *c* son conjuntamente admitidos, entonces, una conclusión correcta (Tweety vuela) puede estar justificada en base a razones incorrectas (vuela porque es un ave). Claramente, la razón por la que Tweety vuela no se debe a una habilidad que las aves por lo general tienen, sino por una habilidad que específicamente tienen los pingüinos genéticamente modificados.

Adicionalmente, es posible considerar otros casos donde no sólo se tienen argumentos con malas razones que sustentan conclusiones correctas, sino que también, el restablecimiento parece llevar a la aceptación de conclusiones que son simplemente incorrectas.

Ejemplo 3.3

a. John es millonario porque es empleado de Microsoft y por lo general ellos tienden a ser millonarios.

b. John tiene menos que medio millón porque es un nuevo empleado de Microsoft y por lo general los nuevos empleados de Microsoft poseen menos que medio millón.

c. John tienen al menos medio millón porque es un nuevo empleado de Microsoft pero en el departamento X y los nuevos empleados de Microsoft que trabajan en el departamento X por lo general tienen al menos medio millón.

Aquí, la conclusión de *a* es más fuerte que la conclusión de *c*. El argumento más general (y por lo tanto el argumento más débil) sustenta una conclusión más específica que la del *mejor* (más específico) argumento *c*, lo cual es contraintuitivo. Además, el contexto parece no legitimar el hecho de que John sea millonario. Por otro lado, desde un punto de vista práctico, no parece razonable defender el argumento *a* con *c*, es decir, no parece una buena idea usar como defensa para sostener *'John es millonario'* un argumento que justificadamente concluye *'John tiene al menos medio millón*'.

A partir de estos ejemplos cabe la pregunta ¿hay algo mal con el restablecimiento? Frente a esta pregunta varias respuestas han sido dadas en la literatura. Brevemente podrían nombrarse a las siguientes.

Según Horty (2001) el hecho de contar con contraejemplos es una buena razón para abandonar el restablecimiento.

Por otro lado, se han dado razones a favor de que no hay nada de malo con el restablecimiento. Esta última idea, defendida por Prakken en (2002) se justifica en los siguientes puntos: el problema no es debido al restablecimiento en sí, sino al:

i. *Lenguaje formal utilizado, como el propuesto en (Prakken y Sartor, 1997; 1998 ó Simari y Loui, 1992) que no es lo suficientemente expresivo como para bloquear los defaults que justamente generan los resultados inadecuados; y*

ii. Hay varios contextos (como el legal o moral) que parecen legitimar la utilización del restablecimiento.

Un ejemplo con respecto al último punto, debido a Prakken (2002), es el siguiente:

Ejemplo 3.4

a: El sospechoso es rubio porque el testigo α lo dice.

b: El sospechoso es morocho porque el testigo β así lo dice.

c: El sospechoso no es morocho porque el testigo γ así lo dice.

Asumiendo que γ es más fiable que β y β más fiable que α, es fácil observar que el argumento *a* parece razonablemente restablecido por *c*.

Según Prakken el inconveniente con el restablecimiento aparece cuando se modela información estadística o default (2002). Ahora bien, si el sistema es dotado de mayor expresividad, como por ejemplo, mediante el empleo de cláusulas de anormalidad, i.e. reglas que establecen explícitamente condiciones de cancelación de la aplicación de reglas default (*Por lo general A es B*) o estadísticas (*La mayoría de los A son B*), sigue diciendo Prakken, el sistema puede adecuadamente modelar los ejemplos propuestos.

Aunque Prakken da buenas razones de que los contraejemplos propuestos por Horty no son tales, su propuesta adolece de un problema práctico. Para implementar un mecanismo de cláusulas de anormalidad, cada excepción a las reglas debe ser explícitamente representada en el sistema, lo cual simplemente es difícil de hacer, imagine la lista de excepciones para la regla: *por lo general las aves vuelan*.

Alternativamente a la propuesta de Prakken, en un trabajo relativamente reciente (Bodanza y Autor, 2010) se ha indicado que el origen del comportamiento anómalo observado en los ejemplos no se debe ni al restablecimiento en sí, ni, al parecer, al lenguaje formal utilizado. En tal trabajo se sostiene que la causa se debe al hecho de que las relaciones de derrota usuales no permiten capturar adecuadamente los casos extraños. En consecuencia, los autores se proponen detectar condiciones de inhibición del restablecimiento bajo ciertas circunstancias especiales. Como resultado sugieren una relación de derrota que permite derrotar a ciertos argumentos a pesar de que no haya inconsistencia que active su aparición (como es usual en argumentación rebatible). Tal condición, denominada S1, consiste en que un argumento *a* derrotará a un argumento *b* cuando *a* sea estrictamente más específico que *b* y la conclusión de *b* juntamente con la base de conocimiento implique la conclusión de *a*.

En los ejemplos 3.2 y 3.3 es fácil observar que S1 se da entre los argumentos *c* y *a*. Por otro lado, S1 no se verifica en los ejemplos 3.1 y 3.4 como es esperable. Todo ello sugiere que la propuesta parece adecuada y por su simplicidad aventaja considerablemente a la propuesta de Prakken (2002). Al mismo tiempo que es un abordaje que contradice la sugerencia de Horty (2001).

Ahora bien y a pesar de que S1 modela adecuadamente los casos indicados y es interesante por razones que exceden al presente artículo, existen ejemplos en los que parece no comportarse correctamente, como el ejemplo 3.5.

Ejemplo 3.5
a. Ana es adinerada porque vive en Brentwood y por lo general los que viven en Brenthood son adinerados.
b. Ana no es adinerada porque renta en Brentwood y por lo general los que rentan en Brentwood no son adinerados.
c. Ana es adinerada porque es hija de millonarios y por lo general los hijos de millonarios son adinerados.

En este ejemplo, 3.5, *c* y *a* no se relacionan mediante S1 por lo que se esperaría que el restablecimiento se comportara adecuadamente. Esto es así

porque S1 cumple una función inhibitoria del restablecimiento en los casos en los que no debe aplicarse, sin embargo, no parece ser así. Las razones que llevan a creer que Ana es adinerada no tienen que ver con que ella viva en Brentwood sino en que es hija de millonarios. De modo que, en este caso, S1 no parece inhibir adecuadamente al restablecimiento dado que, intuitivamente, es más razonable estar dispuestos a aceptar o bien *c* o bien *b* pero nunca *a*.

Estando así las cosas, parecería que las únicas opciones posibles son las de o bien rechazar restablecimiento, siguiendo a Horty (2001) o bien enriquecer el lenguaje de los sistemas tal como lo propone Prakken (2002). Se ha dicho brevemente ya que ambos abordajes no son satisfactorios. Además es interesante considerar que el supuesto de ambas tesis con respecto a los ejemplos. En la tesis de Horty se supone que los ejemplos son suficientes para invalidar el restablecimiento en general, pero Horty no discute, cuestiona o limita la noción de restablecimiento en sí. Para Prakken son una excusa con vistas a explorar e identificar representaciones adecuadas de los argumentos, y el restablecimiento en sí sigue siendo incuestionable.

Para este artículo en cambio, los ejemplos son una buena excusa para cuestionar el principio del restablecimiento tal como está definido, más allá de si los ejemplos están o no bien representados, pues ello dependerá siempre del sistema que se considere.

En lo que sigue dos ideas centrales, además de la propuesta del artículo, podrán encontrarse.

La primera relacionada con un supuesto sobre el restablecimiento en los sistemas basados en argumento. Se asume que lo único importante es que cada derrotador del derrotado este a su vez derrotado, sin importar el tipo de derrota, puesto que se asume que cualquiera derrota es relevante o conveniente para restablecer un argumento. Aquí tal supuesto será cuestionado.

Lo segundo se relaciona con la propuesta de Prakken. Al respecto, una cosa más debe ser atendida. Prakken dice que los argumentos default (los ejemplos 3.2; 3.3; 3.5) o estadísticos (no tratados aquí) generan problemas al restablecimiento debido a una mala representación, pero no dice nada de cómo se restablece un argumento default o estadístico que ha sido derrotado. Aquí se hará una propuesta que podrá servir para ello.

Para instrumentar ambos aportes, se propone la noción de desactivación, que será desarrollada en la sección siguiente.

DERROTA Y DESACTIVACIÓN DE ARGUMENTOS

En los sistemas argumentativos, el restablecimiento es el principio que asegura que un argumento, a pesar de haber perdido un estado epistémico positivo puede recuperarlo (vale la pena aclarar que no se restablece un estado negativo). Lo pierde cuando es derrotado por algún argumento y lo recupera cuando para cualquier argumento que lo derrota (en un marco determinado) existe al menos uno que lo defiende. Atendiendo a este principio, es claro que los ejemplos considerados suponen un problema grave a los sistemas argumentativos,

puesto que en los ejemplos, sucede lo enunciado por el principio pero no parece que haya una restauración del estado original del argumento pretendidamente restablecido. El objetivo de la presente sección será responder a la pregunta: ¿qué debe suceder para que un argumento recupere su estado original? La respuesta estará mediada por lo que se dirá en base a la pregunta ¿Habrá alguna diferencia entre la derrota del derrotador de *a* en el ejemplo 3.1 con respecto a lo que sucede en los ejemplos 3.2, 3.3 y 3.5?

Un análisis de los ejemplos que se comportan adecuadamente permitirá aclarar la noción de restablecimiento, porque según se afirmará en este trabajo, para la recuperación de un estado epistémicamente positivo, un argumento debe contar con un tipo peculiar de defensa y la idea ampliamente aceptada

restablecer = derrotar a todos los derrotadores

debe ser precisada. La propuesta también permitirá entender por qué en los ejemplos mal comportados no se verifica el restablecimiento.

Repasando, la situación es la siguiente: para restablecer un argumento, es preciso derrotar a todos los derrotadores de un argumento. Esta idea es verificada en el ejemplo 3.1. Ahora bien, frente a los ejemplos 3.2, 3.3 y 3.5, el derrotador de un argumento se encuentra a su vez derrotado pero no parece adecuado decir que está restablecido. Será interesante saber si hay o no una diferencia relevante en el tipo de derrotas que se dan en los ejemplos que permita explicar el resultado paradójico.

¿Qué ocurre en el ejemplo 3.1? Como primer característica se tiene que el argumento *b* derrota a *a* y que el argumento *c* lleva a la no aceptación de *b*, porque si se cree en *c* no se puede creer, desde un punto de vista racional, en *b*. Esto claramente es compartido por todos los ejemplos considerados.

El tipo de derrota de *c* contra *b*, en este ejemplo, es una derrota tal que neutraliza la derrota que *b* ejerce contra *a* en el siguiente sentido: Si se acepta “la observación de que Tweety es pingüino no es confiable”, entonces no se está en condiciones de aceptar “Tweety no vuela” puesto que tal afirmación se basa en que Tweety es pingüino, y dado el conocimiento disponible en ese momento, no se sabe si Tweety es un pingüino. Si no se puede creer que Tweety es un pingüino, y no se ha desmentido que sea ave, entonces, se puede creer que Tweety vuela. Por lo tanto, la derrota de *b* contra *a* ha quedado neutralizada, sin efecto.

Lehrer y Paxson (1969) propusieron un ejemplo que se ha hecho clásico en la literatura sobre la discusión en torno a la concepción tripartita del conocimiento. Independientemente de tal planteo, aquí puede ser útil porque permite ilustrar las condiciones bajo las cuales una creencia puede ser restaurada. En el análisis del ejemplo se seguirán algunas de las ideas discutidas por Sudduth (2011) que contribuirán a lo dicho para el ejemplo 3.1.

Ejemplo 4.1

Supongamos que tengo razones para creer T: He visto a Tom Grabit robar un libro en la biblioteca por lo que puedo concluir que Tom Grabit robó un libro de la

biblioteca. Ahora supongamos que cuento con un derrotador D de T, es decir, la Señora Grabit dice que Tom está a miles de kilómetros de distancia y su hermano gemelo, que es cleptómano estaba en la biblioteca en el momento en cuestión. Si luego me entero, por su psiquiatra, que la Señora Grabit es un mentirosa compulsiva y desquiciada, y que el hermano gemelo de Tom es un invento suyo, entonces he adquirido un derrotador D para D. Mientras que D hace que T sea considerado injustificado, D* restaura mi justificación para creer en T.*

El ejemplo 4.1 es similar al ejemplo 3.1 en el sentido de que el restablecimiento funciona correctamente. Esquemáticamente, D* es un derrotador de D, y D lo es de T.

El ejemplo permite considerar en un primer momento, que se tienen razones para creer que Tom Grabit robó un libro porque se lo ha visto (T). Ahora bien, D sugiere, *prima facie,* que es falso T o que la sustentación de T no es buena, es decir, acepar D significa aceptar que, *prima facie,* "Tom robó el libro" es falso o que, *prima facie,* el observador no puede confiar en sus sentidos, o tal vez significa que se carece de información relevante, pues no se sabía que Tom tenía un hermano gemelo. Aceptar D también supone aceptar, *prima facie,* que la Señora Grabit es mentalmente sana y que su testimonio es confiable.

Sin embargo, cabe señalar que D es un derrotador engañoso o aparente dado que, a la luz de la evidencia total disponible (al considerar toda la información del ejemplo) no podría haberse considerado a D como un derrotador de T. Esto puede saberse gracias a la función del argumento D*. El rol del argumento D* es mostrar que D es un derrotador engañoso, i.e. un derrotador que presupone, sugiere o depende de suposiciones falsas y en consecuencia, neutralizar a D, dejarlo sin efecto. En este caso, por ejemplo se supuso que la madre de Tom decía la verdad, que Tom tiene un hermano gemelo, que el agente que vio a Tom no tuvo en cuenta información relevante, etc. Esta situación permite restaurar a T (y también restaura creencias implícitas como que el agente tuvo en cuenta información relevante, que puede confiar en sus sentidos, etc.), obviamente dependerá que D* sea un derrotador genuino, es decir, que tal derrotador no presuponga, sugiera o dependa de alguna suposición falsa (Sudduth, 2011) y lo será mientras el marco de la información disponible se conserve.

Ambos ejemplos permiten concluir que para que un argumento derrotado sea restaurado a su estado original es necesario que todos sus derrotadores estén desactivados, i.e. se cuente con información que lleve al rechazo de los supuestos en los que estos se basan; información que permite entender que el argumento rival del que se defiende se basa en información falsa, incorrecta o implausible.

En los ejemplos, 3.2 y 3.3 la cuestión es notablemente diferente a lo dicho recién. La derrota de *c* contra *b* lleva a la no aceptación de *b*. Pero el argumento *c* no neutraliza la derrota que *b* ejerce contra *a*.

El hecho de que Tweety sea un pingüino (ejemplo 3.2) es una buena razón para inhibir al argumento *a,* pues se basa en que las aves normales vuelan, y se cuenta con una evidencia que afirma que no es el caso (pues es un ave

excepcional). Ahora bien, se podría restaurar la creencia original en el argumento *a,* según lo dicho recientemente, si se contara con una pieza de información que señalara que "Tweety es pingüino" es falso o improbable. El problema aquí es que efectivamente, *Tweety es pingüino,* y más aún, tal información es consistente con el argumento *c.* En ningún caso, el argumento *c,* señala que algún supuesto en el que *b* está basado, sea falso, incorrecto o implausible. Nuevamente ¿Qué debería suceder para poder restaurar el argumento *a*? Debería saberse que Tweety no es pingüino.

En el ejemplo 3.3 la cuestión es similar al 3.2. El argumento *a*, "*John es millonario porque es empleado de Microsoft y por lo general tienden a ser millonarios*", se encuentra derrotado por el argumento *b* pero no puede considerarse restablecido por *c* porque para restaurar el estado original hace falta que John no sea un novel empleado de Microsoft o saber que los nuevos empleados de Microsoft también son millonarios. Sin embargo, el argumento *c* es consistente con la afirmación *"novel empleado de Microsoft"* y no permite saber si "*los nuevos empleados de Microsoft también son millonarios".*

Unas figuras ilustrarán la idea o el efecto de desactivar un argumento. Cuando un derrotador desactiva a un argumento deja sin efecto no solo a la *prueba* para la conclusión que sustenta sino también a las derrotas ejercidas por este (cuando corresponda). Por otro lado, cuando un argumento derrota a otro, este lleva a la no aceptación de la conclusión sustentada por este pero no las derrotas ejercidas, estas no son neutralizadas y permanecen activas.

En estas figuras los argumentos serán representados por letras, y las derrotas por flechas. El ejemplo 3.1 puede ser representado como en la Figura 1.

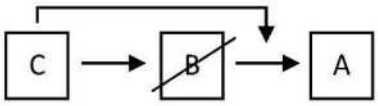

Figura 1: *b* derrota a *a* y *c* desactiva a *b.*

Por otro lado, en los ejemplo 3.2 y 3.3 las relaciones son de tal modo que el argumento *a* no debe considerarse aceptado. El motivo, la derrota de *b* contra *a* está activa.

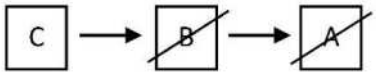

Figura 2: derrota sin defensa.

El ejemplo 3.5 puede esquemáticamente representarse de la siguiente manera (Figura 3), obviamente, dependerá de si se adopta una teoría crédula (Figura 3) o escéptica (Figura 4) de base:

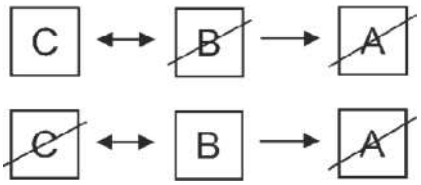

Figura 3. Derrota sin defensa para el ejemplo 3.5 bajo una teoría crédula.

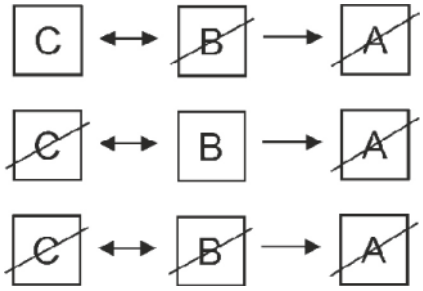

Figura 4. Derrota sin defensa para el ejemplo 3.5. bajo una teoría escéptica.

En esta sección se mostró que para que haya restablecimiento se debe esencialmente contar con piezas de información que revelan que los argumentos rivales del defendido presuponen, sugieren o dependen de información falsa o incorrecta. Cuando un argumento que es derrotador de otro, se encuentra a su vez derrotado y se hace patente alguna de las suposiciones falsas que lleva la aceptación de tal argumento, el restablecimiento es natural e intuitivo, cuando se dan situaciones como en el ejemplo 3.2, 3.3 y 3.5 no es adecuado restablecer a los argumentos puesto que las derrotas ejercidas por los rivales permanecen activas debido a que son derrotadores genuinos.

En suma, si se modelan razonamientos como se hace en (Prakken y Sartor, 1997; 1998; Simari y Loui, 1992; García y Simari, 2004; Dung y Son, 2000) y se mantiene la noción de restablecimiento como derrota de derrotadores el problema será endémico. Las nociones de desactivación y derrota permiten entender cuándo un argumento puede considerarse restablecido y cuando no. Como se notará, no todas las derrotas son equivalentes. Hay derrotas que defienden y otras que no lo hacen. Respondiendo a la pregunta que abrió esta sección: ¿qué debe suceder para que un argumento recupere su estado original (sea o no default o estadístico)? Se debe decir que es preciso contar con información que permita saber que el argumento rival del argumento que se pretende defender se basa en información falsa, incorrecta o improbable. Haciendo más claro lo dicho, dos nociones han sido propuestas, la primera, *la desactivación,* la segunda, una nueva versión del *restablecimiento*.

- Un argumento *a desactiva* a otro *b* si y sólo si aceptar *a* lleva a aceptar que *b* se encuentra basado en información falsa, incorrecta o improbable (atendiendo a la información total disponible).

- Un argumento *a* se dice *restablecido* si y solo si para todo argumento que lo derrota existe al menos un argumento que desactiva a los derrotadores (atendiendo a la información total disponible).

CONCLUSIÓN

Los sistemas argumentativos seleccionan a los *buenos* argumentos mediante un proceso de comparación y evaluación. Todos los argumentos que pueden construirse a partir de un conjunto de información determinado son sometidos a una suerte de *combate* donde los argumentos se comparan y se seleccionan según ciertos criterios. Aquellos que prevalecen frente a sus rivales son los ganadores y se entiende que ello alcanza para considerarlos como argumentos justificados. Una noción fundamental involucrada en tal proceso es la del restablecimiento. Brevemente, si un argumento *'a'* es derrotado por *'b',* y a su vez, *'c'* derrota a *'b',* se dirá que *'a'* es restablecido por *'c'*.

En la literatura se ha discutido la validez del restablecimiento a través de la propuesta de una serie de ejemplos que muestran que el principio no parece funcionar adecuadamente. Estos ejemplos han sido analizados en el artículo con vistas a poder determinar las causas de tal situación. Como resultado, se ha propuesto que el inconveniente aparece por no distinguir dos tipos de derrota: las que defienden y las que no lo hacen. A las primeras se las llamó *desactivación*.

Un argumento *a* desactiva a otro *b* cuando aceptar al argumento *a* lleva a aceptar que *b* depende de información, falsa, incorrecta o improbable. Si *b* está basado en información de ese tipo entonces:

i. no sólo es incapaz de brindar una adecuada prueba a la conclusión que sustenta, sino que

ii. no puede considerarse que las derrotas ejercidas por este tengan el efecto de invalidar a otros argumentos.

Cuando un argumento que es derrotador de otro, es derrotado y se hace patente alguna de las suposiciones falsas que lleva la aceptación de tal argumento, aparece el restablecimiento de los argumentos derrotados por aquel. Por otro lado, si un argumento *a* derrota a *b,* pero no lo desactiva, entonces, el argumento *b* se encuentra sustentado en información, cierta, probable y correcta. Aceptar *a* lleva a no elegir *b,* pero el efecto de su derrota permanece activo.

El problema restante consiste en definir ambos tipos de derrotas de manera formal y aplicarla en sistemas específicos, como también estudiar una manera de conjugarlas con el proceso de justificación que usualmente se basa sólo en la noción de derrota.

BIBLIOGRAFÍA

ALESSIO, C (2015) Restablecimiento y especificidad en sistemas argumentativos. Disertación Doctoral. Universidad Nacional del Sur. Bahía Blanca-Argentina.

ALESSIO, C (2013) Un sistema argumentativo para razonamiento default: discusión

y propuesta. En Algañaraz V et al. (comp.) II Encuentro de Jóvenes Investigadores: consolidando espacios del quehacer científico en San Juan (Sd.)

ALESSIO, C (2012) Una propuesta para modelar razonamiento default en sistemas argumentativos. En Milone, R & Torres, J Ciencia y Metafísica: Selección de Trabajos (pp.19-25). Mendoza, Argentina: FFYL-UNCuyo.

ALESSIO, C (2012) Aceptabilidad extendida para marcos argumentativos abstractos. Epistemología e Historia de la Ciencia, 18: Sd.

ALESSIO, C (2010) Extensiones no justificadas en Lógica Default. Epistemología e historia de la ciencia, 16: 13-19.

ALESSIO, C (2009) Razonamiento rebatible: origen, historia, estrategias. Revista Cuadernos de la Universidad, 42: 15-35.

ANTONIOU, G (2006) Defeasible reasoning: A discussion of some intuitions. International Journal of Intelligent Systems, 21 (6): 545-558.

BODANZA, G (2015) La argumentación abstracta en inteligencia artificial. Problemas de interpretación y adecuación de las semánticas para la toma de decisiones. Theoria, 30: 395-414.

BODANZA, G (2011) Dudas razonables sobre el restablecimiento como principio de la argumentación rebatible. Epistemología e Historia de la Ciencia, 17: 94 – 101

BODANZA, G & ALESSIO, C (2010) Sobre la aceptabilidad de argumentos en un marco argumentativo con especificidad. En Santibañez, C (Ed.) Actas de la II Conferencia Internacional Lógica, Argumentación y Pensamiento Crítico (pp. 74-81). CEAR. Santiago, Chile.

BODANZA, G & ALESSIO, C (2014) Reinstatement and the Requirement of Maximal Specificity in Argument Systems. Logic, Language, Information, and Computation, 8652: 81-93.

CAMINADA, M (2004) For the sake of the Argument. Explorations into argument-based reasoning. Doctoral dissertation: Free University Amsterdam.

CHESÑEVAR, C, MAGUITMAN, A & LOUI, R (2000) Logical models of argument. ACM Computing Surveys, 32 (4): 337-383.

DIMOPOULOS, Y, MORAITIS, P & AMGOUD, L (2009) Extending Argumentation to Make Good Decisions. En Rossi, E; Tsoukias, F and Alexis, R (Eds.) Algorithmic Decision Theory (pp. 225-236). Berlin: Springer Berlin Heidelberg

DUNG, P (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.

DUNG, P & SON, T (1995) Nonmonotonic inheritance, argumentation and logic programming. En Wiktor Marek, V, Nerode, A and Truszczyński, M (Eds) Logic Programming and Nonmonotonic Reasoning (pp. 316-329). Berlin: Springer Berlin Heidelberg

DUNG, P & SON, T (2000) Default reasoning with specificity. En John Lloyd et al. (Eds) Computational Logic—CL 2000 (pp. 792-806). Berlin: Springer Berlin Heidelberg.

DUNG, P & SON, T (2001) An argument-based approach to reasoning with specificity. Artificial Intelligence, 133 (1-2): 35–85.

GARCÍA, A & SIMARI, G (2004) Defeasible Logic Programming: An argumentative approach. Journal of Theory and Practice of Logic Programming, 4 (1-2): 95-138.

HORTY, J (1994) Some Direct Theories of Nonmonotonic Inheritance. In: Gabbay, D, Hobber, C and Robinson, J (Eds.) Handbook of Logic in Artificial Intelligence and Logic Programming. Nonmonotonic Reasoning and Uncertain Reasoning, vol. 3 (pp. 111–187). Oxford: Oxford University Press.

HORTY, J (2001) Argument construction and reinstatement in logics for defeasible reasoning. Artificial Intelligence and Law, 9 (1): 1-28.

HORTY, J (2012) Reasons as Defaults. Oxford: Oxford University Press.

HORTY, J Thomason, R & Touretzky, D (1990) A skeptical theory of inheritance in nonmonotonic semantic networks. Artificial intelligence, 42 (2): 311-348.

KACI, S, VAN DER TORRE, L and WEYDERT, E (2006) Acyclic Argumentation: Attack = Conflict +Preference. In Brewka, G; Coradeschi, S; Perini, A and Traverso, P (Eds) Proceedings of the 17th European Conference on Artificial Intelligence, (pp. 725-26). Amsterdam: IOS Press

KOWALSKI, R & TONI, F (1996) Abstract argumentation. Artificial intelligence and law, 4 (3-4): 275-296.

LEHRER, K & PAXSON, T (1969) Knowledge: Undefeated Justified True Belief. Journal of Philosophy, 66: 225-37.

LOUI, R (1987) Defeat among arguments: a system of defeasible inference. Computational intelligence, 3 (1): 100-106.

LOUI, R & STIEFVATER, K (1992) Corrigenda to Poole's rules and a lemma of Simari-Loui, Department of Computer Science, Washington University, St. Louis.

POLLOCK, J (1987) Defeasible reasoning. Cognitive science, 11 (4): 481-518.

POLLOCK, J (1995). Cognitive carpentry: A blueprint for how to build a person. Cambridge, Massachusetts: MIT Press.

PRAKKEN, H & SARTOR, G (1996[a]) A dialectical model of assessing conflicting arguments in legal reasoning. Artificial Intelligence and Law, 4: 331-368.

PRAKKEN, H & SARTOR, G (1996[b]) A system for defeasible argumentation, with defeasible priorities. En Gabbay, D & Ohlbach, H (Eds.) Practical Reasoning, International Conference on Formal and Applied Practical Reasoning (pp. 510-524). Berlin: Springer Berlin Heidelberg.

PRAKKEN, H & SARTOR, G (1997) Argument-based extended logic programming with defeasible priorities. Journal of applied non-classical logics, 7 (1-2): 25-75.

PRAKKEN, H & SARTOR, G (1998) Modelling reasoning with precedents in a formal dialogue game. Artificial Intelligence and Law, 6: 231–287.

PRAKKEN, H (2002) Intuitions and the Modelling of Defeasible Reasoning: Some Case Studies. En Benferhat, S and Giunchiglia, E (eds.) Proceeding of 9th International Workshop on Non-Monotonic Reasoning (pp. 91-102). Tolouse, France.

PRAKKEN, H & VREESWIJK, G (2000) Logics for defeasible argumentation. En Gabbay, D & Franz, G (Eds.) Handbook of philosophical logic, Vol. IV (pp. 219-318). Berlin: Springer.

SIMARI, G & LOUI, R (1992) A mathematical treatment of defeasible reasoning and its implementation. Artificial intelligence, 53 (2): 125-157.

SUDDUTH, M (2008) Defeaters in Epistemology En Dowden, B (Ed.) Internet Encyclopedia of Philosophy. http://www.iep.utm.edu/ep-defea/ Recuperado el 10/03/2015.