UNIVERSIDAD DE BUENOS AIRES
Facultad de Ciencias Exactas y Naturales
Departamento de Matemática

Tesis de Licenciatura

SÓLO SÉ QUE NO SÉ NADA

Lucila Schmidt

Director: Santiago Figueira

2 de marzo 2015

Gracias totales

Primeramente agradezco a mis padres por darme la oportunidad de estudiar aún cuando la situación económica era desfavorable, y por nunca cuestionarme el camino que elegí. También a mis hermanos, y a Natalia que fue una gran incorporación para la familia. A Amadeo que no lo conozco pero ya me cae bien. A mis tíos y primos, que siempre estuvieron presentes aún en la distancia. Agradezco a mis amigas de toda la vida Dana, Catalina y Carolina por acompañarme desde siempre hasta que me mudé del otro lado de la General Paz. Gracias a Magdalena, en quien redescubrí una gran amiga (pero que no le gusta mi vestidito), y a Lupe y More, a quienes les perdí el rastro pero tengo grandes recuerdos. A Facundo, gracias a quien pude conocer mi tema de tesis, a través de quien conocí a mi director y quien me acompañó en mi desarrollo académico, profesional y personal. También debo agradecer con mucho énfasis *a Santiago* que me tuvo paciencia infinita con mis caprichos epistémicos. Aunque la UBA no vaya a leer esto: Gracias a la UBA por brindar educación pública, gratuita y de calidad sin la cual mucha gente no podría explotar su potencial, por el mero hecho de no pertenecer a una determinada clase social.

Ahora quizás voy a decepcionar, con una sola frase, a muchos (o no) que esperan encontrar sus nombres: Gracias a la inmensa cantidad de gente copada que conocí en la Facultad. Si usted pretendía encontrar su nombre en esta sección, dese por agradecido. Pero este agradecimiento es genuino: el conjunto de la gente que va a leer esto está incluída en el de la gente que tengo algo para agradecerle. Así que sepa usted que le estoy agradecida **de verdad** aunque no figure su nombre. Si, aún así, usted no se encuentra conforme puede completar aquí con su nombre y apellido: muchas gracias a ..... y presentar este papel en la Secretaría de Agradecimientos los sábados de 8.30 a 9.10 o los lunes de 23.30 a 23.40 para legalizar el agradecimiento.

No puedo dejar de agradecer a la gente de Mercap donde encontré el grupo de gente más increíble de la historia, y a la gente de Despegar con quienes puedo seguir aprendiendo día a día. Gracias a todas las personas que dejaron huellas en mí. Gracias Piero. Gracias a mis profesores de baile, que me enseñaron a descubrir que el cuerpo no es el envoltorio del cerebro. Gracias a la música por darme lo que me faltaba.

Índice general

Capítulo 1

Introducción

> Todo hombre, por naturaleza,
> desea saber.
>
> Aristóteles.

Desde el origen de la especie, el conocimiento jugó un rol importante en el desarrollo de las civilizaciones. Es por este carácter ancestral e inherente a la existencia del hombre, que delimitar qué es el conocimiento con una definición que lo caracterice apropiadamente se convierte en una tarea ardua. Esta es la pregunta fundacional de la epistemología.

Las primeras consideraciones y problemas acerca de este tópico se remontan a los orígenes de la filosofía. Se considera como pionero en el debate epistémológico a Platón, quien aborda la temática en sus textos Teeteto, Menón y La República. Fue Platón quien definió por primera vez el conocimiento como creencia verdadera y justificada. Esta definición fue el punto de partida para filósofos y epistemólogos a lo largo de la historia e incluso hoy sigue siendo motivación de distintos estudios y textos filosóficos.

En esta tesis estudiaremos la lógica epistémica: una herramienta que nació para formalizar el concepto de conocimiento y poder razonar acerca él, ofreciendo una aproximación lógica al problema central de la epistemología. Si bien vamos a tratar el enfoque moderno de la lógica epistémica, cabe destacar que tiene sus orígenes en la Antigua Grecia, cuando Aristóteles comienza a observar que frases como "Yo sé que..." tienen propiedades sistemáticas que las hacen sometibles a un estudio formal. Posteriormente, en la Edad Media, Buridan, Duns Scotus y Ockham extendieron las ideas aristotélicas de una manera muy similar al enfoque lógico actual.

La lógica epistémica formalizada como la conocemos hoy surgió como un caso particular de la lógica modal a mediados del siglo XX y continuó su evolución como un área de estudio en sí misma. Por otro lado, desde los textos platónicos, la filosofía ha intentado determinar puntos en común y diferencias entre las creencias y los conocimientos. Por este motivo, las mismas motivaciones filosóficas que inspiraron la creación de la lógica epistémica también dieron lugar a la lógica doxástica, la cual sirve para razonar acerca de las creencias.

Si bien las lógicas epistémicas y doxásticas surgieron por motivos meramente filosóficos, a partir de la década del 80 (cuando investigadores de Inteligencia Artificial comenzaron a

interesarse en formalismos y lógicas como recursos para su área) fueron resignificadas como mecanismos especialmente útiles para las Ciencias de la Computación. Además, dado que las lógicas epistémicas y doxásticas son útiles para describir contextos de información, también han ganado terreno en disciplinas como Teoría de Juegos y Economía, entre otras.

Cada una de estas áreas tiene sus propias necesidades de manejo de información, y en consecuencia surgen distintas nociones de conocimiento que tienen diferentes enfoques y matices. En el Capítulo 3 vamos a estudiar algunas de estas nociones que contribuyeron al crecimiento de las lógicas epistémicas y doxásticas, a través de varias lógicas con operadores que formalizan distintos tipos de conocimiento y creencia.

Pero estudiar los conocimientos y las creencias de manera estática, es poco realista: la información puede cambiar, especialmente en contextos en los que varios sujetos o computadoras, interactúan intercambiando información. Por ejemplo: un partido de Clue, un partido de chinchón, etcétera. En general hay especial interés en desarrollar teorías acerca de cómo y por qué cambian los conocimientos. Este es probablemente el aspecto más rico de las lógicas epistémicas y doxásticas, dando lugar a la lógica epistémica dinámica y a la teoría de revisión, que serán estudiadas en los Capítulos 4 y 5 respectivamente. Estas áreas son de particular interés para la Teoría de Juegos por su capacidad para modelar tanto juegos de información perfecta como de información imperfecta.

Por ejemplo, un juego clásico que puede ser estudiado con lógicas epistémicas dinámicas es de los niños embarrados: un padre que tiene k hijos, algunos de ellos tienen barro en la frente. Ninguno sabe si está embarrado o no porque no pueden ver sus propias frentes, sin embargo todos pueden ver las frentes ajenas. El padre dice en voz alta que al menos uno de ellos tiene la frente embarrada. Y luego ordena: "El que en este momento sepa que tiene su frente embarrada que de un paso al frente ya mismo. El resto que permanezca en su lugar". Mientras nadie de un paso al frente, el padre seguirá repitiendo el anuncio. ¿Cómo hacen los niños para saber si están embarrados o no?

En esta tesis nos centraremos mayormente en los cambios de información debido a comunicación. La principal característica de la comunicación es que no produce efectos en la realidad, sino que afecta únicamente a la cantidad de información que se tiene sobre ella. Es decir, no cambian los hechos sino lo que se sabe o lo que se cree sobre ellos.

Analizar los cambios en el conocimiento producidos por comunicaciones no se limita a recibir información e incorporarla. En muchos contextos comunicativos, existen resultados poco intuitivos, como lo es el caso en el que aprender una cierta información la convierte en falsa. Por ejemplo: Una mujer le informa a un hombre "estás borracho y creés que no lo estás". Luego de recibir la información transmitida, ¿puede el hombre seguir creyendo que no está borracho? Lo esperable sería que no, sin embargo el mensaje afirma que sí lo cree. Por lo tanto al recibir el mensaje, este último se convierte en falso por el mero hecho de ser recibido.

Pero además de la veracidad de la información, en los contextos comunicativos también son de importancia otros factores. Uno de los más influyentes es la actitud del receptor hacia el emisor. Es esperable que el receptor no incorpore la información de un mensaje del mismo modo si proviene de una fuente infalible (por ejemplo: sus propios ojos) que

si proviene de una fuente que no sólo es falible, sino que además es desconfiable (por ejemplo: la revista Barcelona). Por estos motivos es necesario formalizar distintos grados de confianza entre los participantes de una comunicación.

Pero si un participante de una comunicación sabe cuánto confían en él los otros, puede utilizar esta información para que los mensajes que transmita produzcan los efectos deseados. Un emisor puede querer confundir a un receptor, comunicando un mensaje que sabe que es falso, sabiendo que el efecto en el receptor será que creer fielmente lo que dijo. Por ejemplo: un par de hermanos quieren utilizar un mismo juguete al mismo tiempo y el mayor le miente al menor diciendo "te llama papá", para quedarse con el juguete. Esto representaría un ejemplo de una comunicación deshonesta: una comunicación en la que el emisor sabe que el receptor va a creer la información del mensaje, mientras que el emisor no.

Por otro lado, también cabe la posibilidad de que un emisor conociendo cuáles son los efectos que producen sus mensajes en un receptor sepa más fácilmente qué decirle para convencerlo de pensar sobre un asunto φ lo mismo que él. Puede saber que anunciando ψ, que no necesariamente es igual a φ, logrará que el receptor tome la misma postura que él y usar esto a su favor.

Un ejemplo cotidiano es la "psicología inversa": una madre considera que una buena opción para colegio secundario de su hijo adolescente (que siempre le lleva la contra) es un colegio técnico. Entonces decide decirle "Los colegios técnicos no son una buena opción para vos". Esto es lo que llamaremos un anuncio persuasivo: la madre logra que el hijo piense como ella, a pesar de que para lograrlo tuvo que anunciarle otra cosa. Esto no es un anuncio deshonesto: ya que el hijo comenzará a creer que los colegios técnicos son una buena opción para él, y la madre también lo cree. Sin embargo, la madre no cree en el mensaje que transmite. Cuando esto sucede se considera que la comunicación no es sincera.

En el Capítulo 5 veremos formalizaciones de la confianza, sinceridad, honestidad y persuasión, conceptos cotidianos relacionados con las actitudes hacia la información recibida, emitida, su emisor o su receptor. Finalmente vamos a estudiar de qué manera se pueden fusionar creencias de un grupo de modo que finalmente todos los participantes compartan creencias, a través de mensajes honestos, sinceros y persuasivos. Vamos a estudiarlo tanto de un modo estático, como de un modo dinámico. La motivación para estudiar estas fusiones proviene de Social Choice Theory, donde también se encuentran aplicaciones de la lógicas epistémicas y doxásticas.

Por la cantidad y heterogeneidad de aplicaciones, la lógica epistémica se encuentra en una etapa de desarrollo fuerte. Trabajos recientes incluyen herramientas topológicas y probabilísticas para los modelos epistémicos y doxásticos, enriqueciendo el contenido matemático detrás del conceptual. El objetivo de esta tesis es presentar un panorama general de la lógicas epistémicas y doxásticas, así que por motivos cohesivos se incluyeron únicamente los principales enfoques, pero alentamos al lector interesado a explorar por su cuenta estas herramientas útiles aplicables a múltiples contextos. Sin más, nos adentramos en el Capítulo 2 donde se describirán las herramientas necesarias para la comprensión del resto de la tesis.

Capítulo 2

Preliminares y Lógica Modal

> All men are mortal. Socrates was mortal. Therefore, all men are Socrates.
>
> Woody Allen.

En este capítulo se van a mencionar conceptos básicos de lógica modal a modo introductorio, siguiendo los libros [12] y [13]. Algunos de ellos serán usados posteriormente y otros se mencionan para dar un panorama general de estas lógicas que tienen como caso particular a la lógica epistémica.

La lógica modal es una variación de la lógica proposicional clásica que incorpora operadores los cuales también son llamados modalidades. El nombre proviene de que históricamente se usaron estos nuevos operadores para expresar modalidades lingüísticas. Formalmente estas son un tipo de fuerzas ilocucionarias donde el emisor expresa un grado de compromiso con respecto a la credibilidad, obligatoriedad, deseabilidad o realidad de la proposición que transmite en el lenguaje, y generalmente son señaladas por verbos modales: tener, deber, querer, poder, etcétera. Si bien comprender cabalmente estos conceptos lingüísticos no es necesario para la lectura de esta tesis, referimos al lector interesado al libro [21] donde puede profundizar un poco esta noción aún sin conocimientos previos de lingüística.

La lógica modal fue estudiada desde la Antigua Grecia, en particular por Aristóteles en De Interpretatione donde por primera vez se razonó acerca de la relación entre necesariedad y posibilidad. Desde ese entonces muchos filósofos abordaron el tema en mayor o menor profundidad, entre ellos Diodorus, Boecio, Kant, Leibnitz, Lewis, Gödel, entre otros. De acuerdo con [15], quienes más han contribuído en la concepción actual de la lógica modal fueron MacColl, quien incorporó el análisis simbólico, Carnap, que desarrolló una semántica basada en el concepto de Leibnitz de posibibilidad, y Kripke, creador de la semántica de mundos relativos que actualmente se usa.

A continuación se encuentra una breve reseña de la lógica modal proposicional que será el puntapié inicial para el estudio de la lógica epistémica.

2.1. Lenguajes modales

El lenguaje que usa la lógica modal no es otro que el lenguaje proposicional munido de nuevos operadores que serán llamados *operadores modales o modalidades*. Estos lenguajes, van a ser herramientas útiles para describir y razonar acerca de conjuntos en los que hay definidas una o más relaciones, cuya información será accedida a través de los operadores modales.

Definición 2.1.1. El *lenguaje modal básico* se define usando un conjunto de variables proposicionales Φ, los operadores lógicos de la lógica proposicional y un operador modal unario $\Diamond$. Las fórmulas bien formadas de este lenguaje son

$$\varphi ::= p \mid \bot \mid \neg\varphi \mid \psi \vee \varphi \mid \Diamond\varphi$$

donde p es alguna variable proposicional (i.e.: $p \in \Phi$).

Esto significa que una fórmula bien formada es o bien una variable proposicional, o la constante proposicional false, o una fórmula bien formada negada, o una disyunción de fórmulas bien formadas, o una fórmula bien formada precedida por un diamante.

Llamaremos *cuantificador modal existencial* al operador $\Diamond$.

Así como en la lógica de primer orden los cuantificadores universal y existencial son duales el uno del otro, en la lógica modal básica tenemos el operador $\Box$ que queda definido por $\Box\varphi := \neg\Diamond\neg\varphi$ al que llamaremos *cuantificador universal modal*.[1]. Además definimos también $\top := \neg\bot$.

Hay muchas lecturas posibles de los operadores modales. Una de las más tradicionales es interpretar $\Diamond\varphi$ como "es posible que valga φ". Bajo esta interpretación $\Box\varphi$ significa "no es posible que no φ", es decir, "necesariamente vale φ". Con este lenguaje entonces podríamos formular proposiciones como $\Box\varphi \to \Diamond\varphi$ (cualquier cosa que necesariamente vale, entonces es posible que valga), o $\varphi \to \Diamond\varphi$ (si algo es verdadero entonces es posible), que intuitivamente parecen verdaderos para cualquier instanciación posible. Pero al tratarse simplemente de un lenguaje, sin semántica ni axiomas y reglas de deducción, estas fórmulas en sí no tienen valor de verdad o validez. Si consideramos que esos son principios correctos para razonar acerca de la posibilidad y la necesariedad, entonces al momento de axiomatizar la lógica o de dar una semántica al lenguaje debemos tomar alguna precaución para que los símbolos y sus interpretaciones cumplan las propiedades que intuitivamente parecen válidas.

La generalización del lenguaje modal básico es de la manera más esperable: admitiendo múltiples operadores modales, y permitiéndoles tener distintas aridades.

Definición 2.1.2. Un *tipo de similaridad modal* es un par $\tau = (O, \rho)$, donde O es un conjunto finito no vacío y ρ una función $O \to \mathbb{N}$. Los elementos de O serán llamados

[1] La elección de definir el lenguaje con $\Diamond$ como operador primitivo y definir $\Box$ como una abreviación o al revés depende del contexto. Los modelos epistémicos suelen definirse con operador primitivo de tipo $\Box$, pero por motivos didácticos se eligió introducir el lenguaje con cuantificador existencial modal como operador primitivo.

operadores modales y usaremos $\Diamond_0$, $\Diamond_1$,... para denotar los elementos de O. La función ρ asigna a cada operador $\Diamond \in O$ una aridad finita (es decir, define el número de argumentos que precisa $\Diamond$).

Asumimos que la aridad de los operadores es sabida y no distinguimos entre τ y O.

Definición 2.1.3. Un *lenguaje modal* se construye a partir de un tipo de similaridad modal τ y un conjunto de variables Φ donde las fórmulas bien formadas están dadas por:

$$\varphi ::= p \mid \bot \mid \neg\varphi \mid \psi \vee \varphi \mid \Diamond(\varphi_1, \ldots, \varphi_{\rho(\Diamond)})$$

donde $p \in \Phi$ y $\Diamond \in O$.

Definición 2.1.4. Para cada operador $\Diamond \in O$ definimos $\Box$ *el operador dual de* $\Diamond$ como $\Box(\varphi_1, \ldots, \varphi_n) := \neg\Diamond(\neg\varphi_1, \ldots, \neg\varphi_n)$.

Ejemplo 2.1.5. El *lenguaje temporal básico* se construye a partir del conjunto de operadores unarios $O = \{\Diamond_F, \Diamond_P\}$ aunque usualmente se nota F en lugar de $\Diamond_F$ y P en lugar de $\Diamond_P$. La interpretación intuitiva de la fórmula $F\varphi$ es "φ será verdad en algún momento futuro", y de la fórmula $P\varphi$ es "φ fue verdad en algún momento pasado". Los duales de F y P suelen ser notados G y H respectivamente, y sus interpretaciones intuitivas son "φ será verdad en todo momendo futuro" y "φ fue verdad en todo momendo pasado". Este lenguaje está diseñado para expresar distintas afirmaciones acerca del tiempo, por ejemplo, $P\varphi \to GP\varphi$ ("si algo pasó, siempre va a haber pasado") o $F\varphi \to FF\varphi$ ("si algo pasará, en el algún momento futuro va a ser cierto que pasará").

2.2. Frames y modelos

En la sección anterior se mencionaron los términos "interpretar", "verdadero", y otros que no fueron definidos. Empezaremos a echar un poco de luz mediante algunas definiciones y de este modo comenzaremos a formalizar y ponerle contenido matemático a nuestro tema de estudio.

Definición 2.2.1. Un *frame* para el lenguaje modal básico es un par $\mathfrak{F} = \langle W, R\rangle$ tal que

1. W es no vacío
2. R es una relación binaria en W

Nos referiremos a los elementos de W de distintos modos, usualmente *estados, puntos o mundos*, y llamaremos a W el dominio de $\mathfrak{F}$ que notaremos $Dom(\mathfrak{F})$.

Si en un frame $\mathfrak{F}$ queremos distinguir un punto w, definimos un *frame punteado* que consiste en un par $(\mathfrak{F}, w)$.

Por lo tanto, un frame es un objeto que describe cómo se relacionan los distintos mundos de W, lo cuál es completamente estructural, es decir, semántico. Pero aún no tenemos la información contingente, esta información viene de la mano de los modelos.

Definición 2.2.2. Un *modelo de Kripke (o simplemente modelo)* para el lenguaje modal básico es un par $\mathfrak{M} = \langle \mathfrak{F}, \mathtt{i} \rangle$ donde $\mathfrak{F}$ es un frame para el lenguaje modal básico e $\mathtt{i}$ es una función que asigna a cada variable proposicional $p \in \Phi$ un conjunto $\mathtt{i}(p) \subseteq W$. Informalmente pensamos en $\mathtt{i}(p)$ como el conjunto de puntos en los que p es verdadero.

Llamaremos *dominio* al dominio de $\mathfrak{F}$ y lo notaremos como $Dom(\mathfrak{M})$, y en algunas ocasiones notaremos $w \in \mathfrak{M}$ en vez de $w \in Dom(\mathfrak{M})$. También nos referimos a la función $\mathtt{i}$ como *interpretación o valuación* y a $\mathfrak{F}$ como el *frame subyacente de* $\mathfrak{M}$.

Si el frame es punteado obtenemos un *modelo punteado* que notaremos $(\mathfrak{M}, w)$.

Ejemplo 2.2.3. A continuación vemos un ejemplo de un modelo de Kripke para un lenguaje modal básico con dos variables proposicionales p y q. En el diagrama los nodos representan los estados del modelo, su contenido indica si valen las variables proposicionales en el estado, y las flechas representan la relación del modelo. El modelo es punteado, y el estado pintado de turquesa representa el estado distinguido.

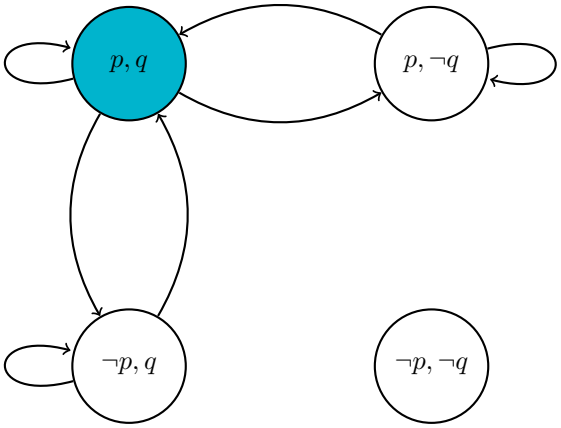


Figura 2.1: Ejemplo de modelo de Kripke para un lenguaje modal básico.

Ahora estamos en condiciones de dar una definición de verdad de una fórmula en un punto de un modelo.

Definición 2.2.4. Sea w un estado de un modelo $\mathfrak{M} = \langle W, R, \mathtt{i} \rangle$ definimos inductivamente la noción de que una fórmula sea satisfecha (o verdadera) en w para el modelo $\mathfrak{M}$ del siguiente modo:

- $w \Vdash_{\mathfrak{M}} p$ sii $w \in \mathtt{i}(p)$ donde $p \in \Phi$
- $w \Vdash_{\mathfrak{M}} \bot$ nunca
- $w \Vdash_{\mathfrak{M}} \neg\varphi$ sii no $w \Vdash_{\mathfrak{M}} \varphi$
- $w \Vdash_{\mathfrak{M}} \varphi \vee \psi$ sii $w \Vdash_{\mathfrak{M}} \varphi$ o $w \Vdash_{\mathfrak{M}} \psi$

- $w \Vdash_{\mathfrak{M}} \Diamond\varphi$ sii para algún $v \in W$ con wRv vale $v \Vdash_{\mathfrak{M}} \varphi$

Se sigue de la definición anterior que $\mathfrak{M}, w \Vdash \Box\varphi$ si y sólo si para todo $v \in W$ tal que wRv se tiene que $\mathfrak{M}, v \Vdash \varphi$.

Nota. Notaremos $\mathcal{P}(X)$ al conjunto de partes de X.

Definición 2.2.5. Dado un modelo $\mathfrak{M} = \langle \mathfrak{F}, \mathtt{i} \rangle$, definimos la *extensión de la interpretación* $\mathtt{i}$ como $\bar{\mathtt{i}} : Form \to \mathcal{P}(W)$ como $\bar{\mathtt{i}}(\varphi) := \{w \in W \ : \ w \Vdash_{\mathfrak{M}} \varphi\}$.

Observación 2.2.6. Vale que $\bar{\mathtt{i}}(p) = \mathtt{i}(p)$, es decir, la extensión de una interpretación es efectivamente una extensión. Se suele nombrar indistintamente como $\mathtt{i}$ a una interpretación y a su extensión.

Si Σ es un conjunto de fórmulas, decimos que Σ *es verdadero en un estado* w *de un modelo* $\mathfrak{M}$ si todos los elementos de Σ son verdaderos en w. En tal caso notamos $\mathfrak{M}, w \Vdash \Sigma$.

Esta definición es local (se evalúa en cada estado) e interna (se evalúa dentro de un modelo). Pero también hay otras nociones de verdad o satisfacción:

Definición 2.2.7. Una fórmula φ es *globalmente o universalmente verdadera* en un modelo $\mathfrak{M}$ si es verdadera en todos los puntos del modelo. En tal caso notamos indistintamente $\Vdash_{\mathfrak{M}} \varphi$ o $\mathfrak{M} \Vdash \varphi$.

Definición 2.2.8. Una fórmula φ es *satisfacible* en un modelo $\mathfrak{M}$ si es verdadera en algún punto del modelo, y es *refutable* en un modelo si es falsa en algún punto del modelo. Un conjunto Σ de fórmulas es universalmente verdadera (respectivamente, satisfacible) en un modelo $\mathfrak{M}$ si $w \Vdash_{\mathfrak{M}} \Sigma$ para todos los estados (respectivamente, algún estado) w de $\mathfrak{M}$.

Recordando la lectura de $\Diamond\varphi$ como "posiblemente vale φ" y de $\Box\varphi$ como "necesariamente vale φ", esta definición de satisfacción es esencialmente un intento de capturar matemáticamente la visión (atribuída a Leibniz) de que "necesidad" significa "verdadero en todos los mundos" y "posible" significa verdadero en algún mundo. De allí viene la palabra "mundo" para denominar los estados del modelo.

Es esperable una generalización de frames, modelos y verdad para lenguajes modales de tipo de similaridad arbitrario:

Definición 2.2.9. Sea τ un tipo de similaridad modal. Un τ-frame es una tupla $\mathfrak{F}$ que tiene

1. Un conjunto W no vacío

2. Para cada operador modal $\Diamond$ de aridad $\rho(\Diamond) = n$, una $R_\Diamond$ una relación n-aria en W

Notaremos $\mathfrak{F} = \langle W, R_\Diamond \rangle_{\Diamond \in \tau}$ o, en caso de que τ tenga finitos operadores modales: $\mathfrak{F} = \langle W, R_{\Diamond_1}, \ldots, R_{\Diamond_n} \rangle$.

Construimos un modelo a partir de un frame, exactamente del mismo modo que para el lenguaje modal básico, muniéndolo de una interpretación.

La noción de verdad en un estado w de un modelo $\mathfrak{M} = \langle W, \mathtt{i}, R_\Diamond \rangle_{\Diamond \in \tau}$ se define inductivamente con las mismas reglas que en el lenguaje modal básico para los casos atómicos y booleanos, junto con esta regla:

- $w \Vdash_{\mathfrak{M}} \Diamond(\phi_1, \ldots, \phi_n)$ sii para algún $v_1, \ldots, v_n \in W$ con $R_\Diamond(w, v_1, \ldots, v_n)$ vale para cada i $v_i \Vdash_{\mathfrak{M}} \varphi$

Escribiremos $w \Vdash \varphi$ en lugar de $w \Vdash_{\mathfrak{M}} \varphi$ cuando $\mathfrak{M}$ quede claro del contexto.

Nota. Por motivos formales se define primero el lenguaje y en base a ellos los modelos de Kripke, pero en realidad el modelo es previo a la construcción del lenguaje, en el sentido de que el lenguaje está al servicio de lo que se quiere modelar con el modelo de Kripke, y no al revés.

Por este motivo, muchas veces se construye un frame o un modelo de Kripke con las relaciones sin referenciar a un operador modal, sino al revés: se construye un modelo $\mathfrak{M} = \langle W, R_1, \ldots, R_n, \mathtt{i} \rangle$, y luego se nombra a los operadores universales asociados a cada una de ellas $[R_i]$ y a los existenciales $\langle R_i \rangle$. Usaremos esta notación en Capítulos posteriores.

2.3. Validez

En la sección anterior hablamos de modelos, que consisten en un frame e información contigente: la interpretación. A menudo queremos ignorar los efectos de la interpretación, y tener una noción de qué fórmulas son válidas independientemente de los valores de verdad otorgados a las variables proposicionales. El concepto de *validez* nos permite captar verdades a nivel frame. Una fórmula es válida en un frame si es verdad en todo punto de todo modelo que puede construirse a partir del frame. Este concepto interpreta fórmulas modales sobre frames por abstracción de los efectos de valuaciones en particular.

Definición 2.3.1. Una fórmula φ es válida en un estado w de un frame $\mathfrak{F}$ si φ es verdadero en w para todo modelo basado en $\mathfrak{F}$, y lo notaremos $w \Vdash_{\mathfrak{F}} \varphi$.

Si φ es válida en todo estado de $\mathfrak{F}$ decimos que φ es válida en $\mathfrak{F}$ y notaremos $\Vdash_{\mathfrak{F}} \varphi$ o $\mathfrak{F} \Vdash \varphi$.

Si $\mathtt{F}$ es una clase de frames, decimos que φ es válida en $\mathtt{F}$ si lo es en cada $\mathfrak{F}$ en $\mathtt{F}$.

Observación 2.3.2. Se puede ver facilmente que si tomamos $\mathtt{F}$ es la clase de todos los frames $\mathtt{F} \Vdash \Diamond\varphi \leftrightarrow \neg\Box\neg\varphi$ y $\mathtt{F} \Vdash \Box\varphi \leftrightarrow \neg\Diamond\neg\varphi$, por este motivo es que a ambas modalidades se las llama duales.

Uno podría preguntarse si nuestras nociones de verdad y validez son equivalentes, y la respuesta es no. Un ejemplo simple e introductorio es el siguiente: cuando $\varphi \vee \psi$ es verdadera en un punto w, esto significa que φ es verdadera en w o ψ lo es en w. Por otro lado, si $\varphi \vee \psi$ es válido en un frame $\mathfrak{F}$ no significa que φ o ψ sean válidos en $\mathfrak{F}$ ($p \vee \neg p$ es

un contraejemplo simple, en el que se ve que no vale φ ni ψ). Sin embargo $\Vdash_{\mathfrak{F}}$ implica $\Vdash_{\mathfrak{M}}$, para todo $\mathfrak{M}$ con frame subyacente $\mathfrak{F}$.

Antes de continuar, recordemos los siguientes tipos de relaciones:

- **Reflexiva:** vRv para todo v del dominio
- **Simétrica:** si vRw entonces wRv
- **Transitiva:** si vRw y wRu entonces vRu
- **Serial:** para todo v existe w tal que vRw
- **Euclideana:** si xRy y xRz entonces yRz
- **de equivalencia:** si es reflexiva, transitiva y simétrica o, equivalentemente, si es reflexiva, transitiva y euclideana

Ejemplo 2.3.3. (i) La fórmula $\Diamond(p \vee q) \to (\Diamond p \vee \Diamond q)$ es válida en todos los frames. Para verlo, tomemos un frame $\mathfrak{F}$ y sea $\mathtt{i}$ una interpretación sobre $\mathfrak{F}$. Tenemos que mostrar que si $\langle \mathfrak{F}, \mathtt{i}\rangle, w \Vdash \Diamond(p \vee q)$ entonces $\langle \mathfrak{F}, \mathtt{i}\rangle, w \Vdash \Diamond p \vee \Diamond q$. Asumamos que $\langle \mathfrak{F}, \mathtt{i}\rangle, w \Vdash \Diamond(p \vee q)$. Por definición hay un estado v tal que wRv cumpliendo $\langle \mathfrak{F}, \mathtt{i}\rangle, v \Vdash p \vee q$. Pero si vale $v \Vdash p \vee q$, vale $v \Vdash p$ o $v \Vdash q$. Por lo tanto $w \Vdash \Diamond p$ o $w \Vdash \Diamond q$, entonces $w \Vdash \Diamond p \vee \Diamond q$.

(ii) La fórmula $\Diamond\Diamond p \to \Diamond p$ no es válida en todos los frames. Para ver esto necesitamos hallar un frame $\mathfrak{F}$, un estado w en $\mathfrak{F}$ y una interpretación en $\mathfrak{F}$ que haga falsa la fórmula en w. Por ejemplo: sea $\mathfrak{F}$ un frame con un universo $W = \{0, 1, 2\}$ y una relación $R = \{(0, 1), (1, 2)\}$. Tomamos una interpretación tal que $\mathtt{i}(p) = \{2\}$, entonces $\langle \mathfrak{F}, \mathtt{i}\rangle, 0 \Vdash \Diamond\Diamond p$, pero $\langle \mathfrak{F}, \mathtt{i}\rangle, 0 \Vdash \Diamond p$ ya que 0 no está relacionado con 2.

(iii) Pero hay una clase de frames en la cual $\Diamond\Diamond p \to \Diamond p$ es válida: la clase de frames transitivos (con relación transitiva). En efecto, sea $\mathfrak{F}$ un frame transitivo, w un estado en $\mathfrak{F}$ e $\mathtt{i}$ una interpretación para $\mathfrak{F}$, asumamos que vale $\langle \mathfrak{F}, \mathtt{i}\rangle, w \Vdash \Diamond\Diamond p$. Tenemos que ver que $\langle \mathfrak{F}, \mathtt{i}\rangle, w \Vdash \Diamond p$. Como vale $\langle \mathfrak{F}, \mathtt{i}\rangle, w \Vdash \Diamond\Diamond p$, por definición se tiene que hay un estado v tal que wRv y $\langle \mathfrak{F}, \mathtt{i}\rangle, v \Vdash \Diamond p$, y esto a su vez significa que hay un estado u tal que vRu y $\langle \mathfrak{F}, \mathtt{i}\rangle, u \Vdash p$. Pero como R es transitiva y valen wRv y vRu entonces vale que wRu con u tal que $\langle \mathfrak{F}, \mathtt{i}\rangle, u \Vdash p$, es decir, vale $\langle \mathfrak{F}, \mathtt{i}\rangle, w \Vdash \Diamond p$. La otra implicación se sigue de aplicar definiciones.

Proposición 2.3.4. *Sea $\mathfrak{F} = \langle W, R\rangle$ un frame, vale:*

1. *R es reflexiva sii $\Vdash_{\mathfrak{F}} \Box\varphi \to \varphi$*
2. *R es euclideana sii $\Vdash_{\mathfrak{F}} \neg\Box\varphi \to \Box\neg\Box\varphi$*
3. *R es serial sii $\Vdash_{\mathfrak{F}} \neg\Box\bot$*

4. R es transitiva sii $\Vdash_{\mathfrak{F}} \Box\varphi \to \Box\Box\varphi$

Demostración. Las implicaciones hacia la derecha son aplicar definiciones, vamos a demostrar las implicaciones hacia la izquierda.

1) Supongamos que R no es reflexiva, entonces existe $w \in W$ tal que no vale wRw. Sea $p \in \Phi$, definimos un modelo $\mathfrak{M} = \langle \mathfrak{F}, \mathtt{i} \rangle$, con $\mathtt{i}(p) = \{w\}^c = W \setminus \{w\}$. Es fácil ver que $w \Vdash_{\mathfrak{M}} \Box p$ pero $w \Vdash_{\mathfrak{M}} \neg p$.

2) Supongamos que R no es euclideana, entonces existen $w, v, u \in W$ tales que wRv y wRu, pero no vale vRu. Sea $p \in \Phi$, definimos un modelo $\mathfrak{M} = \langle \mathfrak{F}, \mathtt{i} \rangle$ con $\mathtt{i}(p) = \{u\}^c$. Se puede ver que $w \Vdash_{\mathfrak{M}} \neg\Box p$ pero $w \Vdash_{\mathfrak{M}} \neg\Box\neg\Box p$.

3) Supongamos que no, entonces existe $w \in W$ tal que no está relacionado con nadie, por lo tanto $\Vdash_{\mathfrak{F}} \Box\bot$.

4) Demostrado en el Ejemplo 2.3.3 iii). □

2.4. Consecuencias modales

Si bien la idea de validez en frames permite comenzar a razonar acerca de fórmulas modales, no hemos dicho nada acerca de qué significa *consecuencia lógica* para lógicas modales.

La consecuencia lógica es la relación que hay entre las premisas y la conclusión que uno puede desprender de ellas vía razonamientos lógicos. Definir el concepto de consecuencia lógica involucra preguntarse qué significa que una conclusión sea consecuencia de sus premisas. En [11] se encuentran 3 propiedades mencionadas por Tarski a tener en cuenta para una caracterización correcta de la consecuencia lógica:

1. La relación de consecuencia lógica reside en la forma lógica de las fórmulas involucradas

2. La relación es a priori, i.e., puede ser determinada si vale o no independientemente de evidencia empírica

3. Tiene un componente modal, en el sentido de que si valen las premisas *necesariamente* vale la conclusión.

Es decir, al no tener una noción de consecuencia no sabemos aún que significa que un conjunto de fórmulas modales Σ implique una fórmula modal φ.

Vamos a introducir dos familias de consecuencias: una local y una global, ambas definidas semánticamente en términos de clases de estructuras. Comenzamos con unas definiciones previas.

Definición 2.4.1. Llamamos *clase de estructuras* a una clase de modelos o de frames.

Si $\mathtt{S}$ es una clase de modelos, llamamos *modelo de* $\mathtt{S}$ es un elemento $\mathfrak{M}$ de $\mathtt{S}$. Si $\mathtt{S}$ es una clase de frames, llamamos *modelo de* $\mathtt{S}$ es un modelo con frame subyacente en $\mathtt{S}$.

Para definir una noción razonable de consecuencia lógica hay dos ítems a tener en cuenta. Primero, parece correcto mantener la idea familiar de que una consecuencia semántica sucede cuando la verdad de las premisas garantizan la verdad de la conclusión. Segundo, las inferencias que podemos adoptar van a depender de la clase de estructuras en la que estemos trabajando. Por ejemplo: las inferencias que van a ser legítimas en un frame transitivo serán diferentes a las de un frame que no lo es. Por lo tanto nuestra definición de consecuencia será ser paramétrica: debe hacer referencia a una clase de estructuras $\mathtt{S}$.

Definición 2.4.2. Sea τ un tipo de similaridad y sea $\mathtt{S}$ una clase de estructuras. Se dice que $\mathtt{S}$ *es de tipo* τ si $\mathtt{S}$ es una clase de frames de tipo τ o clase de modelos sobre frames de tipo τ.

Definición 2.4.3. Sea $\mathtt{S}$ una clase de estructuras de tipo τ, Σ un conjunto de fórmulas y φ una fórmula de un lenguaje de tipo τ. Decimos que φ es *consecuencia semántica local de* Σ *sobre* $\mathtt{S}$ (notación: $\Sigma \Vdash_S \varphi$) si para todos los modelos $\mathfrak{M}$ de $\mathtt{S}$ y todos los puntos w de $\mathfrak{M}$ vale que si $\mathfrak{M}, w \Vdash \Sigma$ entonces $\mathfrak{M}, w \Vdash \varphi$.

Es local en el sentido que miramos que se cumpla algo a nivel estado (o punto) del modelo.

Ejemplo 2.4.4. Supongamos que estamos trabajando con $\mathtt{Tran}$, la clase de frames transitivos. Entonces, por lo visto en el Ejemplo 2.3.3:

$$\{\Diamond\Diamond p\} \Vdash_{\mathtt{Tran}} \Diamond p$$

sin embargo, $\Diamond p$ *no* es consecuencia semántica local de $\{\Diamond\Diamond p\}$ en la clase de *todos* los modelos.

La variante global es la esperable:

Definición 2.4.5. Sean τ, $\mathtt{S}$, Σ y φ como en la Definición 2.4.3. Decimos que φ es una *consecuencia semántica global de* Σ *sobre* $\mathtt{S}$ (notación: $\Sigma \Vdash^g_{\mathtt{S}} \varphi$) si y sólo si para todas las estructuras $\mathfrak{S}$ en $\mathtt{S}$ vale que si $\mathfrak{S} \Vdash \Sigma$ entonces $\mathfrak{S} \Vdash \varphi$.

Ejemplo 2.4.6. Las consecuencias semánticas locales y globales son conceptualmente diferentes. Para fijar ideas: consideremos las fórmulas p y $\Box p$. Es fácil ver que p no implica localmente $\Box p$ (de hecho, que esto no valga hace a la esencia de la noción de localidad). Por otro lado, supongamos que tenemos un modelo $\mathfrak{M}$ en el que p es globalmente verdadero (es decir: $\mathtt{i}(p) = W$). En particular p vale en todos los sucesores de todos los estados, así que vale $\mathfrak{M} \Vdash \Box p$, y por lo tanto vale $p \Vdash^g \Box p$.

Sin embargo, hay una conexión entre estos dos tipos de consecuencias.

Proposición 2.4.7. *Sea Σ un conjunto de fórmulas del lenguaje modal básico, y sea $\mathtt{F}$ la clase de todos los frames. Vale que $\Sigma \Vdash^g_{\mathtt{F}} \varphi$ sii $\{\Box^n\sigma\colon \sigma \in \Sigma, n \in \mathbb{N}\} \Vdash_{\mathtt{F}} \varphi$, donde $\Box^n\varphi$ es una abreviación para $\overbrace{\Box \ldots \Box}^{n \text{ veces}} \varphi$.*

2.5. Equivalencias, bisimulaciones y filtraciones

Definición 2.5.1. Sean w y w' mundos de dos modelos $\mathfrak{M}$ y $\mathfrak{M}'$ para un mismo lenguaje modal. Se dice que w y w' son *equivalentes* si para toda φ vale que $w \Vdash_{\mathfrak{M}} \varphi$ sii $w' \Vdash_{\mathfrak{M}'} \varphi$. En tal caso notamos $\mathfrak{M}, w \equiv \mathfrak{M}', w'$.

Definición 2.5.2. Sean dos modelos de Kripke $\mathfrak{M} = \langle W, R, \mathtt{i} \rangle$ y $\mathfrak{M}' = \langle W', R', \mathtt{i}' \rangle$ basados en el lenguaje modal básico. Una *bisimulación entre* $\mathfrak{M}$ *y* $\mathfrak{M}'$ es una relación $Z \subseteq W \times W'$ que cumple:

- **(forth)** Para todo $w, v \in W$ y $w' \in W'$ si wZw' y wRv entonces existe un $v' \in W'$ tal que vZv' y $w'R'v'$
- **(back)** Para todo $w \in W$ y $w', v' \in W'$ si wZw' y $w'R'v'$ entonces existe un $v \in W$ tal que vZv' y wRv
- **(prop)** Para todo $w \in W$, $w' \in W'$ si wZw' entonces para toda $p \in \Phi$ $w \Vdash_{\mathfrak{M}} p$ sii $w' \Vdash_{\mathfrak{M}'} p$

Si, además, Z cumple que su dominio es W y su imagen es W' se llama *bisimulación zigzag*.

Si existe una bisimulación entre $\mathfrak{M}$ y $\mathfrak{M}'$, se dice que son *bisimilares*.

La idea detrás del concepto de bisimulación es la de identificar modelos que son iguales, en lo que al alcance de la lógica modal se refiere. Si vía lógica modal no podemos distinguirlos, entonces el concepto de bisimulación tiene una definición adecuada. A este propósito viene el siguiente teorema:

Teorema 2.5.3. *Sean* $\mathfrak{M} = \langle W, R, \mathtt{i} \rangle$ *y* $\mathfrak{M}' = \langle W', R', \mathtt{i}' \rangle$ *bisimilares (con bisimulación* Z*). Entonces para todos* $w \in W$ *y* $w' \in W'$ *si* wZw' *vale* $\mathfrak{M}, w \equiv \mathfrak{M}', w'$.

Demostración. Queremos ver que si wZw', y $w \Vdash_{\mathfrak{M}} \varphi$ sii $w' \Vdash_{\mathfrak{M}'} \varphi$, y esto sale por inducción en φ: si $\varphi \in \Phi$, $\varphi = \varphi_1 \wedge \varphi_2$ o $\varphi = \neg\bar{\varphi}$ se desprende de aplicar definiciones.

Supongamos que $\varphi = \Diamond\psi$ y $w \Vdash_{\mathfrak{M}} \Diamond\psi$ queremos ver que $w' \Vdash_{\mathfrak{M}'} \Box\psi$. Tenemos que existe un v tal que wRv cumpliendo $w \Vdash_{\mathfrak{M}} \psi$, y por la propiedad **forth** tenemos que existe un v' tal que vZv' y $w'R'v'$. Por hipótesis inductiva, y recordando que vale $v \Vdash_{\mathfrak{M}} \psi$, tenemos que $v' \Vdash_{\mathfrak{M}'} \psi$. Por lo tanto, $w' \Vdash_{\mathfrak{M}'} \Diamond\psi$.

La otra implicación es análoga pero utilizando la propiedad **back**. □

Con bisimulación podemos comparar si dos modelos son iguales. Ahora estamos interesados en alguna herramienta que nos permita obtener modelos a partir de otros, que preserven parte de la estructura subyacente.

Definición 2.5.4. Sea $\mathfrak{M} = \langle W, R, \mathtt{i} \rangle$ un modelo de Kripke y Φ un conjunto finito de fórmulas cerrado por subfórmulas. Definimos la relación $\equiv_\Phi$ tal que $w \equiv_\Phi v$ si para toda $\psi \in \Phi$ vale: $w \Vdash \psi$ sii $v \Vdash \psi$. Esta relación resulta de equivalencia y dado un $w \in W$ notaremos $[w]_\Phi$ a la clase de equivalencia de w, o simplemente $[w]$ cuando Φ quede clara en el contexto.

Definición 2.5.5. Sea $\mathfrak{M} = \langle W, R, \mathtt{i}\rangle$ un modelo de Kripke y Φ un conjunto finito de fórmulas cerrado por subfórmulas. Llamamos *filtración de* $\mathfrak{M}$ *a través de* Φ a un modelo $\mathfrak{M}_f = \langle W_f, R^f, \mathtt{i}_f\rangle$ donde $W_f = \{[w] \ : \ w \in W\}$, $\mathtt{i}_f(p) = \mathtt{i}(p)$, donde la relación R^f cumple:

- **Min(R^f/R):** para todo $[w], [v] \in W_f$ si existen $w', v' \in W$ tales que $w'Rv'$, $[w] = [w']$ y $[v] = [v']$, entonces $[w]R^f[v]$
- **Max(R^f):** para todo $[w], [v] \in W_f$ si $[w]R^f[v]$ entonces para toda $\Box\psi \in \Phi$ si vale $w \Vdash_{\mathfrak{M}} \Box\psi$ entonces vale $v \Vdash_{\mathfrak{M}} \psi$

También llamaremos *filtración de* R *a través de* Φ a la relación R^f resultante.

Para modelos multimodales una filtración, es un modelo $\mathfrak{M}_f = \langle W_f, R_1^f, \ldots, R_n^f, \mathtt{i}_f\rangle$ donde cada R_i^f es una filtración de R_i.

Observación 2.5.6. Las filtraciones son modelos finitos, ya que sus estados son las clases de equivalencia de $\equiv_\Phi$ y no puede haber más que las maneras de asignar valores de verdad a las finitas variables proposicionales presentes en las finitas fórmulas de Φ.

De este modo tenemos una herramienta para generar modelos finitos a partir de modelos arbitrarios, cuyas restricciones establecen que se preserven los valores de verdad de las fórmulas de Φ, tal como queda demostrado en este teorema:

Teorema 2.5.7. *Sea* $\mathfrak{M} = \langle W, R_1, \ldots, R_n, \mathtt{i}\rangle$ *y* $\mathfrak{M}_f$ *una filtración de* $\mathfrak{M}$ *a través de* Φ. *Entonces para toda* $\psi \in \Phi$ *y* $w \in W$ *vale* $w \Vdash_{\mathfrak{M}} \psi$ *sii* $[w] \Vdash_{\mathfrak{M}_f} \psi$.

Demostración. Lo veremos por inducción en la estructura de ψ. Para $p \in \Phi$ y los conectivos lógicos se desprenden de la definición de $\mathtt{i}_f$, veamos directamente el caso $\psi = \Box\alpha$ y asumamos que el teorema ya fue demostrado para α.

Supongamos primero que $w \Vdash_{\mathfrak{M}} \Box\alpha$ y sea $[v]$ tal que $[w]R_f[v]$. Queremos ver que $[v] \Vdash_{\mathfrak{M}_f} \alpha$. Por **Max($R^f$)**, tenemos que si $w \Vdash_{\mathfrak{M}} \Box\alpha$ y $[w]R^f[v]$ entonces $v \Vdash_{\mathfrak{M}} \alpha$. Por hipótesis inductiva: $[v] \Vdash_{\mathfrak{M}_f} \alpha$.

Veamos la vuelta. Supongamos que $[w] \Vdash_{\mathfrak{M}_f} \Box\alpha$, entonces para todo $[v]$ tal que $[w]R^f[v]$ vale $[v] \Vdash_{\mathfrak{M}_f} \alpha$. Sea u tal que wRu, por **Min(R^f/R)** tenemos que $[w]R^f[u]$ entonces $[u] \Vdash_{\mathfrak{M}_f} \alpha$. Por hipótesis inductiva: $u \Vdash_{\mathfrak{M}} \alpha$. □

2.6. Sistemas de axiomas

Hasta ahora se ha hablado de los aspectos semánticos, pero la lógica tiene otro componente importante que es la sintaxis. La primera pregunta que surge es si dada una clase de frames $\mathtt{F}$ existe algún mecanismo sintáctico capaz de generar las fórmulas válidas en $\mathtt{F}$. El objetivo de esta sección es responder esa pregunta, y para ello vamos a definir un sistema de axiomas **K**, que veremos que es el sistema adecuado para razonar acerca de frames. De este sistema derivan otros más fuertes que se verán posteriormente.

Definición 2.6.1. Llamamos *sistema de axiomas* **S** a un conjunto de fórmulas junto con un conjunto de reglas de inferencia.

Una *demostración basada en* **S** es una secuencia ordenada de fórmulas, cada una de las cuáles es un axioma, o se deriva de uno o más ítems anteriores de la secuencia por aplicación de alguna de las reglas de inferencia del sistema.

Una fórmula φ se dice *demostrable por* **S**, si existe una demostración basada en **S** cuya última fórmula es φ. En tal caso notamos $\vdash_{\mathbf{S}} \varphi$.

Definición 2.6.2. Definimos *el sistema axiomático* **K**. Sus axiomas son los axiomas de la lógica proposicional:

- **(SP1)** $\varphi \to (\psi \to \varphi)$
- **(SP2)** $(\varphi \to (\psi \to \rho)) \to ((\varphi \to \psi) \to (\varphi \to \rho))$
- **(SP3)** $(\neg\varphi \to \neg\psi) \to (\psi \to \varphi)$

agregándole los siguientes axiomas:

- **(K)** $\Box(p \to q) \to (\Box p \to \Box q)$
- **(Dual)** $\Diamond p \leftrightarrow \neg\Box\neg p$

Y las reglas de inferencia son:

1. *Modus ponens:* dada φ y $\varphi \to \psi$, queda probado ψ.
2. *Sustitución uniforme:* dada φ, queda probado θ, donde θ se obtiene de φ reemplazando uniformemente las variables proposicionales en φ por fórmulas arbitrarias.
3. *Necesitación:* dada φ, queda probado $\Box\varphi$.

Observación 2.6.3. Modus ponens preserva validez, es decir, si $\Vdash \varphi$ y $\Vdash \varphi \to \psi$ entonces $\Vdash \psi$. Dado que pretendemos razonar acerca de frames, esta propiedad es crucial. Pero además, modus ponens preserva verdad global (si $\mathfrak{M} \Vdash \varphi$ y $\mathfrak{M} \Vdash \varphi \to \psi$, entonces vale $\mathfrak{M} \Vdash \psi$) y satisfabilidad (si $\mathfrak{M}, w \Vdash \varphi$ y $\mathfrak{M}, w \Vdash \varphi \to \psi$ entonces vale $\mathfrak{M}, w \Vdash \psi$). Por lo tanto también es una regla de inferencia adecuada para razonar acerca de modelos tanto local como globalmente.

Sin embargo no sucede lo mismo con la regla de sustitución uniforme. Esta regla preserva validez, pero no preserva verdad global y satisfabilidad. Por ejemplo: q se obtiene de p por sustitución uniforme, pero aunque p sea globalmente verdadero en un modelo no necesariamente sucede lo mismo con q.

El axioma **(K)** es el axioma fundamental de este sistema. Es claramente válido, pero además de eso nos permite transformar una fórmula-box $\Box(\varphi \to \psi)$ en una implicación $\Box\varphi \to \Box\psi$. Por otro lado el axioma Dual establece que $\Box$ es una abreviación de $\Diamond$.

Definición 2.6.4. Un sistema de axiomas **S** se dice *correcto con respecto a una clase de frames* $\mathfrak{F}$ si para toda fórmulas φ, $\vdash_{\mathbf{S}} \varphi$ implica $\Vdash_{\mathfrak{F}} \varphi$, y se dice *completo con respecto a una clase de frames* $\mathfrak{F}$ si para toda fórmulas φ, $\Vdash_{\mathfrak{F}} \varphi$ implica $\vdash_{\mathbf{S}} \varphi$.

Vimos que las reglas de inferencia preservan validez y además los axiomas de **K** son válidos en la clase de todos los frames, entonces **K** es correcto con respecto a esa clase. Además, veremos más adelante, resulta ser que el recíproco también vale: si una fórmula es válida en la clase de todos los frames es demostrable por **K**. Es decir, que **K** es correcto y completo respecto de la clase de todos los frames.

El sistema axiomático **K** resulta demasiado débil como para demostrar propiedades que son válidas en algunas clases de frames. Si estamos interesados en frames transitivos, necesitamos un sistema de axiomas que refleje esa propiedad. Por ejemplo, sabemos que $\Diamond\Diamond p \to \Diamond p$ es válida en todos los frames transitivos (como vimos en el Ejemplo 2.3.3), entonces queremos un sistema capaz de demostrar esta fórmula. Y sabemos que **K** no va a poder demostrarla porque esa fórmula no es válida en todos los frames. El sistema que resulta de agregar $\Diamond\Diamond p \to \Diamond p$ al sistema **K** se llama **K4** y es correcto y completo respecto de la clase de los frames transitivos, como se demuestra en [13] utilizando lo que se llama Teoría de Correspondencia: una herramienta estándar para demostrar completitud de lógicas modales basada en conceptos de la siguiente sección.

Maximales consistentes y modelo canónico

En esta sección veremos herramientas estándares desarrolladas en el libro [23] para demostrar completitud de lógicas modales, que serán utilizadas en el siguiente capítulo para las lógicas epistémicas.

Definición 2.6.5. Dado un sistema de axiomas **S**, decimos que:

- φ es *consistente* con **S** si $\nvdash_{\mathbf{S}} \neg\varphi$
- Un conjunto finito $\{\varphi_1, \ldots, \varphi_n\}$ es *consistente* con **S** si $\varphi_1 \wedge \ldots \wedge \varphi_n$ lo es
- Un conjunto infinito de fórmulas $\mathbf{\Phi}$ es *consistente* con **S** si cualquier subconjunto finito lo es
- Un conjunto de fórmulas $\mathbf{\Phi}$ es *maximal consistente* si es consistente y $\mathbf{\Phi} \cup \{\psi\}$ no lo es, para cualquier fórmula $\psi \notin \mathbf{\Phi}$
- Una fórmula o conjunto de fórmulas se dice *inconsistente con* **S** si no es consistente con **S**

Proposición 2.6.6. *Sea* ***S*** *un sistema de axiomas que extiende a la axiomatización de la lógica proposicional. Entonces valen:*

1. *Todo conjunto de fórmulas* $\mathbf{\Phi}$ *consistente con un sistema de axiomas* ***S*** *puede ser extendido a un maximal consistente*

2. Sea $\mathbf{\Phi}$ *maximal consistente, para todo* φ, ψ *valen:*

a) *o bien* $\varphi \in \mathbf{\Phi}$ *o bien* $\neg\varphi \in \mathbf{\Phi}$

b) $\varphi \wedge \psi \in \mathbf{\Phi}$ *sii* $\varphi \in \mathbf{\Phi}$ *y* $\psi \in \mathbf{\Phi}$

c) *Si* $\varphi \in \mathbf{\Phi}$ *y* $\varphi \to \psi \in \mathbf{\Phi}$ *entonces* $\psi \in \mathbf{\Phi}$

d) *Si* $\vdash_{\boldsymbol{S}} \varphi$ *entonces* $\varphi \in \mathbf{\Phi}$

Demostración. 1) Como Φ es finito, podemos numerar las fórmulas: $\varphi_0, \varphi_1, \ldots$. Definimos una secuencia creciente de conjuntos $\mathbf{\Phi}_n$ inductivamente comenzando con $\mathbf{\Phi}_0 := \mathbf{\Phi}$ y tomando $\mathbf{\Phi}_{n+1} := \mathbf{\Phi}_n \cup \{\varphi_n\}$ si esta unión resulta consistente con $\mathbf{S}$ o $\mathbf{\Phi}_{n+1} := \mathbf{\Phi}$ en caso contrario, y definimos $\mathbf{\Phi}_\omega := \bigcup \mathbf{\Phi}_n$.

$\mathbf{\Phi}_\omega$ resulta consistente. En efecto, si no lo fuera existe algún subconjunto finito $\mathbf{\Psi} \subseteq \mathbf{\Phi}_\omega$ inconsistente, pero como es finito existe un n tal que $\mathbf{\Psi} \subseteq \mathbf{\Phi}_n$, lo cual no es posible porque $\mathbf{\Phi}_n$ es consistente.

Más aún, $\mathbf{\Phi}_\omega$ es maximal consistente. Si no lo fuera, existiría $\varphi \notin \mathbf{\Phi}_\omega$ tal que $\mathbf{\Phi}_\omega \cup \{\varphi\}$ es consistente. Como habíamos numerado las fórmulas existe un m tal que $\varphi = \varphi_m$, entonces $\mathbf{\Phi}_m \cup \{\varphi_m\}$ es inconsistente (de lo contrario $\varphi \in \mathbf{\Phi}_\omega$). Entonces $\mathbf{\Phi}_\omega \cup \{\varphi\}$ no es consistente.

$2a$) Supongamos que no, entonces existe φ tal que $\varphi \notin \mathbf{\Phi}$ y $\neg\varphi \notin \mathbf{\Phi}$. Como $\mathbf{\Phi}$ es maximal consistente, $\mathbf{\Phi} \cup \{\varphi\}$ y $\mathbf{\Phi} \cup \{\neg\varphi\}$ son inconsistentes con $\mathbf{S}$. Entonces existen subconjuntos finitos $\mathbf{\Psi_1}$ y $\mathbf{\Psi}_2$ tales que son inconsistentes con $\mathbf{S}$ y $\mathbf{\Psi}_1 \subseteq \mathbf{\Phi} \cup \{\varphi\}$ y $\mathbf{\Psi}_2 \subseteq \mathbf{\Phi} \cup \{\neg\varphi\}$. Por lo tanto tenemos que $\mathbf{\Psi}_1 \cup \mathbf{\Psi}_2 \cup \{\varphi\}$ y $\mathbf{\Psi}_1 \cup \mathbf{\Psi}_2 \cup \{\neg\varphi\}$ son inconsistentes con $\mathbf{S}$. Como $\mathbf{\Psi}_1 \cup \mathbf{\Psi}_2$ es finito, supongamos $\mathbf{\Psi}_1 \cup \mathbf{\Psi}_2 = \{\psi_1, \ldots, \psi_r\}$. Entonces tenemos $\vdash_{\mathbf{S}} \neg(\psi_1 \wedge \ldots \wedge \psi_r \wedge \varphi)$ y $\vdash_{\mathbf{S}} \neg(\psi_1 \wedge \ldots \wedge \psi_r \wedge \neg\varphi)$. Usando los axiomas y reglas de inferencia de la lógica proposicional tenemos que:

$$\vdash_{\mathbf{S}} \neg\psi_1 \vee \neg\psi_2 \vee \ldots \vee \neg\psi_r \vee \neg\varphi$$

$$\vdash_{\mathbf{S}} \neg\psi_1 \vee \neg\psi_2 \vee \ldots \vee \neg\psi_r \vee \varphi$$

$$\vdash_{\mathbf{S}} (\psi_1 \wedge \ldots \wedge \psi_r) \to \varphi$$

$$\vdash_{\mathbf{S}} (\psi_1 \wedge \ldots \wedge \psi_r) \to \neg\varphi$$

$$\vdash_{\mathbf{S}} (\psi_1 \wedge \ldots \wedge \psi_r) \to (\varphi \wedge \neg\varphi)$$

$$\vdash_{\mathbf{S}} (\psi_1 \wedge \ldots \wedge \psi_r) \to \bot$$

$$\vdash_{\mathbf{S}} \neg(\psi_1 \wedge \ldots \wedge \psi_r)$$

Pero si $\vdash_{\mathbf{S}} \neg(\psi_1 \wedge \ldots \wedge \psi_r)$, resulta que $\mathbf{\Phi}$ es inconsistente, porque $\mathbf{\Psi}_1 \cup \mathbf{\Psi}_2 \subseteq \mathbf{\Phi}$.

$2b$) $\Rightarrow$) Supongamos $\varphi \wedge \psi \in \mathbf{\Phi}$ y $\varphi \notin \mathbf{\Phi}$. Entonces $\mathbf{\Phi} \cup \{\varphi\}$ es inconsistente, es decir, para algún subconjunto finito $\mathbf{\Psi} \subseteq \mathbf{\Phi}$, $\mathbf{\Psi} \cup \{\varphi\} = \{\psi_1, \ldots, \psi_r, \varphi\}$ es inconsistente. Es decir: $\vdash_{\mathbf{S}} \neg(\psi_1 \wedge \ldots \wedge \psi_r \wedge \varphi)$.

Como $\mathbf{\Psi} \cup \{\varphi \wedge \psi\} \vdash_{\mathbf{S}} \mathbf{\Psi} \cup \{\varphi\}$, entonces también resulta que $\mathbf{\Psi} \cup \{\varphi \wedge \psi\}$ es inconsistente contradiciendo el hecho de que $\mathbf{\Phi}$ es maximal consistente.

$\Leftarrow$) Supongamos que $\varphi, \psi \in \mathbf{\Phi}$ pero $\varphi \wedge \psi \notin \mathbf{\Phi}$. Entonces $\mathbf{\Phi} \cup \{\varphi \wedge \psi\}$ es inconsistente, es decir, para algún conjunto finito $\{\psi_1, \ldots, \psi_r\}$ vale $\vdash_{\mathbf{S}} \neg(\psi_1 \wedge \ldots \wedge \psi_r \wedge \varphi \wedge \psi)$, pero esto afirma que $\{\psi_1, \ldots, \varphi, \psi\}$ es inconsistente, pero es un subconjunto de $\mathbf{\Phi}$.

$2c$) Supongamos que $\varphi, \varphi \to \psi \in \mathbf{\Phi}$, pero $\psi \notin \mathbf{\Phi}$. Resulta que existe un subconjunto finito $\{\psi_1, \ldots, \psi_r\, \psi\}$ inconsistente, es decir, $\vdash_{\mathbf{S}} \neg(\psi_1 \wedge \ldots \wedge \psi_r \wedge \psi)$. Pero esto último a su vez implica que $\{\psi_1, \ldots, \psi_r, \varphi, \varphi \to \psi\}$ es inconsistente (porque $\{\psi_1, \ldots, \psi_r, \varphi, \varphi \to \psi\} \vdash_{\mathbf{S}} \{\psi_1, \ldots, \psi_r, \varphi, \psi\}$). Y eso contradice el hecho de que $\mathbf{\Phi}$ es un maximal consistente.

$2d$) Supongamos que $\vdash_{\mathbf{S}} \varphi$ pero $\varphi \notin \mathbf{\Phi}$. Como $\mathbf{\Phi}$ es maximal consistente $\mathbf{\Phi} \cup \{\varphi\}$ es inconsistente, existen $\psi_1, \ldots, \psi_r \in \mathbf{\Phi}$ tales que $\vdash_{\mathbf{S}} \neg(\psi_1 \wedge \ldots \wedge \psi_r \wedge \varphi)$ y de esto se tiene que $\vdash_{\mathbf{S}} \neg\varphi$. Entonces podemos realizar la siguiente demostración:

$$\vdash_{\mathbf{S}} (\psi_1 \wedge \ldots \wedge \psi_r) \to \neg\varphi$$

$$\vdash_{\mathbf{S}} \varphi$$

Por el axioma **SP3** de la lógica proposicional:

$$\vdash_{\mathbf{S}} \varphi \to (\psi_1 \wedge \ldots \wedge \psi_r)$$

Por modus ponens:

$$\vdash_{\mathbf{S}} \neg(\psi_1 \wedge \ldots \wedge \psi_r)$$

Y esto contradice que $\mathbf{\Phi}$ es consistente. □

Definición 2.6.7. Un sistema axiomático **S** se dice *inconsistente* si existe φ tal que $\vdash_{\mathbf{S}} \varphi$ y $\vdash_{\mathbf{S}} \neg\varphi$. Cuando **S** no es inconsistente lo llamaremos *consistente*.

El siguiente es un resultado clásico de lógica cuya demostración puede hallarse en [19].

Proposición 2.6.8. *Si un sistema axiomático* $\boldsymbol{S}$ *es inconsistente, entonces para toda* φ *vale* $\vdash_{\boldsymbol{S}} \varphi$.

Corolario 2.6.9. *Sea* $\boldsymbol{S}$ *un sistema axiomático inconsistente, entonces el único conjunto de fórmulas consistente con* $\boldsymbol{S}$*es el vacío.*

Definición 2.6.10. Dado un sistema axiomático **S** consistente, llamamos *modelo canónico para* $\boldsymbol{S}$ al modelo $\mathfrak{M}^C := \langle W^C, R^C, \mathtt{i}^C \rangle$ dado por

- $W^C := \{w_\Theta \ : \ \Theta \text{ es un conjunto maximal consistente de fórmulas}\}$
- $\mathtt{i}^C(p) := \{w_\Theta \ : \ p \in \Theta\}$
- $R^C := \{(w_\Theta, w_\Psi) \ : \ \Theta/\Box \subseteq \Psi\}$ con $\Theta/\Box := \{\varphi \ : \ \Box\varphi \in \Theta\}$

Proposición 2.6.11. *Sea* $\boldsymbol{S}$ *un sistema axiomático que extienda el de la lógica proposicional que contenga al axioma* $\boldsymbol{K}$ *y la regla de necesitación.*

Entonces se tiene que:

$$w_\Theta \Vdash_{\mathfrak{M}^C} \varphi \textit{ sii } \varphi \in \Theta$$

Demostración. Veamos que $w_\Theta \Vdash_{\mathfrak{M}^C} \varphi$ sii $\varphi \in \Theta$, por inducción en la longitud de la fórmula φ.

`Caso 1:` $\varphi \in \Phi$ se desprende de la definición de $\mathtt{i}^C$

`Caso 2:` $\varphi = \varphi_1 \wedge \varphi_2$

$w_\Theta \Vdash_{\mathfrak{M}^C} \varphi_1 \wedge \varphi_2$ sii $w_\Theta \Vdash_{\mathfrak{M}^C} \varphi_1$ y $w_\Theta \Vdash_{\mathfrak{M}^C} \varphi_2$ sii (por hipótesis inductiva) $\varphi_1 \in \Theta$ y $\varphi_2 \in \Theta$ sii (por ser Θ maximal consistente) $\varphi_1 \wedge \varphi_2 \in \Theta$.

`Caso 3:` $\varphi = \neg\bar{\varphi}$

$w_\Theta \Vdash_{\mathfrak{M}^C} \neg\bar{\varphi}$ sii $w_\Theta \nVdash_{\mathfrak{M}^C} \bar{\varphi}$ sii (por hipótesis inductiva) $\bar{\varphi} \notin \Theta$ sii (por ser Θ maximal consistente) $\varphi = \neg\bar{\varphi} \in \Theta$

`Caso 4:` $\varphi = \Box\psi$

$\Leftarrow$) Si $\varphi \in \Theta$, entonces $\psi \in \Theta/\Box$. Sea w_Ψ tal que $w_\Theta R^C w_\Psi$, entonces $\psi \in \Theta/\Box \subseteq \Psi$ y por ende (por hipótesis inductiva) $w_\Psi \Vdash_{\mathfrak{M}^C} \psi$. Entonces $w_\Theta \Vdash_{\mathfrak{M}^C} \varphi$

$\Rightarrow$) Supongamos que $w_\Theta \Vdash_{\mathfrak{M}^C} \Box\psi$. Por el lema que se demostrará después de este teorema vale que $\Theta/\Box \cup \{\neg\psi\}$ es inconsistente. Entonces existe un subconjunto finito tal que $\vdash \neg(\psi_1 \wedge \ldots \wedge \psi_r \wedge \neg\psi)$. Usando axiomas proposicionales tenemos que $\vdash \psi_1 \to (\psi_2 \to (\ldots(\psi_r \to \psi)\ldots))$, y por necesitación obtenemos $\vdash \Box(\psi_1 \to (\psi_2 \to (\ldots(\psi_r \to \psi)\ldots)))$.

Renombremos $\epsilon = \psi_2 \to (\psi_3 \to (\ldots(\psi_r \to \psi)\ldots))$,y junto con el axioma **K** tenemos que $\vdash \Box\psi_1 \to (\Box(\psi_1 \to \epsilon) \to \Box\epsilon)$.

Dado que Θ es maximal consistente contiene a todos los teoremas, en particular:

$$\Box\psi_1 \to (\Box(\psi_1 \to \epsilon) \to \Box\epsilon) \in \Theta \tag{2.1}$$

Pero como $\psi_1 \in \Theta/\Box$, $\Box\psi_1 \in \Theta$ y por la Proposición 2.6.62c tenemos que:
$\Box(\psi_1 \to \epsilon) \to \Box\epsilon \in \Theta$, lo que junto con 2.1 y la consistencia maximal de Θ nos garantiza que $\Box\epsilon \in \Theta$. Iterando obtenemos el resultado. □

Proposición 2.6.12. *Sea $\boldsymbol{S}$ un sistema axiomático y $\mathfrak{M}^C$ su modelo canónico. Valen:*

- $\Box\varphi \to \varphi \in \boldsymbol{S}$ *entonces* $\mathfrak{M}^C$ *es reflexivo*
- $\Box\varphi \to \Box\Box\varphi \in \boldsymbol{S}$ *entonces* $\mathfrak{M}^C$ *es transitivo*
- $\neg\Box\varphi \to \Box\neg\Box\varphi \in \boldsymbol{S}$ *entonces* $\mathfrak{M}^C$ *es euclideano*

Demostración.

`Reflexividad:` sea $w_\Theta \in W^C$, queremos ver que $\Theta/\Box \subseteq \Theta$. Sea $\varphi \in \Theta/\Box$, como Θ es maximal consistente contiene a todos los teoremas, en particular al axioma $\Box\varphi \to \varphi$. Entonces como Θ es maximal consistente por la Proposición 2.6.6 ítem

impli, tenemos que $\varphi \in \Theta$. Por lo tanto $\Theta/\Box \subseteq \Theta$, es decir $w_\Theta R^C w_\Theta$.

`Transitividad`: Sean $w_\Theta R^C w_\Psi$ y $w_\Psi R^C w_\Phi$, queremos ver que $w_\Theta R^c W_\Phi$, i.e. que $\Theta/\Box \subseteq \Phi$. Sea φ tal que $\Box\varphi \in \Theta$, como el axioma $\Box\varphi \to \Box\Box\varphi$ también está en Θ y este es maximal consistente, tenemos que $\Box\Box\varphi \in \Theta$. Por lo tanto $\Box\varphi \in \Psi$ (ya que $\Theta/\Box \subseteq \Psi$) y $\varphi \in \Phi$ (ya que $\Psi/\Box \subseteq \Phi$). Como φ era arbitrario, tenemos que $\Theta/\Box \subseteq \Phi$.

`Euclideana`: Sean $w_\Theta R^C w_\Psi$ y $w_\Theta R^C w_\Phi$ queremos ver que $w_\Psi R^C w_\Phi$, es decir, que $\Psi/\Box \subseteq \Phi$. Sea φ tal que $\Box\varphi \in \Psi$ y supongamos que $\varphi \notin \Phi$, entonces $\varphi \notin \Theta/\Box$ (porque $\Theta/\Box \subseteq \Phi$), i.e. $\Box\varphi \notin \Theta$. Como Θ es maximal consistente esto implica que $\neg\Box\varphi \in \Theta$. Por el axioma $\neg\Box\varphi \to \Box\neg\Box\varphi$ y la Proposición 2.6.6 ítem $2c$ también tenemos que $\Box\neg\Box\varphi \in \Theta$, y como $\Theta/\Box \subseteq \Psi$, $\neg\Box\varphi \in \Psi$ lo que contradice la consistencia de Ψ. □

Teorema 2.6.13. ***K*** *es completo respecto de la clase de modelos de Kripke.*

Demostración. Queremos ver que $\Vdash \varphi$ implica $\vdash \varphi$, o recíprocamente que $\nVdash \varphi$ implica que $\nvdash \varphi$, lo que también se puede expresar como $\nVdash \varphi$ implica que existe un mundo w de un modelo $\mathfrak{M}$ tal que $w \Vdash_{\mathfrak{M}} \neg\varphi$. Reemplazando φ por $\neg\varphi$ lo que queremos demostrar es que si φ es consistente entonces es satisfacible. Para ello alcanza con probar que todo conjunto consistente es satisfacible, y por la Proposición 2.6.6 ítem1 alcanza con probar que todo maximal consistente es satisfacible.

Para ello tomamos el modelo canónico, en el que hay un estado por cada maximal consistente, que lo satisface. Para este modelo vale $w_\Theta \Vdash_{\mathfrak{M}^C} \varphi$ sii $\varphi \in \Theta$, lo cual implica que $w_\Theta \Vdash_{\mathfrak{M}^C} \Theta$ quedando demostrada la completitud.

□

Con estos resultados preliminares ya estamos en condiciones de empezar a estudiar la lógica epistémica propiamente dicha.

Capítulo 3

Lógica epistémica y doxástica

Sólo sé que no sé nada

Sócrates.

Creo que creo lo que creo que no creo y creo que no creo lo que creo que creo

Oliverio Girondo.

La lógica epistémica es una lógica modal para razonar acerca de la actitud mental de conocimiento mientras que la lógica doxástica es para hacerlo sobre las creencias. Ambas disciplinas adquirieron popularidad en el área de Inteligencia Artificial para describir el conocimiento en robots, en otras áreas de Ciencias de la Computación para modelar la información distribuida en procesos de cómputos, así como en la Teoría de Juegos para modelar juegos de información imperfecta, en la Epistemología para resolver paradojas filosóficas, etcétera.

En este capítulo se van a introducir generalidades y conceptos básicos de las lógicas epistémicas y algunas variaciones, basándonos principalmente en [20], aunque también se tomaron ideas de [1], [10] y [5]. La idea detrás del enfoque modal de la noción de conocimiento es que además del estado real del mundo hay un montón de otros estados (o mundos) posibles. Dada la información disponible para un *agente epistémico* (un sujeto u objeto que posee información) podría suceder que éste no sea capaz de identificar en cuál de esos posibles estados se encuentra. Pero cuando una proposición vale en todos los mundos que considera posibles, entonces el agente sabe que la proposición vale aunque no sepa exactamente en cuál de esos mundos se encuentra.

Por ejemplo: un agente epistémico ha permanecido en un cuarto aislado sin relojes y con las ventanas cerradas por varios días de manera que pierde noción del tiempo (está escribiendo su tesis). Ya no sabe si es de día o de noche. Además de la incertidumbre de

no saber si es de día, el agente duda acerca de la potencial existencia de un dios o ente creador.

Decide salir a la calle y observa que es de noche. Este hecho descarta posibilidades que barajaba el agente. Ahora considera como mundos posibles sólo los siguientes: uno en el que es de noche y existe ente creador y otro en el que es de noche y no existe ente creador. En todos los mundos posibles (para este agente) es de noche, por lo tanto el agente **sabe** que es de noche. Sin embargo, no puede concluír nada acerca de la existencia de un ente creador, ya que considera posibles mundos en los que sí existe tanto como mundos en los que no.

En definitiva, si un agente no está seguro acerca del valor de verdad de una determinada proposición p (digamos, si existe un dios o ente creador) tiene que contar con la posibilidad de que p sea verdadero y con la posibilidad de que no.

Formalmente esto es capturado por un modelo de Kripke en el cual para un mundo el agente considera múltiples mundos alternativos (capturados por una relación de accesibilidad) para algunas de las cuales vale p mientras que para otras no.

Así, tenemos una primera semántica intuitiva: una fórmula φ es sabida en un mundo por un agente si todas las alternativas barajadas por él satisfacen φ. Dicho un poco más formalmente: un agente sabe φ en un estado w si en todos los estados accesibles vía una relación desde el estado w vale φ.

En definitiva, la información que un agente dispone en cada mundo posible le permitirá distinguirlo de algunos de los otros mundos. Esta información puede ser fehaciente, en tal caso dará lugar a conocimientos, o puede ser es posible/probable y bien justificada (pero aún posiblemente falsa) en cuyo caso dará lugar a creencias.

Ejemplo 3.0.14. Un agente epistémico arroja una moneda y la atrapa, pero no observa el resultado. La situación queda representada en el siguiente diagrama

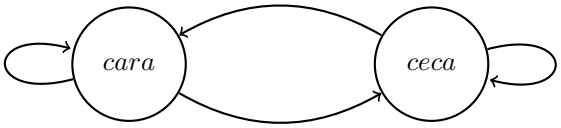


Figura 3.1: Ejemplo introductorio de a modelos epistémicos.

En este diagrama, las flechas representan la relación de *alternativa epistémica*: dado un mundo w, el agente sabe que se encuentra en alguna de sus alternativas epistémicas (pero no sabe en cuál). En otras palabras: si un mundo w está unido por una flecha al mundo v, significa que si el agente se encuentra en el mundo w no puede descartar hallarse en el mundo v. En este ejemplo en particular, dado que el agente no miró la moneda, si salió cara no puede descartar el caso en el que salió cara ni el caso en el que salió ceca. Pero si salió ceca tampoco puede descartar ninguno de los dos casos, porque no tiene ni la más mínima idea de qué lado obtuvo.

Las flechas del diagrama representan esa situación. Si el caso real (desconocido por el agente) fuese que salió ceca, el diagrama sería igual pero con el mundo real coloreado en turquesa:

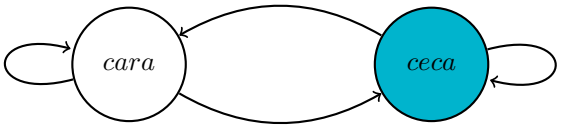


Figura 3.2: Ejemplo introductorio de a modelos epistémicos con mundo real indicado.

Ejemplo 3.0.15. Supongamos ahora que el agente arrojó la moneda, y la atrapó sin observar el resultado. El agente tiene tendencia a creer que salió cara por algún motivo (por ejemplo: porque antes le salieron 57 cecas seguidas). Vamos a representar esta situación en el siguiente diagrama:

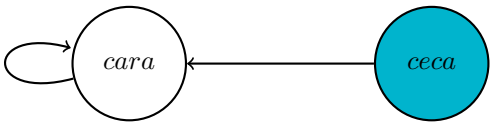


Figura 3.3: Ejemplo introductorio de a modelos doxásticos.

En este diagrama, las flechas representan la relación de *alternativa doxástica*: dado un mundo w el agente cree que se halla en alguno de los mundos relacionados con él. En este caso, si salió cara el agente va a creer que salió cara (porque es el único estado relacionado con el mundo en el que salió cara), y si salió ceca también va a creerlo (porque el mundo en el que salió cara es el único mundo relacionado con el mundo en el que salió ceca).

Avancemos hacia una formalización de esta intuición:

Definición 3.0.16. Llamamos *lenguaje epistémico* (notamos $\mathcal{L}_K$) al lenguaje modal básico renombrando $\Box$ como K y $\Diamond$ como $\hat{K}$. Llamamos *lenguaje doxástico* (notamos $\mathcal{L}_B$) al lenguaje modal básico renombrando $\Box$ como B y $\Diamond$ como $\hat{B}$.

Como mencionamos en el Capítulo 2 los lenguajes modales pueden ser construídos con los operadores modales existenciales y definir como abreviación a los operadores modales universales, o viceversa. En adelante siempre vamos a basarnos en los operadores modales universales, ya que se adaptan mejor a los contextos epistémicos.

Definición 3.0.17. Un *modelo para un lenguaje epistémico o modelo epistémico* es un modelo para tal lenguaje, donde se asume que la relación es de equivalencia.

Un *modelo para un lenguaje doxástico o modelo doxástico* es un modelo para tal lenguaje, donde se asume que la relación es euclideana, transitiva y serial.

En el caso de lógicas epistémicas, el conjunto de estados accesibles desde un estado fijo debe ser pensado como las alternativas epistémicas a él: si el agente en este estado no es capaz de distinguirlo de otro debido a su (falta de) conocimiento, entonces en lo que al agente concierne sus alternativas son iguales. De este modo, un agente sabe algo (K) cuando vale en todas sus alternativas epistémicas.

El motivo para modelar el conocimiento con una relación reflexiva para su accesibilidad puede ser entendido del siguiente modo: el agente, estando en un determinado estado, considera que puede estar en cualquiera de sus alternativas epistémicas. Si no considerase como alternativa epistémica el estado en el que se encuentra, entonces jamás va a poder conocer las verdades de su mundo sino las de aquellos con los que se relaciona.

El motivo de la transitividad es bastante simple: si estando en un estado w, el agente considera posible el estado v y en el estado v considera posible el estado u, como el agente *no puede descartar* encontrarse en el estado v también tiene que considerar las alternativas epistémicas de ese mundo. En particular, el agente tiene que contemplar la posibilidad de encontrarse en el estado u.

La motivación para incluír la simetría está relacionada con la hipótesis de indistinguibilidad de los mundos. En algunos modelos que veremos más adelante esta hipótesis se va a descartar pero, por lo pronto, vamos a considerar que para un agente un mundo y sus alternativas epistémicas son indistinguibles. Por lo tanto, si estando en w considera que puede estar en v o en w, estando en v tiene que contar con las mismas alternativas.

Por otro lado, para modelar creencias la condición de equivalencia para su relación puede ser demasiado fuerte, principalmente por la reflexividad. Por ejemplo: un agente puede no creer una proposición, pero que sin embargo esa proposición sea verdadera en el estado real. Es decir, no necesariamente desde un estado se considera como alternativa doxástica a ese mismo estado, resultando en una relación no necesariamente reflexiva.

Además la serialidad de la relación de accesibilidad para modelos doxásticos nos garantiza que el agente epistémico siempre considera que hay alguna alternativa doxástica, lo cual intuitivamente se corresponde con el hecho de que el agente toma siempre partido por al menos uno de los posibles mundos.

Un poco más complicada de motivar es la condición de euclideaneidad de la relación doxástica. Supongamos que cuando estamos en el mundo w creemos que podría suceder que estemos en el mundo v y también creemos que podría suceder que estemos en el mundo u. Si desde el mundo v no creyésemos posible al mundo u, esto diferencia al mundo w del mundo v (en cuanto a creencias se refiere). Tal caso parecería contradecir el hecho de creer posible encontrarnos en v cuando estamos en w. A pesar de parecer un trabalenguas, sugerimos al lector invertir unos minutos para pensar en la situación descripta arriba.

Las nociones de validez y verdad en los modelos epistémicos y doxásticos, son las mismas que para cualquier otro modelo de Kripke. Aplicando estas definiciones obtenemos la siguiente proposición:

Proposición 3.0.18. *Las siguientes fórmulas son válidas respecto de la clase de los modelos epistémicos:*

1. ***(K)*** $K(\varphi \to \psi) \to (K\varphi \to K\psi)$
2. ***(Veracidad)*** $K\varphi \to \varphi$
3. ***(Introspección positiva)*** $K\varphi \to KK\varphi$
4. ***(Introspección negativa)*** $\neg K\varphi \to K\neg K\varphi$

Demostración. 1) Vale por el simple hecho de usar semántica de Kripke, por lo visto en el Capítulo 2.

Por la Proposición 2.3.4, 2 y 3 valen por ser R reflexiva y transitiva. Y 4 vale, ya que una relación transitiva y simétrica es euclideana.

□

¿Qué representan estas fórmulas en lenguaje coloquial?

1. El conocimiento es cerrado bajo implicaciones: si se sabe que vale una implicación, sabiendo que vale el antecedente se puede saber que vale el consecuente. Por ejemplo: sé que si es más de las 14hs no quedan tartas en el sucucho, y sé que son más de las 14hs. No sólo se puede concluír que no hay tartas en el sucucho, sino que también se puede conluír que lo sé.
2. Esta fórmula expresa que el conocimiento es verdadero. Uno no puede saber honesta, verdadera y justificablemente algo falso.
3. Si un agente sabe algo, sabe que lo sabe.
4. Si un agente no sabe algo, sabe que no lo sabe.

La validez de estas fórmulas en la clase de modelos epistémicos es un ejercicio simple. Sin embargo la validez fórmulas en la vida real es discutible en muchos aspectos, pero eso no es objeto de estudio de la matemática ni de la lógica sino de la epistemología propiamente dicha, motivo por el cual no vamos a hacer hincapié en ello, pero remitimos a [24] donde hay consideraciones filosóficas al respecto.

Definición 3.0.19. Llamamos **S5** al sistema que resulta de tomar una axiomatización de la lógica proposicional y agregarle los esquemas de axiomas de la proposición anterior, con reglas de inferencia modus ponens y necesitación (de φ puedo obtener $K\varphi$).

Con respecto a lógica doxástica, obtenemos las siguientes fórmulas válidas en los modelos doxásticos:

1. **(K)** $B(\varphi \to \psi) \to (B\varphi \to B\psi)$
2. **(Consistencia)** $\neg B\bot$

3. **(Introspección positiva)** $B\varphi \to BB\varphi$
4. **(Introspección negativa)** $\neg B\varphi \to B\neg B\varphi$

¿Qué representan estas fórmulas en lenguaje coloquial?

1. Las creencias son cerradas bajo implicaciones: si se cree que vale una implicación, creyendo que vale el antecedente se cree que vale el consecuente.
2. Las creencias de un agente no son inconsistentes (pero esto no implica que sean verdaderas).
3. Si un agente cree algo, cree que lo cree.
4. Si un agente no cree algo, cree que no lo cree.

La diferencia entre las propiedades que cumple K y las que cumple B, radica en el hecho de que el conocimiento es veraz, mientras que la creencia sólo es consistente.

Definición 3.0.20. Llamamos **KD45** al sistema que consiste en una axiomatización para la lógica proposicional junto con los esquemas de axiomas recién enunciados. Las reglas de inferencia son modus ponens y necesitación para B.

COMPLETITUD, CORRECTITUD Y DECIBILIDAD

Tal como lo indica el título, vamos rumbo a resultados de completitud, correctitud y decibilidad. Previamente enunciamos el siguiente resultado que se desprende de la Proposición 2.6.12.

Corolario 3.0.21. *Sea $\mathfrak{M}^c$ el modelo canónico para **S5**, donde renombramos $\Box$ como K.*

Teorema 3.0.22. ***S5** es completo respecto de la clase de modelos epistémicos.*

Demostración. Si repetimos el razonamiento de la demostración del Teorema 2.6.13, nos encontramos con que sólo hace falta ver que $\mathfrak{M}^c$ es un modelo epistémico, es decir que R^c es de equivalencia. Y esto sucede por la Proposición 2.6.12. □

Corolario 3.0.23. ***S5** es correcto y completo respecto de la clase de modelos epistémicos.*

Demostración. Por Proposición 3.0.18 y Teorema 3.0.22. □

Definición 3.0.24. Un *modelo reducido* es un modelo $\mathfrak{M} = \langle W, \mathtt{i}\rangle$ para el lenguaje epistémico, donde la semántica usual para los operadores proposicionales y $w \Vdash_{\mathfrak{M}} K\varphi$ sii para todo $v \in W$ vale $v \Vdash_{\mathfrak{M}} \varphi$

Un *modelo de Kripke reducido* es igual agregándole una relación $R = W \times W$.

Definición 3.0.25. Sea $\mathfrak{M} = \langle W, R, \mathtt{i}\rangle$ un modelo epistémico y $w \in W$, llamamos *modelo reducido de $\mathfrak{M}$ para w* al modelo $\mathfrak{M}_{red} = \langle W', R', \mathtt{i}'\rangle$ donde $W' = \{v \in W : wRv\}$, $R' = W' \times W'$ e $\mathtt{i}' = \mathtt{i}|_{W'}$.

Observación 3.0.26. Un modelo reducido de un modelo $\mathfrak{M}$ para un estado w, es un modelo reducido.

Lema 3.0.27. *En los modelos reducidos vale que $w \Vdash K\varphi$ sii para todo v se tiene $v \Vdash K\varphi$.*

Proposición 3.0.28. *Dado un modelo $\mathfrak{M}$, un mundo w y el modelo reducido para ese mundo $\mathfrak{M}_{red}$, se tiene que $\mathfrak{M}, w \equiv \mathfrak{M}_{red}, w$.*

Esta proposición va en la dirección de buscar un resultado de decibilidad: para determinar si una fórmula es verdadera en un estado alcanza con mirar un modelo más chico (el de su clase de equivalencia).

Pero podemos ir un poco más lejos: podemos identificar los mundos en los que hay la misma asignación de verdad para las variables proposicionales. Es decir, dentro de cada modelo $\mathfrak{M} = \langle W, R, \mathtt{i}\rangle$ podemos definir una relación de equivalencia dada por $w \sim w'$ para w y w' tales que para todo $p \in \Phi$ vale $w \Vdash p$ sii $w' \Vdash p$. ¿Pero esto tiene sentido? La siguiente proposición nos lo garantiza.

Proposición 3.0.29. *Sea $\mathfrak{M} = \langle W, i\rangle$ un modelo reducido. Sean w y w' tales que para toda variable proposicional $p \in \Phi$ vale $w \Vdash p$ sii $w' \Vdash p$. Entonces para toda fórmula φ vale $w \Vdash \varphi$ sii $w' \Vdash \varphi$.*

Esta proposición muestra la redundancia de información en los modelos reducidos, por lo tanto la identificación de estados no sólo queda bien definida sino que además no pierde información.

Definición 3.0.30. Llamamos *modelo de Kripke simple para un conjunto de variables proposicionales Φ (o modelo de Kripke simple)* a un modelo $\mathfrak{M} = \langle W, R, \mathtt{i}\rangle$ donde $W \subseteq \mathcal{P}(\Phi)$, $R = W \times W$ y con $\mathtt{i}$ tal que $w \in \mathtt{i}(p)$ sii $p \in w$.

Definición 3.0.31. Llamamos *modelo simple obtenido de un modelo reducido* $\mathfrak{M}$ al modelo que resulta de $\mathfrak{M}$ luego de cocientar por la relación de equivalencia $\sim$. Vamos a identificar cada clase de equivalencia con el subconjunto de variables proposicionales que vale en ella, es decir, $W = \{C \subseteq \Phi :$ existe $w \in \mathfrak{M}$ con $w \in \mathtt{i}(p)$ sii $p \in C\}$.

Teorema 3.0.32. *Cada modelo reducido $\mathfrak{M}$ es bisimilar a un modelo simple $\mathfrak{M}'$. Más aún, la bisimulación puede tomarse como una bisimulación zigzag.*

Demostración. Sea $\mathfrak{M} = \langle W, R, \mathtt{i}\rangle$ un modelo de Kripke reducido. Como $\sim$ es de equivalencia (dada por $w \Vdash p$ sii $v \Vdash p$ para todo $p \in \Phi$) podemos tomar un representante de cada clase de equivalencia. Llamemos W' al conjunto de los representantes.

Para cada $w \in W'$, llamaremos $[w] := \{p \in \Phi \ : \ w \in \mathtt{i}(p)\}$ y $[W'] = \bigcup_{w \in W'}$. Tomamos el modelo simple de Kripke $\mathfrak{M}' = \langle [W'], R', \mathtt{i}'\rangle$, y definimos una relación $Z \subseteq \mathfrak{M} \times \mathfrak{M}'$ dada por $wZ[v]$ sii $[v] = [w]$. Que el dominio de Z es W y la imagen es $[W']$ no cabe duda. Veamos que es bisimulación.

forth) Sean $w \in W$ y $[v'] \in [W']$ tales que $wZ[v']$, y sea $u \in W$ tal que wRu . Sabemos que existe un $w' \in W'$ tal que $w' \sim u$, y para ese w' vale que $[w'] = [u]$, es decir, $uZ[w']$. Además como $[w'], [v'] \in [W']$, tenemos que $[v']R'[w']$.

back) Sean $w \in W$ y $[v'] \in [W']$ tales que $wZ[v']$, y sea $[u'] \in [W']$ tal que $[v']R'[u']$ (sucede para todo par en $[W']$). Como $u' \in W' \subseteq W$ y $\mathfrak{M}$ es un modelo reducido, entonces tenemos que wRu' y claramente vale $u'Z[u']$.

prop) Sean $w \in W$ y $[v'] \in [W']$. Entonces $wZ[v']$ sii $[w] = [v']$ sii $w \sim v'$ sii para todo $p \in \Phi$ vale que $w \Vdash p$ sii $v' \Vdash p$. □

Definición 3.0.33. Definimos *longitud de* φ (y notamos $|\varphi|$) inductivamente como $|p| = 1$ para $p \in \Phi$ (el conjunto de variables proposicionales), $|\neg\varphi| = |\varphi| + 1$, $|\varphi \wedge \psi| = |\varphi| + |\psi| + 1$ y $|K\varphi| = |\varphi| + 1$.

Proposición 3.0.34. *Una fórmula φ es satisfacible en modelos epistémicos sii φ es satisfacible en un mundo de un modelo que contiene a lo sumo $|\varphi|$ mundos.*

Demostración. $\Leftarrow$) es trivial. Probemos $\Rightarrow$).

Supongamos que φ es satisfacible en un mundo w de un modelo, y tomemos su modelo reducido $\mathfrak{M} = \langle W, R, \mathtt{i}\rangle$, entonces φ también es satisfacible en w para $\mathfrak{M}$. Por el Teorema 3.0.32 podemos asumir que $\mathfrak{M}$ es simple.

Llamemos $Sub(\varphi)$ a las subfórmulas de φ. Definimos una función $f : Sub(\varphi) \to \mathcal{P}(W)$, del siguiente modo:

- $f(K\psi) = \{v\}$ con v tal que $\Vdash_{\mathfrak{M}} \neg\psi$ si $w \nVdash_{\mathfrak{M}} K\psi$
- $f(K\psi) = \emptyset$ si $w \Vdash_{\mathfrak{M}} K\psi$
- $f(\neg\psi) = f(\psi)$
- $f(\varphi_1 \wedge \varphi_2) = f(\varphi_1) \cup f(\varphi_2)$
- $f(p) = \emptyset$

Y definimos $W' = \{w\} \cup \bigcup_{\psi \in Sub(\varphi)} f(\psi)$. Lo que hicimos fue crear un subconjunto $W' \subseteq W$ del siguiente modo: primero colocamos w en W', luego agregamos mundos tales que $\neg\psi$ para los ψ en las subfórmulas de φ de la forma $K\psi$ que no son verdaderos en w.

Ahora definimos un modelo $\mathfrak{M}' = \langle W', R', \mathtt{i}' \rangle$ con $R' = W' \times W'$ e $\mathtt{i}' = \mathtt{i}|_{W'}$. Observemos que $\mathfrak{M}'$ es un modelo simple (es un submodelo de uno simple). Como f agrega a lo sumo 1 elemento a W' por cada $K\psi \in Sub(\varphi)$, tenemos que $|W'| \leq |\varphi|$.

Además tenemos que para todo $\psi \in Sub(\varphi)$ y cada $v \in W'$ vale $v \Vdash_{\mathfrak{M}} \psi$ sii $v \Vdash_{\mathfrak{M}'} \psi$. Esto lo podemos ver inductivamente: si $\psi = p \in \Phi$ es trivial ($\mathfrak{M}'$ es simple, entonces los mundos se identifican con sus valuaciones), los casos de conectivos lógicos porque las restricciones no afectan su valor de verdad. Consideremos el caso $\psi = K\alpha$, y supongamos que ya fue demostramos para α.

Sea $u \in W'$ tal que $u \Vdash_{\mathfrak{M}'} K\alpha$ y supongamos que no vale $u \Vdash_{\mathfrak{M}} K\alpha$, entonces existe un $t \in W$ tal que uRt y $t \Vdash \neg\alpha$. Como $w \in W$ y $\mathfrak{M}$ es un modelo reducido (de hecho es simple) tenemos que $w \Vdash_{\mathfrak{M}} \neg K\alpha$, entonces por construcción de f tenemos un estado $v \in W'$ tal que $v \nVdash_{\mathfrak{M}} \alpha$. Por hipótesis inductiva $v \nVdash_{\mathfrak{M}'} \alpha$ y como $\mathfrak{M}'$ también es un modelo reducido tenemos que $u \nVdash_{\mathfrak{M}'} K\alpha$, lo que contradice la hipótesis de la que partimos.

Ahora sea $u \in W'$ tal que $u \Vdash_{\mathfrak{M}} K\alpha$. Esto implica que para todo $v \in W$ vale $v \Vdash_{\mathfrak{M}} \alpha$, pero $W' \subseteq W$ entonces $x \Vdash_{\mathfrak{M}} \alpha$ para todo $x \in W'$. Por hipótesis inductiva (que vale para cualquier estado) tenemos que para todo $x \in W'$ vale $x \Vdash_{\mathfrak{M}'} \alpha$, como $\mathfrak{M}'$ es un modelo reducido tenemos que $u \Vdash_{\mathfrak{M}'} K\alpha$. □

Corolario 3.0.35. ***S5*** *es decidible.*

Demostración. Por completitud sabemos que si φ no es demostrable por **S5**, entonces su negación $\neg\varphi$ es satisfacible. Pero por lo que vimos si es satisfacible alcanza con chequear los modelos que contienen a lo sumo $|\neg\varphi|$ mundos para chequear si se satisface, por lo cual el proceso es finito. □

3.1. Lógica epistémica multiagente

Hasta ahora hemos definido y considerado todo en términos de un único agente epistémico. En esta sección vamos a generalizar al caso de múltiples.

Para empezar, la primera modificación que se puede hacer es introducir operadores epistémicos (respectivamente doxásticos) para cada agente. Se tienen operadores indexados K_i y B_i para el conocimiento y las creencias del agente i, pero uno puede ir un poco más lejos y tratar de definir de algún modo operadores que involucren a un grupo de agentes y la interacción de información de distintos agentes. Esto da lugar a las nociones de conocimiento común y conocimiento compartido.

La noción más simple es la de conocimiento compartido: $E_K\varphi$ se lee como "todos saben φ". Pero uno también puede agregar un operador C_K para el conocimiento común, que es mucho más fuerte: que φ sea conocimiento común no sólo significa que todos saben φ sino que además todos saben que todos saben φ, y que todos saben que todos saben que todos saben φ, etcétera.

El lenguaje es el mismo que el de una lógica epistémica extendida del siguiente modo:

Definición 3.1.1. Las fórmulas bien formadas del lenguaje epistémico multimodal son

$$\varphi ::= p \mid \bot \mid \neg\varphi \mid \psi \vee \varphi \mid K_i\varphi \mid E_K\varphi \mid C_K\varphi$$

donde p es alguna variable proposicional (i.e. $p \in \Phi$).

Definición 3.1.2. Un *modelo para n agentes epistémicos con conocimiento común* está dado por

$$\mathfrak{M} = \langle W, R_1 \ldots, R_n, R_E, R_C, \mathtt{i} \rangle$$

donde

- W es un conjunto de mundos no vacío
- $\mathtt{i}$ es una función de asignación de estados en los que es verdad cada variable proposicional
- R_i son las relaciones (de equivalencia) en W para interpretar los operadores modales K_i
- $R_E = \bigcup_i R_i$
- $R_C = R_E^*$ la clausura transitiva y reflexiva de R_E

Denotamos $\mathcal{S}5\mathcal{EC}_n$[1] a la clase de modelos epistémicos de n agentes.

Definición 3.1.3. Sea $\mathfrak{M} = \langle W, R_1, \ldots, R_n, R_E, R_C, \mathtt{i} \rangle$ un modelo para n agentes epistémicos. Llamamos un *camino* en $\mathfrak{M}$ a una secuencia finita de estados relacionados por las relaciones R_i: $w_0 R_{i_1} w_1 R_{i_2} \ldots R_{i_k} w_k$.

Se dice que dos estados w y v están *conectados por un camino* si existe un camino que empieza en w y termina en v.

Dada una fórmula φ, llamamos *camino-φ* a un camino en el que todos los w_i cumplen $w_i \Vdash \varphi$.

Observación 3.1.4. Dos mundos están conectados vía la relación R_C sii existe un camino que los conecta.

Como es esperable, a los operadores K_i se les da la semántica del operador modal universal para la relación R_i, a E la del operador modal universal para R_E y a C la del operador modal universal para R_C.

Proposición 3.1.5. *Usando las nociones análogas de validez tenemos que:*

1. ***(Todos)*** $E_K\varphi \leftrightarrow K_1\varphi \wedge \ldots \wedge K_n\varphi$
2. ***(Veracidad C)*** $C_K\varphi \rightarrow \varphi$
3. ***(K para C)*** $C_K(\varphi \rightarrow \psi) \rightarrow (C_K\varphi \rightarrow C_K\psi)$

[1] Esta nomenclatura viene de que son modelos caracterizados por el sistema axiomático **S5EC**, que veremos más adelante, el cuál es una extensión de **S5** con axiomas para E y C

4. ***(Común a todos)*** $C_K\varphi \to E_K C_K \varphi$
5. ***(Escalar)*** $C_K(\varphi \to E_K\varphi) \to (\varphi \to C_K\varphi)$
6. ***(introspección positiva para C)*** $C_K\varphi \to C_K C_K \varphi$
7. ***(introspección negativa para C)*** $\neg C_K\varphi \to C_K \neg C_K \varphi$

(todos) muestra que la modalidad "todos saben" es, de hecho, lo que sugiere su nombre. **(veracidad para C), (K para C), (introspección positiva) e (instrospección negativa)** afirman que el conocimiento común tiene al menos la propiedades del conocimiendo individual: es verídico, cerrado por implícación e introspectivo. **(común a todos)** dice que el conocimiento común es sabido por todo el mundo. **(escalar)** es una especie de principio de inducción: el antecedente da la condición bajo la cual uno puede escalar una verdad φ a la categoría de conocimiento común, ese antecedente es que es conocimiento común que si φ es verdad es sabido por todos.

Nota. Por supuesto, se pueden definir las nociones de "todos creen" y "creencia común" para una lógica doxástica multiagente, introduciendo de manera completamente análoga los operadores E_B y C_B sujeto a las relaciones $R_F = \bigcup_i R_i$ (donde R_i representa las relaciones de alternativas doxásticas para el agente i, por lo tanto es serial, transitiva y euclideana), y $R_D = R_F^+$ la clausura transitiva (pero no reflexiva) de R_F. Esto se debe a que creencia común no necesariamente es verdadera.

Definición 3.1.6. Definimos el sistema $\mathbf{S5EC}_n$ como una copia de **S5** para cada operador K_i y agregarle como axiomas **todos**, **veracidad para C**, **K para C**, **común a todos** y **escalar**. Las reglas de inferencia son modus ponens, y necesitación para K_i y para C.

Cuando n queda claro por contexto notamos simplemente **S5EC**.

Proposición 3.1.7. ***S5EC*** *es correcto respecto de los modelos epistémicos multiagentes.*

Demostración. Se desprende de que los modelos epistémicos multiagentes validan los axiomas y de que las reglas de inferencia preservan validez.

□

COMPLETITUD

Ahora que tenemos más modalidades hay que ver si nuestra demostración de completitud se adapta al nuevo escenario.

Definición 3.1.8. Definimos el *modelo canónico para n agentes*, de manera análoga a la Definición 2.6.10, como $\mathfrak{M}^c = \langle W^c, R_1^c, \ldots, R_n^c, R_C^c, R_E^c, \mathtt{i}^c \rangle$ donde W^c e $\mathtt{i}^c$ son los de antes, $R_i^c = \{(w_\Theta, w_\Psi) : \text{con } \Theta, \Psi \text{ maximales consistentes y } \Theta/K_i \subseteq \Psi\}$, y R_E^c y R_C^c se definen análogamente.

Repitiendo el procedimiento de la sección anterior se puede ver que este modelo cumple que para toda φ vale $w_\Theta \Vdash_{\mathfrak{M}^c} \varphi$ sii $\varphi \in \Theta$, pero lo que faltaría ver es que el modelo es un modelo epistémico multiagente y, lamentablemente no lo es.

Teorema 3.1.9. *El modelo canónico para sistemas multiagente satisface:*

1. R_i^c *es de equivalencia*
2. $R_E^c = R_1^c \cup \ldots \cup R_n^c$
3. $(R_E^c)^* \subseteq R_C^c$

Demostración. 1) Exactamente del mismo modo que en la demostración de completitud para **S5**.

2 $\subseteq$) Sean w_Θ y w_Ψ, con Θ y Ψ maximales consistentes, tales que $w_\Theta R_E^c w_\Psi$, entonces $\Theta/E_k \subseteq \Psi$. Supongamos que no existe i tal que $w_\Theta R_i w_\Psi$, es decir, que para todo i vale que $\Theta/K_i \not\subseteq \Psi$. Esto es lo mismo que decir que para todo i existe un $K\varphi_i \in \Theta$ pero $\varphi_i \notin \Psi$, entonces por la consistencia maximal de Ψ tenemos que $\neg\varphi_i \in \Psi$ y en consecuencia $\neg\varphi_1 \wedge \ldots \wedge \neg\varphi_n \in \Psi$.

Sea $\varphi = \varphi_1 \vee \ldots \vee \varphi_n$, como $\vdash \varphi_i \to \varphi$ se tiene (por necesitación de K_i) $\vdash K_i(\varphi_i \to \varphi)$. Y por esto último junto con el axioma **K** y modus ponens $\vdash K_i\varphi_i \to K_i\varphi$, que está en Θ por su consistencia maximal. Pero además $K_i\varphi_i \in \Theta$ con lo que tenemos que $K_i\varphi \in \Theta$ para todo i, resultando en $K_1\varphi_1 \wedge \ldots \wedge K_n\varphi_n \in \Theta$ (porque Θ es maximal consistente). Usando el axioma **todos** tenemos que $E_k\varphi \in \Theta$, es decir $\varphi \in \Theta/E \subseteq \Psi$. Pero al principio demostramos que $\neg\varphi \in \Psi$, absurdo!

2 $\supseteq$) Supongamos que $w_\Theta(R_1^c \cup \ldots \cup R_n^c)w_\Psi$, con Θ y Ψ maximales. Supongamos que están relacionados por R_j^c, entonces $\Theta/K_j \subseteq \Psi$. Queremos ver que $\Theta/E_k \subseteq \Psi$. Sea $\varphi \in \Theta/E_k$, es decir $E_k\varphi \in \Theta$, por el axioma **todos** y la consistencia maximal de Θ tenemos que $K_1\varphi \wedge \ldots \wedge K_n\varphi \in \Theta$. Esto implica que $K_i\varphi \in \Theta$ para todo i, en particular para $i = j$ es decir $\varphi \in \Theta/K_j \subseteq \Psi$.

3 Sea $w_\Theta(R_E^c)^* w_\Psi$, esto es que existe una cadena finita $w_\Theta = w_{\Delta_1}, \ldots, w_{\Delta_r} = w_\Psi$ tal que $\Delta_i/E_k \subseteq \Delta_{i+1}$ para todo $0 \leq i < r$. Queremos ver que $w_\Theta R_C^c w_\Psi$, es decir, que $\Theta/C_k \subseteq \Psi$. Sea $\varphi \in \Theta/C_k$, entonces por definición vale $C_k\varphi \in \Theta$. Esto junto con el axioma **común a todos** y la maximalidad de Θ resulta en que $E_kC_k\varphi \in \Theta = \Delta_0$. Como $\Delta_0/E_k \subseteq \Delta_1$, tenemos que $C_k\varphi \in \Delta_1$. Repitiendo el argumento $r-1$ veces obtenemos $\varphi \in \Psi$. □

Proposición 3.1.10. *Para el modelo canónico multiagente se tiene que* $R_C^c \not\subseteq (R_E^c)^*$.

Demostración. Definimos inductivamente $E_k^n\varphi$ del siguiente modo $E_k^0\varphi = \varphi$ y $E_k^{n+1} = E_kE_k^n\varphi$. Consideramos $\Phi = \{\neg C_kp\} \cup \{E_kp, E_k^2p, E_k^3p, \ldots\}$, y veamos que es consistente.

Supongamos que no es consistente, entonces hay un $\Delta \subseteq \Phi$ finito tal que es inconsistente, digamos $\Delta = \{E_k^{n_1}, \ldots, E_k^{n_r} \ : \ \text{con } n_i < n_{i+1}\} \cup \{\neg Cp\}$. Como Δ es inconsistente, $\psi := E_k^{n_1} \wedge \ldots \wedge E_k^{n_r} \wedge \neg C_kp$ lo es y por la correctitud de $\mathbf{S5EC}_n$ tenemos que ψ es insastisfacible.

Sin embargo, consideremos el siguiente modelo: $\mathfrak{M} = \langle \mathbb{N}, R_1, R_2, R_E, R_C, \mathtt{i} \rangle$, donde $R_1 = \{(n,n) \ : \ n \in \mathbb{N}\} \cup \{(n, n+1) \ : \ n \text{ es par}\} \cup \{(n+1, n) \ : \ n \text{ es par}\}$, $R_2 = \{(n,n) \ : \ n \in \mathbb{N}\} \cup \{(n, n+1) \ : \ n \text{ es par}\} \cup \{(n+1, n) \ : \ n \text{ es impar}\}$, $R_E = R_1 \cup R_2$ y $R_C = R_E^*$. Definimos la función de interpretación como $\mathtt{i}(p) = \{n \in \mathbb{N} \ : \ n \leq n_r + 1\}$. Para este modelo vale $1 \Vdash_{\mathfrak{M}} \psi$.

Acabamos de ver la consistencia de Φ, entonces por la proposición 2.6.6 tenemos que $\Phi \subseteq \Phi'$ con Φ' maximal consistente. Sea $\Psi = \{\neg p\} \cup \{\psi \ : \ C\psi \in \Phi'\}$, resulta consistente (porque $p \notin \Phi'$) por lo que existe un maximal consistente $\Psi' \supseteq \Psi$.

Tenemos que en el modelo canónico $w_{\Phi'} R_C^c w_{\Psi'}$ pero no $w_{\Phi'} (R_E^c)^* w_{\Psi'}$.

□

El modelo canónico no está en $\mathcal{S}5\mathcal{EC}$, ya que R_C^c no es la clausura reflexiva y transitiva de R_E^c, por lo tanto buscamos una variación de este modelo que nos permita conservar algunas propiedades para poder alcanzar un teorema de completitud como en el caso de agente único. Para ello introducimos algunas definiciones y resultados previos, todos tomados de [20]:

Definición 3.1.11. Dada una fórmula φ definimos $\Phi_1 = \{\psi \ : \ \psi \text{ es subfórmula de } \varphi\}$, $\Phi_2 = \{K_i\psi, \neg K_i\psi \ : \ E_k\psi \in \Phi_1\}$, $\Phi_3 = \{E_k C_k \psi, \neg E_k C_k \psi, K_i C_k \psi, \neg K_i C_k \psi \ : \ C\psi \in \Phi_1\}$. Llamamos a $adec(\varphi) = \Phi_1 \cup \Phi_2 \cup \Phi_3$ el *conjunto adecuado de subfórmulas de* φ.

Observación 3.1.12. $adec(\varphi)$ contiene a φ, es finito y cerrado por subfórmulas, para cada fórmula $\psi \notin adec(\varphi)$ hay una fórmula $\chi \in adec(\varphi)$ que es equivalente a $\neg\psi$, para cada $C_k\psi \in adec(\varphi)$ también se tiene $E_k C_k \psi \in adec(\varphi)$ y para cada $E\psi \in adec(\varphi)$ tenemos $K_i\psi \in adec(\varphi)$.

Definición 3.1.13. Sea $\mathfrak{M}^c = \langle W^c, R_1^c, \ldots, R_n^c, R_E^c, R_C^c, \mathtt{i} \rangle$ el modelo canónico multiagente y sea φ una fórmula consistente. Definimos la *filtración del modelo canónico a través de* $adec(\varphi)$ como el modelo $\mathfrak{M}_\varphi^c = \langle \bar{W}, \bar{R}_1, \ldots, \bar{R}_n, \bar{R}_E, \bar{R}_C, \bar{\mathtt{i}} \rangle$ dado por:

- $\bar{W} = \{[w] \ : \ [w]_{adec(\varphi)} \text{ es clase de equivalencia de } \equiv_{adec(\varphi)}\}$
- $\bar{\mathtt{i}}(p) = \mathtt{i}(p)$
- Para todo $i \leq n$ valen: $[w]\bar{R}_i[v]$ sii para todo $K\psi \in adec(\varphi)$ se tiene que $w \Vdash_{\mathfrak{M}^c} K_i\psi$ es equivalente a $v \Vdash_{\mathfrak{M}^c} K_i\psi$
- $\bar{R_E} = \bar{R}_1 \cup \ldots \cup \bar{R}_n$
- $\bar{R_C} = (\bar{R_E})^*$

donde $\equiv_{adec(\varphi)}$ está definida como en 2.5.4.

Lema 3.1.14. *El modelo* $\mathfrak{M}_\varphi^c$ *es un modelo epistémico multiagente.*

Demostración. Sólo resta probar que las relaciones $\bar{R}_i$ son de equivalencia y esto es inmediato por la definición de $\bar{R}_i$. □

Lema 3.1.15. *Sea $\bar{W}$ el conjunto de mundos de la filtración del modelo canónico. Para cada $A \subseteq \bar{W}$ existe una fórmula σ_A tal que para toda $[w] \in \bar{W}$ vale $w \Vdash_{\mathfrak{M}^c} \sigma_A$ sii $[w] \in A$.*

Demostración. Sea $Form_\varphi(w) := \{\phi \in adec(\varphi) \ : \ w \Vdash_{\mathfrak{M}^c} \phi\}$. Por definición de $[w]$ vale que $v \Vdash_{\mathfrak{M}^c} Form_\varphi(w)$ sii $[w] = [v]$. Tomamos $\sigma_A := \bigvee_{[v] \in A} Form_\varphi(v)$, que está bien definida porque A es finito ($A \subseteq \bar{W}$, y $\bar{W}$ es finito). Entonces tenemos que:

$$w \Vdash_{\mathfrak{M}^c} \sigma_A \text{ sii } w \Vdash_{\mathfrak{M}^c} \bigvee_{[v] \in A} Form_\varphi(v)$$

$$\text{sii } w \Vdash_{\mathfrak{M}^c} Form_\varphi(v) \text{ para algún } v \text{ con } [v] \in A \text{ sii } [v] \in A.$$

□

Teorema 3.1.16. *Sea φ una fórmula consistente, el modelo $\mathfrak{M}^c_\varphi$ es efectivamente una filtración de $\mathfrak{M}^c$ a través de $adec(\varphi)$.*

Demostración. Queremos ver que la relación $\bar{R}_i$ es una filtración de R^c_i a través de $adec(\varphi)$ para todo i, y análogamente para R^c_E y R^c_C.

- Los $\bar{R}_i$ son filtraciones de R^c_i a través de $adec(\varphi)$

Sea $i \leq n$, veamos que R^c_i satisface $\mathbf{min}(\bar{R}_i/R^c_i)$: sean $[w], [v] \in \bar{W}$ y $w', v' \in W^C$ con $w' R^c_i v'$, $[w] = [w']$ y $[v] = [v']$, y sea $K_i\psi \in adec(\varphi)$. Tenemos que:

$w \Vdash_{\mathfrak{M}^c} K_i\psi$ sii ($[w] = [w']$) $w' \Vdash_{\mathfrak{M}^c} K_i\psi$ sii (R^c_i es de equivalencia y $w' R_i v'$) $v' \Vdash_{\mathfrak{M}^c} K_i\psi$ sii ($[v] = [v']$) $v \Vdash_{\mathfrak{M}^c} K_i\psi$.

Entonces vale $w \bar{R}_i v$, es decir, se cumple $\mathbf{min}(\bar{R}_i/R^c_i)$. Ahora veamos que se cumple $\mathbf{max}(\bar{R}_i)$:

Sean $[w], [v] \in \bar{W}$ con $[w]\bar{R}_i[v]$ y $w \Vdash_{\mathfrak{M}^c} K_i\psi$. Por definición de $\bar{R}_i$ se tiene que $v \Vdash_{\mathfrak{M}^c} K_i\psi$ y como $\bar{R}_i$ es reflexiva también tenemos que $v \Vdash_{\mathfrak{M}^c} \psi$.

- $\bar{R_E}$ es filtración de R^c_E a través de $adec(\varphi)$

Veamos que vale $\mathbf{Min}(\bar{R_E}/R^c_E)$, para ello tomamos w, v tales que $w R^c_E v$. Como vimos que $R^c_E = R^c_1 \cup \ldots \cup R^c_n$, existe un i tal que $w R^c_i v$. Como $\bar{R}_i$ es filtración de R^c_i a través de $adec(\varphi)$ tenemos que $[w]\bar{R}_u[v]$ y por definición de $\bar{R}_E$ resulta que $[w]\bar{R_E}[v]$.

Ahora veamos que vale $\mathbf{Max}(\bar{R_E})$: sean $[w]$ y $[v]$ tales que $[w]\bar{R_E}[v]$ y $w \Vdash_{\mathfrak{M}^c} E\psi$ con $E\psi \in adec(\varphi)$. Como $adec(\varphi)$ es el conjunto adecuado para φ tenemos que $K_i\psi \in adec(\varphi)$ para todo i y, dado que $\bar{R}_i$ es una filtración de R^c_i a través de $adec(\varphi)$, tenemos que $v \Vdash_{\mathfrak{M}^c} \psi$ (que es lo que queríamos ver).

- $\bar{R_C}$ es una filtración de R^c_C a través de $adec(\varphi)$

Veamos que vale $\mathbf{Min}(\bar{R}_C/R^c_C)$, para ello sean w y v tales que wR^c_Cv y sea $A = \{[u] \in \bar{W} : [w]\bar{R}_E{}^*[u]\}$. Si vemos que $w \Vdash_{\mathfrak{M}^c} C_k\sigma_A$ tendríamos que $v \Vdash_{\mathfrak{M}^c} \sigma_A$, que es lo mismo que $[v] \in A$, que a su vez significa que $[w]\bar{R}_E{}^*[v]$ y recordemos que $\bar{R}_C = \bar{R}_E{}^*$.

Así que vamos a demostrar que $w \Vdash_{\mathfrak{M}^c} C_k\sigma_A$, para ello usaremos el hecho de que $\mathfrak{M}^c$ valida el esquema **escalar** instanciado en σ_A. En particular:

$$w \Vdash_{\mathfrak{M}^c} C_k(\sigma_A \to E\sigma_A) \to (\sigma_A \to C\sigma_A),$$

y ahora veremos que además vale el antecedente.

Sea x tal que wR^c_Cx y $x \Vdash_{\mathfrak{M}^C} \sigma_A$ y queremos ver que $x \Vdash_{\mathfrak{M}^c} E^c_k\sigma_A$. Tomamos y tal que xR^c_Ey y como $x \Vdash_{\mathfrak{M}^c} \sigma_A$ se tiene $[x] \in A$ con lo que $[w]\bar{R}_E{}^*[x]$. Esto es que existe n tal que $[w]\bar{R}^n_E[x]$, pero como $\bar{R}_E$ es filtración de R^c_E tenemos que $[x]\bar{R}_E[y]$, por lo tanto $[w]\bar{R}_E{}^{n+1}[y]$. Entonces $[y] \in A$ y en consecuencia $y \Vdash_{\mathfrak{M}^c} \sigma_A$.

Ya tenemos que $w \Vdash_{\mathfrak{M}^c} C_k(\sigma_A \to E_k\sigma_A) \to (\sigma_A \to C_k\sigma_A)$ y $w \Vdash_{\mathfrak{M}^c} C_k(\sigma_A \to E_k\sigma_A)$, y además vale que $w \Vdash_{\mathfrak{M}^c} \sigma_A$ (poque $[w] \in A$, y esto significa que $w \Vdash_{\mathfrak{M}^c} \sigma_A$), entonces se tiene que $w \Vdash_{\mathfrak{M}^c} C_k\sigma_A$.

Ahora veamos que vale $\mathbf{Max}(\bar{R}_C)$, para ello tomamos w y v tales que $[w]\bar{R}_C[v]$ y $w \Vdash_{\mathfrak{M}^c} C\psi$ con $C\psi \in adec(\varphi)$. Queremos ver que $v \Vdash_{\mathfrak{M}^c} \psi$. Por definición tenemos que existe un n tal que $[w]\bar{R}_E{}^n[v]$, esto significa que exiten $[v_0]\bar{R}_E[v_1]\bar{R}_E \ldots \bar{R}_E[v_n]$ con $[w] = [v_0]$ y $[v_n] = [v]$. Como el modelo $\mathfrak{M}^c$ cumple **común a todos)**, en particular se tiene que $C_k\psi \to E_kC_k\psi$). Pero además tenemos $w \Vdash_{\mathfrak{M}^c} C_k\psi$, por lo tanto se desprende que $w \Vdash_{\mathfrak{M}^c} E_kC_k\psi$. Es más, como $C_k\psi \in adec(\varphi)$ y $adec(\varphi)$ es el conjunto adecuado para φ también $E_kC_k\psi \in adec(\varphi)$. Esto, junto con el hecho de que $\bar{R}_E$ es una filtración de R^c_E a través de $adec(\varphi)$, nos garantiza que $v_i \Vdash_{\mathfrak{M}^c} E_kC_k\psi$ implica $v_{i+1} \Vdash_{\mathfrak{M}^c} C_k\psi$ (por la condición $\mathbf{Max}(\bar{R}_E)$).

En conclusión, se tiene que $v_i \Vdash_{\mathfrak{M}^c} C_k\psi$ implica $v_{i+1} \Vdash_{\mathfrak{M}^c} C_k\psi$. Aplicando esto n veces tenemos que de $w \Vdash_{\mathfrak{M}^c} C_k\psi$ se desprende $v \Vdash_{\mathfrak{M}^c} C_k\psi$, y como R^c_C es reflexiva obtenemos que $v \Vdash_{\mathfrak{M}^c} \psi$.

□

Corolario 3.1.17. *Sea φ una fórmula, se tiene que $\vdash_{\boldsymbol{S5EC}_n} \varphi$ sii $\Vdash_{\mathcal{S}5\mathcal{EC}_n} \varphi$.*

Demostración. $\Rightarrow$) se desprende de aplicar definiciones, veamos $\Leftarrow$): Supongamos que $\nvdash_{\mathbf{S5EC}} \varphi$, i.e. $\neg\varphi$ es consistente, por lo tanto $\neg\varphi \in \Theta$ para algún maximal consistente Θ. Esto último es equivalente a $w_\Theta \Vdash_{\mathfrak{M}^c} \neg\varphi$. Pero podemos tomar el modelo $\mathfrak{M}^c_\varphi$, que es una filtración de $\mathfrak{M}^c$ a través de $adec(\varphi)$ entonces por el Teorema 2.5.7 tenemos que $w \Vdash_{\mathfrak{M}^c_\varphi} \varphi$. Entonces $\nVdash_{\mathcal{S}5\mathcal{EC}_n} \varphi$. □

Corolario 3.1.18. ***S5EC*** *tiene la propiedad de modelo finito. Es decir, toda fórmula consistente con* ***S5EC*** *es satisfacible en algún modelo finito.*

Demostración. Se desprende de que las filtraciones son finitas. □

Corolario 3.1.19. ***S5EC*** *es dicidible*

Demostración. Es consecuencia de que **S5EC** tenga finitos esquemas de axiomas y cumpla la propiedad de modelo finito.

□

En algunos contextos no precisamos tener operadores E y C (se verá en capítulos posteriores). Definimos los modelos y el sistema axiomático correspondiente a esas lógicas:

Definición 3.1.20. Definimos los *modelos epistémicos para n agentes aislados* como modelos de Kripke $\mathfrak{M} = \langle W, R_1, \ldots, R_n, \mathtt{i} \rangle$ donde las relaciones son todas reflexivas.

Definición 3.1.21. Definimos el sistema axiomático $\mathbf{S5}_n$ que consiste en los axiomas de **S5** para cada operador K_i, junto con las reglas de inferencia modus ponens y necesitación para K_i.

Teorema 3.1.22. *$\boldsymbol{S5}_n$ es correcto y completo respecto de los modelos epistémicos para n agentes aislados.*

Demostración. Consiste en repetir la demostración del caso de un único agente tomando un i arbitrario y utilizando los operadores K_i. □

3.2. Interacción de K y B

Hemos mencionado los modelos epistémicos por un lado, los doxásticos por otro, pero una pregunta natural es si las modalidades de conocimiento y creencia están relacionadas de algún modo significativo. La respuesta es más complicada de lo que puede parecer a primera vista. Desde el nacimiento de estas modalidades se investigaron distintos tipos de interacciones. Resulta ser que hay que ser cuidadoso en poner varias propiedades de interacción juntas porque uno termina teniendo propiedades no deseadas en su sistema, por ejemplo: el colapso de conocimiento y creencia.

Definición 3.2.1. Llamamos *frame epistémico-doxástico (o frame KB)* a una terna $\mathfrak{F} = \langle W, S, T \rangle$ que satisface:

- S es una relación de equivalencia
- T es serial
- $T \subseteq S$
- T es transitiva sobre (S, T) (i.e. si wSv y vTu entonces wTu)

Estos frames dan lugar a modelos para un lenguaje epistémico doxástico, donde K es el operador modal universal inducido por S, y B el inducido por T. Llamamos $\mathcal{KB}$ a la clase de frames epistémico-doxásticos. Es un ejercicio simple ver que en $\mathcal{KB}$ son válidos los axiomas de **S5** (para K) y de **KD45** (para B).

Definición 3.2.2. Definimos el *sistema epistémico-doxástico* **KB** para un lenguaje modal con operadores K y B que consiste en una axiomatización para lógica proposicional más:

1. $K(\varphi \to \psi) \to (K\varphi \to K\psi)$
2. $K\varphi \to \varphi$
3. $\neg K\varphi \to K\neg K\varphi$
4. $B(\varphi \to \psi) \to (B\varphi \to B\psi)$
5. $B\varphi \to BB\varphi$
6. $\neg B\varphi \to B\neg B\varphi$
7. $\neg B\bot$
8. $K\varphi \to B\varphi$
9. $B\varphi \to KB\varphi$

junto con las reglas de inferencia modus ponens, necesitación para K y necesitación para B.

Estos axiomas son los que ya conocemos para creencias y conocimiento junto con dos nuevos axiomas de interacción de las modalidades: el axioma 8 que expresa que el conocimiento es más fuerte que las creencias, y el axioma 9 cuya interpretación es que las creencias son conscientes, siendo ambas propiedades de interés para la interacción entre K y B.

Ejemplo 3.2.3. La introspección positiva no está explícitamente mencionada en **KB** pero vamos a ver que es **KB**-demostrable. Para no entorpecer la lectura, vamos a utilizar la transitividad de $\to$ sin demostrarla.

1. $K\neg K\varphi \to \neg K\varphi$ (Veracidad aplicada $\neg K\varphi$)
2. $K\varphi \to \neg K\neg K\varphi$ (Axioma del contrarrecíproco aplicado al ítem 1)
3. $\neg K\neg K\varphi \to K\neg K\neg K\varphi$ (Introspección negativa aplicada a $\neg K\varphi$)
4. $\neg K\varphi \to K\neg K\varphi$ (Introspección negativa)
5. $\neg K\neg K\varphi \to K\varphi$ (Contrarrecíproco aplicado a 4)
6. $K(\neg K\neg K\varphi \to K\varphi)$ (Necesitación de K)
7. $K(\neg K\neg K\varphi \to K\varphi) \to (K\neg K\neg K\varphi \to KK\varphi)$ (Axioma K aplicado a 6)
8. $K\neg K\neg K\varphi \to KK\varphi$ (Modus ponens de 6 y 7)

9. $K\varphi \to KK\varphi$ (por transitividad de las implicaciones 2, 3 y 8)

Teorema 3.2.4. ***KB*** *es correcto y completo respecto de* $\mathcal{KB}$.

La demostración no será incluída porque no aporta ideas diferentes de las usadas hasta ahora, pero puede consultarse en [27].

Este sistema verifica $B\varphi \to KB\varphi$, es decir: las creencias son conscientes, en el sentido de que el agente sabe que está adoptándolas. Un concepto interesante que no es capturado por este sistema es $B\varphi \to BK\varphi$, sin embargo agregarlo como axioma sin más, resulta en que $\vdash B\varphi \leftrightarrow K\varphi$. Por supuesto, queremos evitar que esto suceda. A este problema de permitir que valga $B\varphi \to BK\varphi$ sin que colapsen las modalidades se lo conoce como problema BCB [2].

Los siguientes resultados nos permiten ver cómo incorporar $B\varphi \to BK\varphi$ a un sistema, sin que resulte en el colapso sintáctico de K y B.

Proposición 3.2.5. *Sea* $\mathfrak{F}$ *un frame con dos relaciones* S *y* T*, que inducen los operadores modales universales* K *y* B*, respectivamente. Valen:*

1. $\Vdash_{\mathfrak{F}} B\varphi \to BK\varphi$ *sii* T *es transitiva sobre* (T, S).

2. $\Vdash_{\mathfrak{F}} B\varphi \to KB\varphi$ *sii* T *es transitiva sobre* (S, T)

Demostración. 1) $\Leftarrow$ sale aplicando definiciones. Veamos $\Rightarrow$: si T no es transitiva sobre (T, S), entonces existen x, y, z tales que xTy, ySz pero no vale xTz. Tomamos una variables proposicional p y definimos una interpretación $\mathtt{i}$, tal que $\mathtt{i}(p) = \{z\}^c$, si $\mathfrak{M} = \langle \mathfrak{F}, \mathtt{i} \rangle$ entonces $\nVdash_{\mathfrak{M}} B\varphi \to BK\varphi$.

2) De manera análoga al 1. □

Teorema 3.2.6. *Sean* S *y* T *dos relaciones sobre* W*. Si* S *es euclideana y* T *es serial, transitiva sobre* (T, S) *y está incluída en* S *entonces* $S = T$.

Demostración. Sean x, y tales que xSy, como T es serial existe un z tal que xTz y por lo tanto xSz. Dado que S es euclideana zSy, y por la transitividad de T sobre (T, S) se obtiene que xTy. □

Existen modelos con relaciones que no cumplen alguna de las 4 condiciones del Teorema anterior en los que vale $S \neq T$. Esto nos permite identificar con qué se corresponde semánticamente el colapso de K y B. Por las correspondencias entre propiedades de las relaciones y fórmulas válidas en un frame y la Proposición 3.2.5, para incorporar el axioma $B\varphi \to BK\varphi$ sin que colapsen B y K es necesario y suficiente descartar alguno de los axiomas 3, 7 u 8.

Descartar el axioma 8 ($K\varphi \to B\varphi$) da lugar a una noción más implícita de creencia y en tal caso el axioma 9 ($B\varphi \to KB\varphi$) no sería el apropiado si queremos representar una

[2] BCB proviene de "Believed counciousness of beliefs".

forma implícita de creencia, y habría que descartarlo a pesar de que semánticamente no sea necesario para la incorporación de $B\varphi \to BK\varphi$.

Si renunciamos al axioma 7 ($\neg B\bot$) sucede algo similar: si conservamos el axioma 9 obtenemos $B\bot \to KB\bot$ ¿y por qué motivo un agente epistémico conservaría una creencia falsa sabiéndolo?.

Podríamos considerar descartar el axioma 3 ($\neg K\varphi \to K\neg K\varphi$), pero es una mejor opción sustituírlo por algún axioma más débil porque es de interés conservar alguna propiedad relacionada con la introspección de K. Vimos en el ejemplo 3.2.3 que el axioma 3 (junto con el 2: $K\varphi \to \varphi$) implican la introspección positiva.

Definición 3.2.7. Llamamos $\mathbf{KB}^-$ al sistema que consiste en todos los axiomas de **KB** pero reemplazando 3 por 3': $K\varphi \to KK\varphi$ y agregando el axioma 8: $B\varphi \to BK\varphi$.

Corolario 3.2.8. $\boldsymbol{KB}^-$ *es correcto para los frames en los que* S *es transitiva y reflexiva (pero no necesariamente euclideana) y* $T \subseteq S$ *es serial y transitivo sobre* (T, S) *y sobre* (S, T).

Demostración. Como las reglas de inferencia preservan validez, si los axiomas son válidos en un frames entonces todas las fórmulas demostrables con el sistema $\mathbf{KB}^-$ serán válidos.

De las correspondencias entre fórmulas validas en frames y propiedades demostradas en la Proposición 2.3.4, como S es transitiva el axioma 3' es válido, y como también es reflexiva el axioma 2 también lo es. La serialidad de T garantiza la validez del axioma 7 y su transitividad sobre (S, T) y (T, S), nos dan la validez de los axiomas 9 y 8 (por la proposición 3.2.5). Pedimos que S no sea euclideana para que no colapsen conocimiento y creencia (porque por la Teorema 3.2.6 resultaría $T = S$), T es serial (para que valga el axioma 4) y transitivo (para que valga el axioma 7). □

Teorema 3.2.9. $\nvdash_{\boldsymbol{KB}^-} B\varphi \to K\varphi$

Demostración. Alcanza con encontrar un frame con las propiedades mencionadas en el Corolario 3.2.8 en el que no sea válida $B\varphi \to K\varphi$. Representamos uno en el siguiente diagrama donde las flechas violetas y punteadas son para S y las turquesas y rellenas para T:

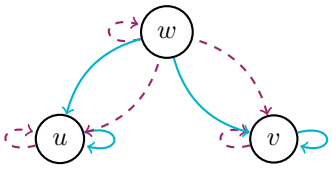


Figura 3.4: Modelo que demuestra que en $\mathbf{KB}^-$ no colapsan K y B

□

Definición 3.2.10. Sean $X, Y, Z \in \{B, K\}$ operadores. Las fórmulas de la forma:

- $X\varphi \to YZ\varphi$ se denominan *fórmulas de introspección positiva*

- $\neg X\varphi \to Y\neg Z\varphi$ se denominan *fórmulas de introspección negativa*
- $XY\varphi \to Z\varphi$ se denominan *fórmulas de extraspección positiva*
- $X\neg Y\varphi \to \neg Z\varphi$ se denominan *fórmulas de extraspección negativa*
- $X(Y\varphi \to \varphi)$ se denominan *fórmulas de confianza*

Al conjunto de fórmulas introspección o extraspección lo llamamos *fórmulas de inspección*, y al de la unión de todas las enumeradas lo llamamos *fórmulas IT*[3].

¿Pueden agregarse alguno de estos axiomas al sistema axiomático $\mathbf{KB}^-$? A este respecto viene el siguiente teorema.

Teorema 3.2.11. ***KB*** *es maximal en el sentido que agregando cualquier fórmula IT que no sea parte del sistema, resulta en que* $\vdash K\varphi \leftrightarrow B\varphi$.

Este resultado nos asegura que no hay mucho más para agregar al sistema axiomático, para hacerlo hay que renunciar a otros axiomas. Omitimos la demostración, que consiste en verificar todos los casos. Puede consultarse en [27].

3.3. Omnisciencia lógica

Quizás la mayor crítica a la lógica epistémica y doxástica se refiere a las paradojas de *omnisciencia lógica*. Ellas son inherentes al uso de semánticas de Kripke. La propiedad básica (válida en la clase de modelos de Kripke) $\Vdash K(\varphi \to \psi) \to (K\varphi \to K\psi)$ supone que un agente conoce todas las consecuencias de su propio conocimiento. Otra consecuencia de los modelos de Kripke para la lógica epistémica es que los agentes saben todas las fórmulas válidas en la lógica proposicional inducida por el mismo conjunto de variables proposicionales. Para el caso de capacidad de razonamiento limitada esto es un poco inverosímil. Por ejemplo: la mayoría de los estudiantes de lógica conocen los axiomas y reglas de inferencia de la lógica proposicional, y sin embargo no conocen todas las fórmulas derivables por ese sistema.

También se critica el axioma de introspección negativa para razonar acerca de proposiciones de las que el agente ni siquiera tiene consciencia: supongamos que hay una persona que no incorporó nuevas tecnologías a su vida desde 1960 en adelante, es discutible afirmar que esta persona sabe que no sabe que Telegram es open source.

Uno puede o bien aceptar esta visión del conocimiento vía operadores modales como una idealización o bien involucrar semánticas no normales [4] para evitar la validez de proposiciones no deseadas.

Vamos a mencionar brevemente una de estas lógicas, las cuales son útiles en el área de inteligencia artificial ya que son mucho más apropiadas para estudiar situaciones epistémicas y doxásticas con agentes que no son perfectamente racionales.

[3] Introspection-trust

[4] Aquellas en las no vale el axioma K.

LÓGICA DE CONSCIENCIA

La lógica de consciencia permite no sólo razonar acerca de las creencias sino acerca de la consciencia de existencia distintas propiedades creíbles. En esta lógica no hay un operador para conocimiento: hay operadores de creencia explícita y creencia implícita. Esto se debe a que el enfoque relacionado con una visión de los agentes sesgada no es compatible completamente con la noción de conocimiento definida hasta ahora. ¿Podemos decir que un agente sabe algo cuando no está considerando en sus razonamientos todas las variables posibles? Por este motivo, esta lógica trabaja con creencias de distintos niveles dejando de lado el conocimiento.

Definición 3.3.1. Una *estructura de consciencia* es una tupla $\mathfrak{M} = \langle W, A, R, \mathtt{i} \rangle$ donde, como en los modelos doxásticos clásicos, W es un conjunto de mundos posibles, $\mathtt{i}$ es una interpretación y R es una relación en W serial, transitiva y euclideana. El nuevo componente es una función A que asocia a cada mundo w un conjunto de variables proposicionales. Intuitivamente $A(w)$ representa el conjunto de proposiciones de las cuales el agente tiene consciencia de su existencia.

Otra restricción de un agente que no es perfectamente racional, es que podría no considerar siempre todas las variables proposicionales. Contemplando esto y la consciencia del agente, dado un conjunto de variables proposicionales Ψ de los que el agente está consciente, definimos relaciones de soporte $\Vdash_T^\Psi$ y $\Vdash_F^\Psi$, cuya motivación es capturar en un mismo estado distintas situaciones parciales en las que no se consideran todas las variables.

Definición 3.3.2. Dado una estructura de consciencia $\mathfrak{M}$ y un conjunto de variables proposicionales Ψ, definimos recursivamente las relaciones de soporte $\Vdash_T^\Psi$ y $\Vdash_F^\Psi$ y la noción de satisfacción $\Vdash$ del siguiente modo:

- $\mathfrak{M}, w \Vdash_T^\Psi \top$
- $\mathfrak{M}, w \nVdash_F^\Psi \top$
- $\mathfrak{M}, w \Vdash \top$

- $\mathfrak{M}, w \Vdash_T^\Psi p$ para $p \in \Phi$ sii $w \in \mathtt{i}(p)$ y $p \in \Psi$
- $\mathfrak{M}, w \Vdash_F^\Psi p$ para $p \in \Phi$ sii $w \notin \mathtt{i}(p)$ y $p \in \Psi$
- $\mathfrak{M}, w \Vdash p$ para $p \in \Phi$ sii $w \in \mathtt{i}(p)$

- $\mathfrak{M}, w \Vdash_T^\Psi \neg\varphi$ sii $\mathfrak{M}, w \Vdash_F^\Psi \varphi$
- $\mathfrak{M}, w \Vdash_F^\Psi \neg\varphi$ sii $\mathfrak{M}, w \Vdash_T^\Psi \varphi$
- $\mathfrak{M}, w \Vdash \neg\varphi$ sii $\mathfrak{M}, w \nVdash \varphi$

- $\mathfrak{M}, w \Vdash_T^\Psi \varphi_1 \wedge \varphi_2$ sii $\mathfrak{M}, w \Vdash_T^\Psi \varphi_1$ y $\mathfrak{M}, w \Vdash_T^\Psi \varphi_2$
- $\mathfrak{M}, w \Vdash_F^\Psi \varphi_1 \wedge \varphi_2$ sii $\mathfrak{M}, w \Vdash_F^\Psi \varphi_1$ o $\mathfrak{M}, w \Vdash_F^\Psi \varphi_2$
- $\mathfrak{M}, w \Vdash \varphi_1 \wedge \varphi_2$ sii $\mathfrak{M}, w \Vdash \varphi_1$ y $\mathfrak{M}, w \Vdash \varphi_2$

- $\mathfrak{M}, w \Vdash_T^\Psi B\varphi$ sii $\mathfrak{M}, v \Vdash_T^{\Psi \cap A(w)} \varphi$ para todo v tal que wRv
- $\mathfrak{M}, w \Vdash_F^\Psi B\varphi$ sii $\mathfrak{M}, v \Vdash_F^{\Psi \cap A(w)}$ para algún v tal que wRv
- $\mathfrak{M}, w \Vdash B\varphi$ sii $\mathfrak{M}, w \Vdash_T^\Phi B\varphi$

- $\mathfrak{M}, w \Vdash_T^\Psi L\varphi$ sii $\mathfrak{M}, v \Vdash_T^\Psi \varphi$ para todo v tal que wRv
- $\mathfrak{M}, w \Vdash_F^\Psi L\varphi$ sii $\mathfrak{M}, v \Vdash_F^\Psi \varphi$ para algún v tal que wRv
- $\mathfrak{M}, w \Vdash L\varphi$ sii $\mathfrak{M}, v \Vdash \varphi$ para todo v tal que wRv

La interpretación intuitiva de $\mathfrak{M}, w \Vdash_T^\Psi \varphi$ es que, considerando las variables Ψ, el mundo w soporta la verdad de φ. Análogamente $\mathfrak{M}, w \Vdash_F^\Psi \varphi$ vamos a entenderlo como que en la misma situación soporta la falsedad de φ. En cambio $w \Vdash \varphi$ es interpretado, al igual que antes, como que en el mundo w vale φ considerando que el agente puede contemplar todas las variables proposicionales al mismo tiempo. Además de las nuevas relaciones de soporte, también contamos con un nuevo operador y nuevas nociones de creencias: mientras B descarta las variables de las que no tiene consciencia el agente, L no. Por este motivo, leemos $B\varphi$ como "el agente cree explícitamente φ" y $L\varphi$ como "el agente cree implícitamente φ".

Como usualmente, decimos que φ es *válida* en $\mathfrak{M}$, si $w \Vdash_{\mathfrak{M}} \varphi$ para todos los $w \in W$, y que es *satisfacible* en $\mathfrak{M}$ si $w \Vdash_{\mathfrak{M}} \varphi$ para algún $w \in W$.

Proposición 3.3.3. *La lógica de consciencia cumple las siguientes propiedades:*

1. $\Vdash$ *es completa, es decir, para cada* $\mathfrak{M}, w$ *y* φ *vale* $\mathfrak{M}, w \Vdash \varphi$ *o* $\mathfrak{M}, w \Vdash \neg\varphi$
2. a) *Si* $\Psi \subseteq \Psi'$ *y* $\mathfrak{M}, w \Vdash_T^\Psi \varphi$ *entonces* $\mathfrak{M}, w \Vdash_T^{\Psi'} \varphi$

b) *Si* $\Psi \subseteq \Psi'$ *y* $\mathfrak{M}, w \Vdash_F^{\Psi} \varphi$ *entonces* $\mathfrak{M}, w \Vdash_F^{\Psi'} \varphi$

3. a) *Para cada* $\Psi \subseteq \Phi$ *si* $\mathfrak{M}, w \Vdash_T^{\Psi} \varphi$ *entonces* $\mathfrak{M}, w \Vdash \varphi$

   b) *Para cada* $\Psi \subseteq \Phi$ *si* $\mathfrak{M}, w \Vdash_F^{\Psi} \varphi$ *entonces* $\mathfrak{M}, w \Vdash \neg\varphi$

4. $\Vdash B\varphi \to L\varphi$

Demostración. 1 es por definición, 2 y 3 se demuestra por inducción en la estructura de φ y 4 se desprende de $3a$.

□

En la proposición de recién se expresa una relación clara entre las dos nociones de creencia: creer algo explícitamente es más fuerte que creerlo implícitamente.

Al igual que en los modelos epistémicos tradicionales, un agente cree todas las fórmulas válidas y consecuencias de sus creencias: pero ahora lo hace implícitamente. Sin embargo, un agente no necesariamente cree explícitamente todas las fórmulas válidas: en particular $\neg B(p \vee \neg p)$ es satisfacible. Además las creencias explícitas no son cerradas por implicación, por ejemplo: $Bp \wedge \neg B(p \wedge (q \vee \neg q))$ es satisfacible.

El propósito de esta sección no es hacer hincapié en los resultados técnicos de esta lógica, sino mencionarla como alternativa a la lógica epistémica tradicional. Por este motivo, no vamos a ir más allá de estas definiciones, pero en caso de interés el lector interesado puede encontrar resultados interesantes en [14], entre ellos los de completitud y complejidad.

Capítulo 4

Lógica epistémica dinámica

> El conocimiento de algo, dado que todas las cosas tienen causa, no es adquirido o completado salvo a través de sus causas.
>
> Avicena.

En este capítulo vamos a abordar el aspecto dinámico de la lógica epistémica: cómo cambian los conocimientos luego de una determinada acción, de un anuncio público, de un anuncio privado, etcétera. Estos conceptos que se encuentran con facilidad en diversos contextos donde múltiples agentes interactúan. Ejemplos clásicos de lógica epistémica dinámica se encuentran en distintas disciplinas: computación, juegos, economía, política. Estos sistemas multiagentes son dinámicos: los agentes cambian el sistema ejecutando acciones como movimientos en un juego, comunicación entre agentes, compra y venta de activos financieros, y demás.

Por otro lado, estos también son sistemas de información: los agentes adquieren, guardan, procesan e intercambian información acerca de ellos mismos y del contexto. Para expresar cambios en la información de los agentes, la lógica epistémica toma una idea de PDL[1]: usar modalidades dinámicas. Vamos a incorporar operadores $[\alpha]\varphi$ cuya lectura intuitiva es "si la acción α es ejecutada en el estado actual, la fórmula φ será verdadera después de la ejecución".

En PDL, las modalidades dinámicas son tratadas como cualquier otra modalidad de un modelo de Kripke: los modelos $\mathfrak{M}$ vienen munidos de una relación binaria R_α para cada acción α, y la semántica de $[\alpha]$ está dada por

$$w \Vdash_{\mathfrak{M}} [\alpha]\varphi \text{ sii para todo } v \text{ con } wR_\alpha v \text{ vale } v \Vdash_{\mathfrak{M}} \varphi.$$

La diferencia con la lógica epistémica dinámica (en adelante: DEL) es que las relaciones R_α tienen que ser interpretadas como relaciones entre *distintos* modelos. La razón es

[1] Lógica proposicional dinámica

que un modelo epistémico representa la situación epistémica en un momento dado y una acción cambia esta situación epistémica, es decir, el modelo. Por lo tanto un modelo puede no contener estados correspondientes a todos los outputs de todas las posibles acciones ejecutables en él. Esto es lo que se llama un *sistema de información abierta*: un sistema de información en el que las acciones futuras no están predeterminadas.

Se podría, desde una perspectiva matemática, convertir un sistema abierto en uno cerrado, agregando al modelo desde un principio todos los estados necesarios para representar todos los outputs posibles de todas las posibles cadenas de acciones. Pero hay razones prácticas por las cuales esta no es una buena idea: en un verdadero sistema abierto, el número posible de acciones es no sólo infinito sino que está más allá de cualquier cardinalidad (es una clase propia, en el sentido de teoría de conjuntos). Aún así, ese fue el enfoque que en un principio adoptó Gebrandy para modelar las lógicas epistémicas dinámicas, pero autores posteriores advirtieron la inconveniencia de esta formulación y adoptaron una visión más simple en la que se comienza con modelos base, y se agregan o quitan nuevos estados sólo cuando es necesario. El precio a pagar es que las transiciones provocadas por las acciones van de un modelo a otro diferente en lugar de quedar dentro de un mismo modelo.

4.1. Actualizaciones epistémicas

Definición 4.1.1. Una *acción o actualización epistémica* es una transformación de modelos α, que consiste en:

- Una función que manda un modelo $\mathfrak{M}$ a otro modelo $\mathfrak{M}^\alpha$
- Una relación binaria de transición de estados $R^\alpha \subseteq \mathfrak{M} \times \mathfrak{M}^\alpha$

La semántica de las modalidades dinámicas entonces va a estar dada por:

$$w \Vdash_{\mathfrak{M}} [\alpha]\varphi \text{ sii para todo } v \in \mathfrak{M}^\alpha \text{ con } wR^\alpha v \text{ vale que } v \Vdash_{\mathfrak{M}^\alpha} \varphi$$

Esta es una definición muy general que nos permite empezar a modelar algunos casos particulares de actualizaciones epistémicas. Vamos a comenzar por la más tradicional: el anuncio público.

4.1.1. Lógica de anuncios públicos (PAL)

El anuncio público captura la situación en la que una información que viene de una fuente certera (por ejemplo: la realidad) es anunciada públicamente a todos los agentes.

La lógica de anuncios públicos, también llamada PAL (PAL: public announcement logic), se obtiene agregando a una lógica multimodal operadores para los anuncios públicos. Fue formulada y axiomatizada en [22], y posteriormente dio lugar a muchas otras lógicas.

Definición 4.1.2. Definimos el *lenguaje de anuncios públicos multiagente*, al que denotaremos con $\mathcal{L}_!(N, \Phi)$, inductivamente del siguiente modo:

$$\varphi ::= p \mid \bot \mid \neg\varphi \mid \psi \vee \varphi \mid \Box_i\varphi \mid [!\varphi]\psi$$

donde p es alguna variable proposicional (i.e. $p \in \Phi$) e i es un agente ($i \in N$).

Estas fórmulas son interepretadas en modelos de Kripke multiagente. Si en la sintaxis de recién usamos K en lugar de $\Box$ e interpretamos las fórmulas en modelos epistémicos obtenemos la *lógica epistémica con anuncios públicos*. En cambio, si usamos B en lugar $\Box$ e interpretamos las fórmulas en modelos doxásticos obtenemos la *lógica doxástica con anuncios públicos*.

Definición 4.1.3. Un *anuncio público* $!\varphi$ es un actualización epistémica que mapea un modelo $\mathfrak{M} = \langle W, (R_i)_{i\in N}, \mathtt{i}\rangle$ a un modelo $\mathfrak{M}^{!\varphi} = \langle W^\varphi, (R_i^\varphi)_{i\in N}, \mathtt{i}^\varphi\rangle$ dado por:

- $W^\varphi := \mathtt{i}(\varphi)$ (i.e. los elementos de W en los que vale φ para la interpretación $\mathtt{i}$)
- $wR_i^\varphi v$ sii wR_iv para $v, w \in W^\varphi$
- $\mathtt{i}^\varphi(p) := \mathtt{i}(p) \cap W^\varphi$

Es decir, transforma el modelo en uno igual pero en el que sólo quedan los estados en los que vale φ, y las relaciones son las de antes restringidas al nuevo universo W^φ. Esta restricción del modelo se debe a que, como la fuente de información es absolutamente certera, los mundos en los que no vale φ ya no son mundos posibles luego del anuncio.

Dado un modelo epistémico $\mathfrak{M}$, la semántica de los conectivos y operadores epistémicos tienen la misma interpretación que en el Capítulo 3 y los nuevos operadores de anuncios públicos tienen la siguiente semántica:

$$w \Vdash_{\mathfrak{M}} [!\varphi]\psi \text{ sii } w \Vdash_{\mathfrak{M}} \varphi \text{ implica } w \Vdash_{\mathfrak{M}^{!\varphi}} \psi.$$

Con esta semántica $w \Vdash_{\mathfrak{M}} [!\varphi]\psi$ puede ser leído como "si se puede anunciar $!\varphi$ en el mundo w, después del anuncio valdrá ψ".

Y definimos $\langle!\varphi\rangle$, el dual de $[!\varphi]$, con la siguiente semántica:

$$w \Vdash_{\mathfrak{M}} \langle!\varphi\rangle\psi \text{ sii } w \Vdash_{\mathfrak{M}} \varphi \text{ y } w \Vdash_{\mathfrak{M}^{!\varphi}} \psi.$$

La interpretación intuitiva de $w \Vdash_{\mathfrak{M}} \langle!\varphi\rangle\psi$ es "Se puede anunciar $!\varphi$ en el mundo w, y si sucede después del anuncio valdrá ψ".

Proposición 4.1.4. *Los anuncios son parciales y funcionales, i.e.*

$$\nVdash \langle!\varphi\rangle\top \;\; y \;\; \Vdash \langle!\varphi\rangle\psi \to [!\varphi]\psi.$$

Demostración. La parcialidad se debe a que si $w \Vdash \neg\varphi$, entonces $w \nVdash \langle !\varphi\rangle \top$.

Para ver la funcionalidad, sean $\mathfrak{M}$ y w un mundo de $\mathfrak{M}$. Por definición $w \Vdash_{\mathfrak{M}} \langle\varphi\rangle\psi$ sii $w \Vdash_{\mathfrak{M}} \varphi$ y $w \Vdash_{\mathfrak{M}^{!\varphi}} \psi$, pero de esto último, razonando proposicionalmente, se obtiene que $w \Vdash_{\mathfrak{M}^{!\varphi}}$ implica $w \Vdash_{\mathfrak{M}^{!\varphi}} \psi$.

□

Para comenzar a avanzar hacia una axiomatización correcta y completa primero vamos a ver algunas fórmulas válidas que nos permitirán escribir fórmulas con anuncios como fórmulas sin anuncios, y que posteriormente vamos a utilizar en nuestro sistema de axiomas.

Proposición 4.1.5. *La siguiente fórmula es válida*

$$[!\varphi]\neg\psi \leftrightarrow (\varphi \rightarrow \neg[!\varphi]\psi)$$

Demostración. Sea $\mathfrak{M}$ un modelo y w un mundo de $\mathfrak{M}$.

Supongamos primero que $w \Vdash_{\mathfrak{M}} [!\varphi]\neg\psi$. Esto significa que $w \Vdash_{\mathfrak{M}} \varphi$ implica $w \Vdash_{\mathfrak{M}^{!\varphi}} \neg\psi$, o dicho de otro modo: $w \Vdash_{\mathfrak{M}} \varphi$ implica $w \nVdash_{\mathfrak{M}^{!\varphi}} \psi$. Pero valiéndonos de la lógica proposicional, tenemos que $w \Vdash_{\mathfrak{M}} \varphi$ implica que no vale ($w \Vdash_{\mathfrak{M}^{!\varphi}}$ o $w \nVdash_{\mathfrak{M}} \varphi$), y esto último es lo mismo que decir que no vale ($w \Vdash_{\mathfrak{M}} \varphi$ implica $w \Vdash_{\mathfrak{M}^{!\varphi}} \psi$). Por lo tanto $w \Vdash_{\mathfrak{M}} \varphi \rightarrow (\varphi \rightarrow \neg[!\varphi]\psi)$.

Ahora supongamos que vale $w \Vdash_{\mathfrak{M}} \varphi \rightarrow \neg[!\varphi]\psi$. Es decir que si $w \Vdash_{\mathfrak{M}} \varphi$ entonces vale $w \nVdash_{\mathfrak{M}} [!\varphi]\psi$. Como esto último es una implicación, tenemos que $w \Vdash_{\mathfrak{M}} \varphi$ implica ($w \Vdash_{\mathfrak{M}} \varphi$ y $w \nVdash_{\mathfrak{M}^{!\varphi}} \psi$), i.e. $w \Vdash_{\mathfrak{M}} \varphi$ implica $[!\varphi]\neg\psi$. Si $w \Vdash_{\mathfrak{M}} \varphi$ es falso, vale $w \Vdash_{\mathfrak{M}} [!\varphi]\neg\psi$ (como queríamos ver). Y si $w \Vdash_{\mathfrak{M}} \varphi$ es verdadero, por modus ponens vale $w \Vdash_{\mathfrak{M}} [!\varphi]\neg\psi$. Por lo tanto siempre vale $w \Vdash_{\mathfrak{M}} [!\varphi]\neg\psi$.

□

Proposición 4.1.6. $[!(\varphi \wedge [!\varphi]\psi)]\chi$ *es equivalente a* $[!\varphi][!\psi]\chi$.

Demostración. Sean $\mathfrak{M}$ un modelo y w un mundo de $\mathfrak{M}$. Resulta que:

$$w \in \mathfrak{M}^{!(\varphi\wedge[!\varphi]\psi)}$$

$$sii$$

$$w \Vdash_{\mathfrak{M}} \varphi \wedge [!\varphi]\psi$$

$$sii$$

$$w \Vdash_{\mathfrak{M}} \varphi \; y \; (w \Vdash_{\mathfrak{M}} \varphi \; implica \; w \Vdash_{\mathfrak{M}^{!\varphi}} \psi)$$

$$sii$$

$$w \in \mathfrak{M}^{!\varphi} \; y \; w \Vdash_{\mathfrak{M}^{!\varphi}} \psi$$

$$sii$$

$$w \in (\mathfrak{M}^{!\varphi})^{!\psi}.$$

Y de esto se desprende el resultado. □

Proposición 4.1.7. $[!\varphi]K_i\psi$ *es equivalente a* $\varphi \to K_i[!\varphi]\psi$.

Demostración. Sean $\mathfrak{M}$ un modelo y w un mundo en $\mathfrak{M}$:

$$w \Vdash_{\mathfrak{M}} \varphi \to K_i[!\varphi]\psi$$

$$sii$$

$$w \Vdash_{\mathfrak{M}} \varphi \text{ implica } w \Vdash_{\mathfrak{M}} K_i[!\varphi]\psi$$

$$sii$$

$$w \Vdash_{\mathfrak{M}} \varphi \text{ implica (para todo } v \text{ tal que } wR_iv \text{ vale } v \Vdash_{\mathfrak{M}} [!\varphi]\psi)$$

$$sii$$

$$w \Vdash_{\mathfrak{M}} \varphi \text{ implica (para todo } v \text{ tal que } wR_iv \text{ vale que } v \Vdash_{\mathfrak{M}} \varphi \text{ implica } v \Vdash_{\mathfrak{M}^{!\varphi}} \psi)$$

$$sii$$

$$w \Vdash_{\mathfrak{M}} \varphi \text{ implica(paratodo } v \in \mathfrak{M}^{!\varphi} \text{ vale que } wR_iv \text{ implica } v \Vdash_{\mathfrak{M}^{!\varphi}} \psi)$$

$$sii$$

$$w \Vdash_{\mathfrak{M}} \varphi \text{ implica } w \Vdash_{\mathfrak{M}^{!\varphi}} K_i\psi$$

$$sii$$

$$w \Vdash_{\mathfrak{M}} [!\varphi]K_i\psi$$

□

El resto de las equivalencias no vamos a demostrarlas porque se prueban del mismo modo, pero dejamos todas las fórmulas recopiladas en la siguiente proposición:

Proposición 4.1.8. *Las siguientes fórmulas son válidas:*

1. $[!\varphi]p \leftrightarrow (\varphi \to p)$ *si* $p \in \Phi$.
2. $[!\varphi](\psi \wedge \chi) \leftrightarrow ([!\varphi]\psi \wedge [!\varphi]\chi)$
3. $[!\varphi](\psi \to \chi) \leftrightarrow ([!\varphi]\psi \to [!\varphi]\chi)$
4. $[!\varphi]\neg\psi \leftrightarrow (\varphi \to \neg[!\varphi]\psi)$
5. $[!\varphi]K_i\psi \leftrightarrow (\varphi \to K_i[!\varphi]\psi)$
6. $[!\varphi][!\psi]\chi \leftrightarrow [!(\varphi \wedge [!\varphi]\psi)]\chi$

La validez de estas fórmulas, nos permite reescribir cualquier fórmula hasta eliminar los anuncios por completo y nos permitirá obtener una axiomatización completa y correcta.

Definición 4.1.9. Llamamos **PA** al sistema de axiomas que consiste en:

- Una axiomatización de la lógica proposicional
- **(K para K_i)** $K_i(\varphi \to \psi) \to (K_i\varphi \to K_i\psi)$
- **(Veracidad para K_i)** $K_i\varphi \to \varphi$
- **(Introspección positiva para K_i)** $K_i\varphi \to K_iK_i\varphi$
- **(Introspección negativa para K_i)** $\neg K_i\varphi \to K_i\neg K_i\varphi$
- **(Permanencia atómica)** $[!\varphi]p \leftrightarrow p$ para $p \in \Phi$
- **(Anuncio-negación)** $[!\varphi]\neg\psi \leftrightarrow (\varphi \to \neg[!\varphi]\psi)$
- **(Anuncio-conjunción)** $[!\varphi](\psi \wedge \chi) \leftrightarrow ([!\varphi]\psi \wedge [!\varphi]\chi)$
- **(Anuncio-conocimiento)** $[!\varphi]K_i\psi \leftrightarrow (\varphi \to K_i[!\varphi]\psi)$
- **(Anuncio-anuncio)** $[!\varphi][!\psi]\chi \leftrightarrow [!(\varphi \wedge [!\varphi]\psi)]\chi$

Junto con las reglas de inferencia modus ponens y necesitación para K_i

Corolario 4.1.10. *El sistema **PA** es correcto respecto de la clase de modelos epistémicos.*

Demostración. Se desprende de la Proposición 4.1.8 y de que la semántica de los operadores K_i es la misma que en la lógica epistémica clásica. □

Para ver la completitud se usa fuertemente el hecho de que podemos expresar lo que sucede después de los anuncios en términos de lo que sucedía antes de ejecutarlos, hecho que se ve reflejado en la Proposición 4.1.8. Vamos a formalizarlo a través de una función de traducción:

Definición 4.1.11. Definimos la *función de traducción de PAL a lógica epistémica* $t : Form(\mathcal{L}_!) \to \mathcal{L}$ recursivamente del siguiente modo:

$$\begin{aligned} t(p) &= p \\ t(\neg\varphi) &= \neg t(\varphi) \\ t(\varphi \wedge \psi) &= t(\varphi) \wedge t(\psi) \\ t(K_i\varphi) &= K_i t(\varphi) \\ t([!\varphi]p) &= t(\varphi \to p) \\ t([!\varphi]\neg\psi) &= t(\varphi \to \neg[!\varphi]\psi) \\ t([!\varphi]\psi \wedge \chi) &= t([!\varphi]\psi \wedge [!\varphi]\chi) \\ t([!\varphi]K_i\psi) &= t(\varphi \to K_i[!\varphi]\psi) \\ t([!\psi][!\psi]\chi) &= [!(\varphi \wedge [!\varphi]\psi)]\chi \end{aligned}$$

La correctitud del sistema implica que la traducción preserva validez. Pero para ver que la equivalencia de cada fórmula con su traducción es demostrable en **PA** vamos a precisar definir una herramienta útil: la complejidad de una fórmula, que definimos más abajo.

Hasta ahora hemos demostrado varios resultados por inducción en la estructura de la fórmula pero en este caso necesitamos ir un poco más lejos porque, por ejemplo, $\varphi \to \neg[!\varphi]\psi$ no es una subfórmula de $[!\varphi]\neg\psi$ pero sin embargo nos gustaría poder aplicar hipótesis inductiva. Esa es la motivación detrás de la definición de la función de complejidad.

Definición 4.1.12. La *complejidad de* $\mathcal{L}_!$ es una función $c : \mathcal{L}_! \to \mathbb{N}$ definida recursivamente del siguiente modo:

$$\begin{aligned} c(p) &= 1 \\ c(\neg\varphi) &= 1 + c(\varphi) \\ c(\varphi \wedge \psi) &= 1 + \text{máx}(c(\varphi), c(\psi)) \\ c(K_i\varphi) &= 1 + c(\varphi) \\ c([!\varphi]\psi) &= (4 + c(\varphi)) \cdot c(\psi) \end{aligned}$$

Esta función preserva el orden de la longitud de las fórmulas, pero además tiene algunas propiedades deseables que carece la longitud. En la definición el número 4 puede resultar arbitrario, pero en realidad se eligió por ser el menor que hace que se cumplan estas propiedades:

Lema 4.1.13. *Para todo* φ, ψ *y* χ *valen:*

1. $c(\varphi) \geq c(\psi)$ *si* $\psi \in Sub(\varphi)$
2. $c([!\varphi]p) > c(\varphi \to p)$
3. $c([!\varphi]\neg\psi) > c(\varphi \to \neg[!\varphi]\psi)$
4. $c([!\varphi](\psi \wedge \chi)) > c([!\varphi]\psi \wedge [!\varphi]\chi)$
5. $c([!\varphi]K_i\psi) > c(\varphi \to K_i[!\varphi]\psi)$
6. $c([!\varphi][!\psi]\chi) > c([!(\varphi \wedge [!\varphi]\psi)]\chi)$

Ahora estamos en condiciones de demostrar que sintácticamente una fórmula y su traducción son equivalentes.

Proposición 4.1.14. *Para toda* $\varphi \in \mathcal{L}_!$ *vale que*

$$\vdash \varphi \leftrightarrow t(\varphi)$$

Demostración. Vamos a verlo por inducción global en $c(\varphi)$:
`Caso` $c(\varphi) = 1$: esto sólo puede suceder cuando $\varphi = p \in \Phi$ y en este caso es trivial ya que $\vdash p \leftrightarrow p$.

`Caso inductivo`: Supongamos que ya quedó probado para todas las fórmulas con complejidad menor a $c(\varphi)$. Los casos para negación, conjunción y K_i salen directo de la hipótesis inductiva y el Lema 4.1.13 ítem 1.

Veamos los casos que involucran al operador de anuncio público:

`Caso` $[!\varphi]p$: Se desprende directamente de la hipótesis inductiva aplicada al axioma de permanencia atómica (se puede aplicar por el ítem 2 del Lema 4.1.13).

`Caso` $[!\varphi]\neg\psi$: Se desprende de la hipótesis inductiva aplicada al axioma de anuncio-negación (se puede aplicar por el ítem 3 del Lema 4.1.13).

`Caso` $[!\varphi](\psi \wedge \chi)$: Se desprende de la hipótesis inductiva aplicada al axioma anuncio-conjunción (se puede aplicar por el ítem 4 del Lema 4.1.13).

`Caso` $[!\varphi]K_i\psi$: Se desprende de la hipótesis inductiva aplicada al axioma anuncio-conocimiento (se puede aplicar por el ítem 5 del Lema 4.1.13).

`Caso` $[!\varphi][!\psi]\chi$: Se desprende de la hipótesis inductiva aplicada al axioma anuncio-anuncio (se puede aplicar por el ítem 6 del Lema 4.1.13). □

Lema 4.1.15. *La imagen de la función de traducción es exactamente* $\mathcal{L}_K$.

Demostración. Por inducción en $c(\varphi)$. □

Corolario 4.1.16. *PAL tiene la misma expresividad que la lógica epistémica para n agentes sin operador de conocimiento común.*

La completitud se desprende de la Proposición 4.1.14 junto con el Teorema 3.1.22

Teorema 4.1.17. *Sea* `M` *la clase de modelos epistémicos para n agentes aislados. Para toda* $\varphi \in \mathcal{L}_!$ *vale*

$$\Vdash_{\mathtt{M}} \varphi \text{ sii } \vdash_{\mathbf{PA}} \varphi$$

Demostración. Supongamos $\Vdash_{\mathtt{M}} \varphi$. Por la correctitud de **PA** y que $\vdash_{\mathbf{PA}} \varphi \leftrightarrow t(\varphi)$, tenemos que $\Vdash_{\mathtt{M}} t(\varphi)$. La fórmula $t(\varphi)$ no contiene operadores de anuncio público, es decir: $t(\varphi) \in \mathcal{L}_K$.

Por lo tanto, por el Teorema 3.1.22, tenemos que $\vdash_{\mathbf{S5}_n} t(\varphi)$. Como todos los axiomas y reglas de inferencia de $\mathbf{S5}_n$ están en **PA**, se desprende que $\vdash_{\mathbf{PA}} t(\varphi)$. Por la Proposición 4.1.14 y usando modus ponens tenemos que $\vdash_{\mathbf{PA}} \varphi$. □

4.1.2. Lógica de anuncios públicos y conocimiento común (PAC)

La esencia del anuncio público conlleva naturalmente al conocimiento común: cuando se hace un anuncio público no sólo todos los presentes incorporan la información sino que además todos saben que todos la incorporan, y todos saben que todos saben que todos la incorporan, etcétera. Por lo tanto tiene sentido estudiar una lógica con operadores de anuncio público y conocimiento común, ya que son nociones intrínsecamente relacionadas. En [4] se estableció una axiomatización, que captura sintácticamente un concepto semántico estaba presente desde los orígenes de los anuncios públicos en [22]. Más detalles pueden encontrarse en [26] y [17].

Definición 4.1.18. Definimos el *lenguaje de anuncios públicos con conocimiento común* (que será notado $\mathcal{L}_{C!}$) a partir del lenguaje $\mathcal{L}_!$ agregándole los operadores E y C. Esta lógica es conocida como PAC (public announcement - common).

La semántica que vamos a darles a E_k y C_k es la misma que le dimos en la lógica epistémica para n operadores: son los operadores modales de las relaciones R_E y R_C, donde $R_E = R_1 \cup \ldots \cup R_n$ y $R_C = (R_E)^*$ (la clausura transitiva y reflexiva).

El resto de los operadores tienen la misma semántica que en PAL.

Un primer intento por acercarnos a un resultado de completitud, podría ser considerar que C puede traducirse a lógica epistémica multiagente del mismo modo que K_i. Es decir: incorporar $[!\varphi]C\psi \leftrightarrow (\varphi \to C[!\varphi]\psi)$ como esquema de axiomas. Pero lamentablemente esta fórmula no es válida como se ve en el siguiente ejemplo.

Ejemplo 4.1.19. Tenemos dos agentes epistémicos y dos proposiciones p y q, las cuales son ambas verdaderas en el mundo real. El agente 1 considera indistinguibles en mundo en el que vale p y q y en el que vale sólo q, mientras que el agente 2 no puede distinguir el mundo en el que vale sólo q del mundo en el que vale sólo p.

Vemos un diagrama de esta situación en la Figura 4.1a, donde los nodos representan los mundos, en su interior se encuentran las variables proposicionales que valen en cada uno de ellos, las flechas etiquetadas representan las relaciones R_1 y R_2. Cuando haya un nodo coloreado, representa el mundo distinguido como mundo real.

(a) Antes de $!p$ (b) Después de $!p$

Figura 4.1: Modelo que no valida $[!p]Cq \leftrightarrow (p \to C[!p]q)$

Llamemos a este modelo $\mathfrak{M}$, PQ al mundo en el que valen p y q, P al mundo en el que sólo vale p y Q al mundo en el que sólo vale q. En este modelo vale que $PQ \Vdash [!p]Cq$, porque en $PQ \Vdash_{\mathfrak{M}^{!p}} Cq$. Sin embargo $PQ \nVdash_{\mathfrak{M}} p \to C[!p]q$, porque $PQ \Vdash_{\mathfrak{M}} p$ pero $PQ \nVdash_{\mathfrak{M}} C[!p]q$. Esto último se debe a que PQR_cP y $P \nVdash_{\mathfrak{M}} [!p]q$.

Sin embargo, hay una manera de relacionar anuncio público con conocimiento común:

Proposición 4.1.20. *Si $\chi \to [!\varphi]\psi$ y $(\chi \wedge \varphi) \to E\chi$ son válidas, entonces $\chi \to [!\varphi]C\psi$ es válida.*

Demostración. Sean $\mathfrak{M}$ y w tales que $w \Vdash_{\mathfrak{M}} \chi$, queremos ver que vale $w \Vdash_{\mathfrak{M}} [!\varphi]C\psi$. Para ello supongamos que $w \Vdash_{\mathfrak{M}} \varphi$ y sea v en $\mathfrak{M}^{!\varphi}$ tal que wR_cv, i.e. existe un camino $w = w_0R_{i_1}w_1R_{i_2}\ldots R_{i_k}w_k = v$ donde todos los w_i son elementos de $\mathfrak{M}^{!\varphi}$, por lo tanto

$v \Vdash_{\mathfrak{M}} \varphi$. Queremos ver que $v \Vdash_{\mathfrak{M}^{!\varphi}} \psi$, vamos a verlo por inducción en k, la longitud del camino:

Si $k = 0$, significa que $w = v$ y efectivamente tenemos que $w \Vdash_{\mathfrak{M}^{!\varphi}} \psi$ ya que $\chi \to [!\varphi]\psi$ y que partimos de la suposición $w \Vdash_{\mathfrak{M}} \chi$.

Ahora supongamos que vale hasta k, veamos que vale para $k+1$: entonces tenemos un i_0 y un u en $\mathfrak{M}^{!\varphi}$ (i.e. $u \Vdash_{\mathfrak{M}} \varphi$) tal que $wR_{i_0}u$ y uR_cv. Como $w \Vdash_{\mathfrak{M}} \chi$ y $w \Vdash_{\mathfrak{M}} \varphi$, junto con la validez de $(\chi \wedge \varphi) \to E\chi$ resulta en $w \Vdash_{\mathfrak{M}} E\chi$. En particular como $wR_{i_0}u$ se tiene que $u \Vdash_{\mathfrak{M}} \chi$, y además también valía que $u \Vdash_{\mathfrak{M}} \varphi$. Aplicando hipótesis inductiva en el camino uR_cv (de longitud k) tenemos que $v \Vdash_{\mathfrak{M}^{!\varphi}} \psi$.

□

Esta proposición nos permite capturar una de las principales características de la interacción entre los anuncios públicos y el conocimiento común:

Corolario 4.1.21. *$[!\varphi]\psi$ es válida sii $[!\varphi]C\psi$ es válida.*

Demostración. La implicación $\Leftarrow$ se debe a la veracidad del operador C, y $\Rightarrow$ se desprende de la Proposición 4.1.20 tomando $\chi = \top$. □

Definición 4.1.22. Definimos el sistema de axiomas **PAC** dado por extender al sistema **PA** con los siguientes axiomas:

- **(K para C)** $C(\varphi \to \psi) \to (C\varphi \to C\psi)$
- **(Veracidad para C)** $C\varphi \to \varphi$
- **(Común a todos)** $C\varphi \to EC\varphi$
- **(Escalar)** $C(\varphi \to E\varphi) \to (\varphi \to C\varphi)$

Y agregando como reglas de inferencia necesitación para C y la regla de inferencia de anuncio público y conocimiento común: de $\chi \to [!\varphi]\psi$ y $\chi \wedge \varphi \to E\chi$ inferimos $\chi \to [!\varphi]C\psi$.

Proposición 4.1.23. ***PAC*** *es correcto respecto de la clase de modelos epistémicos multiagente.*

En adelante, todas las definiciones y resultados estarán orientadas a un teorema de completitud para **PAC** respecto de los modelos epistémicos multiagentes. Para obtener esto, combinaremos ideas de la demostración de completitud de **S5EC** y de demostración de completitud de **PA**.

Definición 4.1.24. Sea $\varphi \in \mathcal{L}_{C!}$, decimos que $adec(\varphi)$ es el *conjunto adecuado para φ en $\mathcal{L}_{C!}$* si es el menor conjunto (respecto de la inclusión) que cumple:

- $\varphi \in adec(\varphi)$

- Si $\psi \in adec(\varphi)$ entonces $Sub(\psi) \subseteq adec(\varphi)$
- Si $\psi \in adec(\varphi)$ y ψ no es una negación, entonces $\neg\psi \in adec(\varphi)$
- Si $C\psi \in adec(\varphi)$ $\{K_i C\psi\} \subseteq adec(\varphi)$
- Si $[!\psi]p \in adec(\varphi)$ entonces $\psi \to p \in adec(\varphi)$
- Si $[!\psi]\neg\chi \in adec(\varphi)$ entonces $\psi \to \neg[!\psi]\chi \in adec(\varphi)$
- Si $[!\psi](\chi \wedge \delta) \in adec(\varphi)$ entonces $[!\psi]\chi \wedge [!\psi]\delta \in adec(\varphi)$
- Si $[!\psi]K_i\chi \in adec(\varphi)$ entonces $\psi \to K_i[!\psi]\chi \in adec(\varphi)$
- Si $[!\psi]C\chi \in adec(\varphi)$ entonces $[!\psi]\chi \in adec(\varphi)$ y $\{K_i[!\psi]C\chi\} \subseteq adec(\varphi)$
- Si $[!\psi][!\chi]\delta \in adec(\varphi)$ entonces $[!(\psi \wedge [!\psi]\chi)]\delta \in adec(\varphi)$

Lema 4.1.25. *Dada una $\varphi \in \mathcal{L}_{C!}$ se tiene que su conjunto adecuado $adec(\varphi)$ es finito.*

Demostración. Por inducción global en la complejidad de φ. □

Definición 4.1.26. Sea φ una fórmula y $\Theta \subseteq \mathcal{L}_{C!}$ diremos que es *maximal consistente de* $adec(\varphi)$ sii $\Theta \subseteq adec(\varphi)$, Θ es consistente y no hay un Θ' consistente tal que $\Theta \subsetneq \Theta' \subseteq adec(\varphi)$.

Como $adec(\varphi)$ es finito sus maximales consistentes también son finitos. Dado Θ un maximal consistente de $adec(\varphi)$ vamos a notar $\bar{\Theta} := \bigwedge \Theta = \bigwedge_{\psi \in \Theta} \psi$, que está bien definido por ser finito.

A continuación vamos a definir un modelo canónico de manera ligeramente diferente a lo visto en el Capítulo 3. Las diferencias son simplemente para facilitar las demostraciones para este tipo de modelos, mientras que antes se hizo de un modo más genérico para poder generalizar con facilidad.

Definición 4.1.27. Sea φ una fórmula. Definimos el modelo canónico para φ en $\mathcal{L}_{C!}$ como $\mathfrak{M}^C(\varphi, \mathcal{L}_{C!}) = \langle W^C, R_1^C, \ldots, R_n^C, R_c^C, R_e^C, \mathtt{i}^c \rangle$ donde:

- $W^C = \{w_\Theta \ : \ \Theta \text{ es maximal consistente en } adec(\varphi)\}$
- $w_\Theta R_i^C w_\Delta$ sii $\Theta/K_i = \Delta/K_i$
- $R_e^C = R_1^C \cup \ldots \cup R_n^C$
- $R_c^C = (R_e^C)^*$
- $\mathtt{i}^C(p) := \{w_\Theta \ : \ p \in \Theta\}$

Observación 4.1.28. Sean $adec(\varphi)$ un conjunto adecuado para alguna fórmula y $\mathfrak{M}^C(\varphi, \mathcal{L}_{C!}) = \langle W^C, R_1^C, \ldots, R_n^C, R_e^C, R_c^C, \mathtt{i}^C \rangle$ su modelo canónico. Entonces las relaciones R_i^C resultan reflexivas, transitivas y euclideanas. Es decir, el modelo $\mathfrak{M}^C(\varphi, \mathcal{L}_{C!})$ es un modelo epistémico para n agentes.

Lema 4.1.29. *Sea φ una fórmula. Todo subconjunto consistente de $adec(\varphi)$ está contenido en un maximal consistente de $adec(\varphi)$.*

Demostración. De manera análoga a la Proposición 2.6.6 pero tomando una enumeración de $adec(\varphi)$. □

Definición 4.1.30. Dado Θ un maximal consistente de $adec(\varphi)$, un *camino desde* Θ, es una secuencia $\Theta_0, \ldots, \Theta_r$ de maximales consistentes de $adec(\varphi)$ con $\Theta = \Theta_0$ tal que para todo k hay un agente i_k para el cual vale $w_{\Theta_k} R_{i_k}^C w_{\Theta_{k+1}}$. Llamamos *longitud* del camino desde Θ a r (la cantidad de elementos de la secuencia).

Llamamos *camino-ψ desde* Θ a un camino desde Θ en el que todos los maximales Θ_k por los que pasa cumplen $\psi \in \Theta_k$.

Lema 4.1.31. *Sea φ una fórmula, como $adec(\varphi)$ es finito, tiene finitos maximales consistentes. Por lo tanto $\bigvee\{$*
$bar\Theta$: es maximal consistente de $adec(\varphi)\}$ está bien definido y vale que $\vdash \bigvee\{\bar{\Theta} : \Theta$ es maximal consistente de $adec(\varphi)\}$.

Proposición 4.1.32. *Sea φ una fórmula y sea $\mathfrak{M}^C(\varphi, \mathcal{L}_{C!}) = \langle W^C, R_1^C, \ldots, R_n^C, R_e^C, R_c^C, \mathtt{i}^C \rangle$ el modelo canónico asociado. Si Θ y Δ son maximales consistentes de $adec(\varphi)$, valen las siguientes afirmaciones:*

1. *Θ es cerrado por implicación en $adec(\varphi)$ (para toda $\psi \in adec(\varphi)$ si $\vdash \bar{\Theta} \to \psi$ entonces $\psi \in \Theta$)*
2. *Si $\neg\psi \in adec(\varphi)$, vale $\psi \in \Theta$ sii $\neg\psi \notin \Theta$*
3. *Si $\psi \wedge \chi \in adec(\varphi)$, vale $\psi \wedge \chi \in \Theta$ sii $\psi \in \Theta$ y $\chi \in \Theta$*
4. *Si $\bar{\Theta} \wedge \neg K_i \neg \bar{\Delta}$ es consistente entonces $\Theta R_i^C \Delta$*
5. *Si $K_i\chi \in adec(\varphi)$, vale $\{K_i\psi : K_i\psi \in \Theta\} \vdash \chi$ sii $\{K_i\psi : K_i\psi \in \Theta\} \vdash K_i\chi$*
6. *Si $C\psi \in adec(\varphi)$, vale $C\psi \in \Theta$ sii todo camino desde Θ es un camino-ψ desde Θ*
7. *Si $[!\psi]C\chi \in adec(\varphi)$, vale $[!\psi]C\chi \in adec(\varphi)$ sii todo camino-ψ desde Θ es un camino-$[!\psi]\chi$.*

Demostración. Del 1 al 5 se desprenden de las definiciones. Veamos los casos que tienen un poco más de interés.

6 $\Rightarrow$) Lo demostraremos por inducción en la longitud del camino desde Θ, de hecho probaremos algo más fuerte: si $C\psi \in adec(\varphi)$ todo camino desde Θ es un camino-$C\psi$ y un camino-ψ desde Θ:

`Caso base`: Supongamos que la longitud del camino es 0. Eso es que el camino empieza y termina en Θ, por lo tanto alcanza con ver que vale $w_\Theta \Vdash_{\mathfrak{M}^C(\varphi,\mathcal{L}_{C!})} \psi$. Pero como vale $C\psi \to \psi \in \mathbf{PAC}$ entonces tenemos que tanto $C\psi$ como $\psi \in adec(\varphi)$. Pero como Θ es cerrado por implicaciones en $adec(\varphi)$, tenemos que $C\psi \in \Theta$ y $\psi \in \Theta$.

`Caso inductivo`: Supongamos que si $C\psi \in adec(\varphi)$ vale que los caminos desde Θ de longitud n son caminos-ψ y caminos-$C\psi$. Tomamos un camino desde Θ de longitud $n+1$, por hipótesis inductiva tenemos que $C\psi \in \Theta_n$. Sea i el agente tal que $w_{\Theta_n} R_i^C w_{\Theta_{n+1}}$, como en $\mathbf{PAC}\vdash C\psi \to EC\psi$ y $\vdash EC\psi \to KC\psi$, entonces tiene que suceder que $K_iC\psi \in \Theta_{n+1}$ (por definición de R_i^C y el hecho de que $K_iC\psi \in adec(\varphi)$). Razonando del mismo modo que en el caso base, obtenemos que $C\psi \in \Theta_{n+1}$ y $\psi \in \Theta_{n+1}$.

6 $\Leftarrow$) Supongamos que todo camino desde Θ es un camino-ψ y que $C\psi \in adec(\varphi)$. Llamaremos M^C al conjunto de los maximales consistentes de $adec(\varphi)$ y M_φ el conjunto de los Δ maximales consistentes de $adec(\varphi)$ tales que todo camino desde Δ es un camino-ψ. Consideremos la fórmula $\chi := \bigvee_{\Delta\in M_\psi} \bar{\Delta}$.

Vamos a probar:

(a) $\vdash \bar{\Theta} \to \chi$

(b) $\vdash \chi \to \psi$

(c) $\vdash \chi \to E\chi$

Y de esto se desprende que $C\psi \in \Theta$: de (c) se tiene (por necesitación de C) que $\vdash C(\chi \to E\chi)$ y por el axioma (Escalar) más modus ponens obtenemos que $\vdash \chi \to C\chi$. Por (a) esto implica que $\vdash \bar{\Theta} \to C\chi$ y por (b) junto con necesitación y el axioma K para C tenemos que $\vdash \bar{\Theta} \to C\psi$. Por lo tanto $C\psi \in \Theta$.

Veamos que valen:

(a) $\Theta \vdash \chi$ porque $\bar{\Theta}$ es una de las disyunciones de Θ.

(b) Observemos que si $\Delta \in M_\psi$ entonces $\psi \in \Delta$, en particular: si $\Delta \in M_\psi$ vale que $\bar{\Delta} \to \psi$. Por lo tanto, como está en cada término de la disyunción χ vale que $\vdash \chi \to \psi$.

(c) Supongamos que $\nvdash \chi \to E\chi$, esto es equivalente a que $\chi \wedge \neg E\chi$ sea consistente. Como χ es una disjunción, tiene que haber algún disjunto $\bar{\Delta}_0$ tal que $\bar{\Delta}_0 \wedge \neg E\chi$ es consistente. Razonando similarmente, tiene que haber algún agente i tal que $\bar{\Delta}_0 \wedge \neg K_i\chi = \bar{\Delta}_0 \wedge \neg K_i(\bigvee_{\Delta\in M_\psi} \bar{\Delta})$ es consistente. Es decir que $\nvdash \neg\bar{\Delta}_0 \vee K_i\chi$, lo cual nos dice por un lado que $\nvdash K_i\chi$ y por otro que

$$\nvdash \neg\bar{\Delta}_0. \tag{4.1}$$

.

Si llamamos $\alpha := \bigvee_{\Delta\in M^C} \bar{\Delta}$, por el Lema 4.1.31 tenemos que $\vdash \alpha$, entonces por necesitación: $\vdash K_i\alpha$. Pero $\alpha = \chi \vee \beta$ (con $\beta := \bigvee_{\Delta\in M^C\setminus M_\psi} \bar{\delta}$), entonces tenemos $\vdash K_i(\chi \vee \beta)$ y $\nvdash K_i\chi$.

Si sucediera $\vdash K_i\neg\beta$, como $\vdash (K_i\neg\beta \wedge K_i(\chi \vee \beta)) \rightarrow K\chi$, tendríamos que $\vdash \chi$ y no es el caso.

Por lo tanto, vale $\nvdash K_i\neg\beta$ y esto tiene como consecuencia que existe un $\Delta_1 \in M^C \setminus M_\psi$ tal que $\nvdash K_i\neg\bar{\Delta}_1$ (porque si todos fueran demostrables entonces también lo sería $K_i\neg\beta$). Por ende podemos afirmar que $\nvdash \bigwedge_{\Delta\in M^C\setminus M_\psi} \bar{\Delta}$ o reformulando un poco apropiadamente: $\nvdash \neg\left(\bigvee_{\Delta\in M^C\setminus M_\psi} \neg K_i\neg\bar{\Delta}\right)$. Esto último junto con 4.1.2 resulta en que $\bar{\Delta}_0 \wedge \bigvee_{\Delta\in M^C\setminus M_\psi} \neg K_i\neg\bar{\Delta}$ es consistente.

Por lo tanto debe existir un $\Gamma \in M^C \setminus M_\psi$ tal que $\bar{\Delta}_0 \wedge \neg K_i\neg\bar{\Gamma}$ es consistente. Entonces por el ítem 4) de esta misma Proposición, $w_{\Delta_0} R_i^C w_\Gamma$. Como $\Theta \notin M_\psi$, existe un camino desde Γ que no es un camino-ψ. Pero entonces existe un camino desde Δ_0 que no es un camino-ψ lo que contradice el hecho de que $\Delta_0 \in M_\psi$, lo que sucede porque es uno de los disyuntos de χ.

Por lo tanto $\vdash \chi \rightarrow E\chi$.

7 $\Rightarrow$) Vamos a ver algo más fuerte: si $[!\psi]C\theta \in \Theta$ todo camino-ψ desde Θ también es un camino-$[!\psi]\theta$ y un camino-$[!\psi]C\theta$. Lo veremos por inducción en la longitud del camino.

`Caso base`: Si la longitud del camino es 0, sólo hay que ver que $[!\psi]\theta \in \Theta$ (porque $[!\psi]C\theta \in \Theta$ es hipótesis). Como $\vdash C\theta \rightarrow \theta$, por necesitación y el axioma K de $[!\psi]$ tenemos que $\vdash [!\psi]C\theta \rightarrow [!\psi]\theta$. Por la condición 9) de la Definición 4.1.24 y el hecho de que Θ es cerrado por implicación en $adec(\varphi)$ (por el ítem 1 de esta misma Proposición) tenemos que $[!\psi]\theta \in \Theta$.

`Caso inductivo`: Vamos a suponer que todo camino-ψ desde Θ de longitud n es un camino-$[!\psi]\theta$ y un camino-$[!\psi]C\theta$. Supongamos ahora que tenemos un camino-ψ desde Θ de longitud $n+1$, es decir, tenemos una sucesión de $\Theta = \Theta_0 R_{i_1}^C \ldots R_{i_{n+1}}^C \Theta_{n+1}$ tales que $\psi \in \Theta_k$. Por hipótesis inductiva podemos asumir que $[!\psi]C\theta \in \Theta_n$, y además por los axiomas tenemos que $\vdash C\theta \rightarrow K_iC\theta$ que al aplicarle necesitación y el axioma K para $[!\psi]$ resulta en $\vdash [!\psi]C\theta \rightarrow [!\psi]K_iC\theta$. Por el axioma **Anuncio-Conocimiento** tenemos que $\vdash [!\psi]K_iC\theta \leftrightarrow (\psi \rightarrow K_i[!\psi]C\theta)$. Como $\psi \in \Theta_n$, vale que $\Theta_n \vdash K_i[!\psi]C\theta$. Entonces, por definición de R_i^C tenemos que $K_i[!\psi]C\theta \in \Theta_{n+1}$. Razonando como en el caso base, tenemos que $[!\psi]\theta \in \Theta_{n+1}$.

7 $\Leftarrow$) Supongamos que todo camino-ψ desde Θ es un camino-$[!\psi]\theta$. Sea M^C el conjunto de maximales consistentes de $adec(\varphi)$ y $M_{\psi,[!\psi],\theta} = \{\Delta \in M^C$: tales que todo camino-ψ desde Δ es un camino-$[!\psi]\theta\}$. Ahora consideramos la fórmula $\chi := \bigvee_{\Delta\in M_{\psi,[!\psi]\theta}} \bar{\Delta}$.

Vamos a probar:

(a) $\vdash \bar{\Theta} \to \chi$

(b) $\vdash \chi \to [!\psi]\theta$

(c) $\vdash \chi \wedge \psi \to E\chi$

Si demostramos eso, obtenemos que $[!\psi]C\theta \in \Theta$, porque aplicando la regla de inferencia de anuncio público y conocimiento común a (b) y (c) se sigue que $\vdash \chi \to [!\psi]C\theta$ y por (a) se tiene que $\vdash \bar{\Theta} \to [!\psi]C\theta$. Por lo tanto $[!\psi]C\theta \in \Theta$.

Veamos que valen:

(a) Se desprende de que $\Theta \in M_{\psi,[!\psi]\theta}$ y que $\bar{\Theta}$ es uno de los disyuntos de χ.

(b) Como todo camino-ψ desde Δ es un camino-$[!\psi]\theta$ para todo $\Delta \in M_{\psi,[!\psi]\theta}$, entonces tiene que suceder $[!\psi]\theta \in \Delta$. Por lo tanto, $[!\psi]\theta$ es uno de los términos de la conjunción $\bar{\Delta}$ y en consecuencia, por razonamiento proposicional, se tiene que $\vdash \chi \to [!\psi]\theta$.

(c) Supongamos que no. Esto significa que $\chi \wedge \psi \wedge \neg E\chi$ es consistente. Como χ es una disyunción tiene que existir un disyunto $\bar{\Delta}_0$ de χ tal que $\bar{\Delta}_0 \wedge \psi \wedge \neg E\chi$ es consistente. Como $\bar{\Delta}_0 \wedge \psi$ es consistente y Δ_0 es maximal consistente en $adec(\varphi)$ y $\psi \in adec(\varphi)$ entonces $\psi \in \Delta_0$. En particular $\bar{\Delta}_0 \wedge \psi = \bar{\Delta}_0$. Por lo tanto $\bar{\Delta}_0 \wedge \neg E\chi$ es consistente. Razonando igual que en 6 $\Leftarrow$)(c) existe un i, y un $\Gamma \notin M_{\psi,[!\psi]\theta}$ tal que $\bar{\Delta}_0 \wedge K_i\bar{\Gamma}$ es consistente. Por el ítem 4) de esta misma Proposición tenemos que esto es equivalente a $\Delta_0 R_i^C \Gamma$. Pero como hay un camino-ψ desde Γ que no es un camino-$[!\psi]\theta$ (ya que $\Gamma \notin M_{\psi,[!\psi]\theta}$) entonces hay un camino-ψ desde Δ_0 que no es un camino-$[!\psi]\theta$ y esto no es posible. Por lo tanto $\vdash \chi \wedge \psi \to E\chi$.

□

Definición 4.1.33. Definimos la *función de complejidad para* $\mathcal{L}_{C!}$ como la extensión de la función de complejidad para $\mathcal{L}_!$ en la que vale $c(C\varphi) = 1 + c(\varphi)$.

Lema 4.1.34. *La función de complejidad para* $\mathcal{L}_{C!}$ *cumple todas las propiedades enunciadas en el Lema 4.1.13 y esta:*

7. $c([!\varphi]C\psi) > c([!\varphi]\psi)$

Proposición 4.1.35. *Sea* φ *una fórmula y sea* $\mathfrak{M}^C(\varphi, \mathcal{L}_{C!}) = \langle W^C, R_1^C, \ldots, R_n^C, R_e^C, R_c^C, \mathtt{i}\rangle$ *el modelo canónico para* φ*. Para todo* Θ *maximal consistente de* $adec(\varphi)$ *y todo* $\psi \in adec(\varphi)$ *vale:*

$$\psi \in \Theta \text{ sii } w_\Theta \Vdash_{\mathfrak{M}^C(\varphi,\mathcal{L}_{C!})} \psi$$

Demostración. Sea $\psi \in adec(\varphi)$, vamos a demostrarlo por inducción en $c(\psi)$.

`Caso base`: Se corresponde con el hecho de que $\psi = p \in \Phi$, y por definición de $\mathtt{i}^C$ vale que $w_\Theta \Vdash_{\mathfrak{M}^C(\varphi,\mathcal{L}_{C!})} p$ sii $p \in \Theta$.

`Caso inductivo`: Supongamos que para toda ψ con $c(\psi) < n$ vale $w_\Theta \Vdash_{\mathfrak{M}^C(\varphi,\mathcal{L}_{C!})} \psi$ sii $\psi \in \Theta$, y que $c(\psi) = n$.

Los casos de conjunción, implicación, negación y $K_i\chi$ son exactamente iguales a la demostración del Teorema 3.0.22. Veamos el resto de los casos:

Si $\psi = C\chi$: $C\chi \in \Theta$ sii (Proposición 4.1.32 ítem 6) todo camino desde Θ es un camino-χ sii (por hipótesis inductiva) todo camino desde Θ es a lo largo de maximales Θ_k cumpliendo $w_{\Theta_k} \Vdash_{\mathfrak{M}^C(\varphi,\mathcal{L}_{C!})} \chi$ sii $w_\Theta \Vdash_{\mathfrak{M}^C(\varphi,\mathcal{L}_{C!})} C\chi$.

Si $\psi = [!\chi]p$: Supongamos que $[!\chi]p \in \Theta$, como $[!\chi]p = \psi \in adec(\varphi)$, es equivalente a $\chi \to p \in \Theta$ por el axioma de permanencia atómica. Por hipótesis inductiva (que puede aplicarse por la condición 2 del Lema 4.1.13) esto es equivalente a $w_\Theta \Vdash_{\mathfrak{M}^C(\varphi,\mathcal{L}_{C!})} \chi \to p$ que, por definición, es equivalente a $w_\Theta \Vdash_{\mathfrak{M}^C(\varphi,\mathcal{L}_{C!})} [!\chi]p$.

Si $\psi = [!\chi]\neg\theta$: Supongamos $[!\chi]\neg\theta \in \Theta$, como $[!\chi]\neg\theta \in adec(\varphi)$, es equivalente a $\chi \to \neg[!\chi]\theta \in \Theta$ (por el axioma anuncio-negación). Por hipótesis inductiva (que puede aplicarse por la condición 3 del Lema 4.1.13) esto es equivalente a $w_\Theta \Vdash_{\mathfrak{M}^C(\varphi,\mathcal{L}_{C!})} \chi \to \neg[!\chi]\theta$ que, por definición, es equivalente a $w_\Theta \Vdash_{\mathfrak{M}^C(\varphi,\mathcal{L}_{C!})} [!\chi]\neg\theta$.

Si $\psi = [!\chi](\theta_1 \wedge \theta_2)$: Supongamos que $[!\chi](\theta_1 \wedge \theta_2) \in \Theta$, como $[!\chi](\theta_1 \wedge \theta_2) \in adec(\varphi)$, es equivalente a $([!\chi]\theta_1 \wedge [!\chi]\theta_2) \in \Theta$ (por el axioma anuncio-conjunción). Por hipótesis inductiva (que puede aplicarse por la condición 4 del Lema 4.1.13) esto es equivalente a $w_\Theta \Vdash_{\mathfrak{M}^C(\varphi,\mathcal{L}_{C!})} [!\chi]\theta_1 \wedge [!\chi]\theta_2$ que, por definición, es equivalente a $w_\Theta \Vdash_{\mathfrak{M}^C(\varphi,\mathcal{L}_{C!})} [!\chi](\theta_1 \wedge \theta_2)$.

Si $\psi = [!\chi]K_i\theta$: Supongamos que $[!\chi]K_i\theta \in \Theta$, como $[!\chi]K_i\theta \in adec(\varphi)$, es equivalente a $\chi \to K_i[!\chi]\theta \in \Theta$ (por el axioma anuncio-conocimiento). Por hipótesis inductiva (que puede aplicarse por la condición 5 del Lema 4.1.13) esto es equivalente a $w_\Theta \Vdash_{\mathfrak{M}^C(\varphi,\mathcal{L}_{C!})} \chi \to K_i[!\chi]\theta$ que, por definición es equivalente a $w_\Theta \Vdash_{\mathfrak{M}^C(\varphi,\mathcal{L}_{C!})} [!\chi]K_i\theta$.

Si $\psi = [!\chi][!\theta]\delta$: Supongamos que $[!\chi][!\theta]\delta \in \Theta$, como $[!\chi][!\theta]\delta \in adec(\varphi)$, es equivalente a $[!(\chi \wedge [!\chi]\theta)]\delta \in \Theta$ (por el axioma anuncio-anuncio). Por hipótesis inductiva (que puede aplicarse por la condición 6 del Lema 4.1.13) esto es equivalente a $w_\Theta \Vdash_{\mathfrak{M}^C(\varphi,\mathcal{L}_{C!})} [!(\chi \wedge [!\chi]\theta)]\delta$ que a su vez es equivalente a $w_\Theta \Vdash_{\mathfrak{M}^C(\varphi,\mathcal{L}_{C!})} [!\chi][!\theta]\delta$.

Si $\psi = [!\chi]C\theta$: Supongamos que $[!\chi]C\theta \in \Theta$, como $[!\chi]C\theta \in adec(\varphi)$, es equivalente a que todo camino-χ desde Θ es un camino-$[!\chi]\theta$ (por el ítem 7 de la Proposición 4.1.32). Por la condición 7 del Lema 4.1.34 puede aplicarse hipótesis inductiva, por lo tanto, $[!\chi]C\theta \in \Theta$ resulta equivalente a que todo camino-χ desde Θ es un camino en el que $w_{\Theta_k} \Vdash_{\mathfrak{M}^C(\varphi,\mathcal{L}_{C!})} [!\chi]\theta$. Y por definición esto es equivalente a $w_\Theta \Vdash_{\mathfrak{M}^C(\varphi,\mathcal{L}_{C!})} [!\chi]C\theta$.

□

Ahora ya estamos en condiciones de demostrar la completitud de **PAC**:

Teorema 4.1.36. *Sea* $\mathtt{M}$ *la clase de modelos epistémicos para n agentes. Para toda* $\varphi \in \mathcal{L}_{C!}$ *vale*

$$\Vdash_{\mathtt{M}} \varphi \textit{ sii } \vdash_{\boldsymbol{PAC}} \varphi$$

Demostración. La vuelta es la Proposición 4.1.23. Sea $\varphi \in \mathcal{L}_{C!}$ tal que $\nVdash_{\mathbf{PAC}} \varphi$, y sea $adec(\varphi)$ el conjunto adecuado para $\neg\varphi$. Como $\{\neg\varphi\}$ es consistente, existe Θ un maximal consistente de $adec(\varphi)$ tal que $\neg\varphi \in \Theta$. Sea $\mathfrak{M}^C(\varphi, \mathcal{L}_{C!})$ el modelo canónico para Φ, por el Lema 4.1.35 vale que $w_\Theta \nVdash_{\mathfrak{M}^C(\varphi,\mathcal{L}_{C!})} \varphi$. Como $\mathfrak{M}^C(\varphi, \mathcal{L}_{C!})$ es un modelo de $\mathtt{M}$, $\nVdash_{\mathtt{M}} \varphi$. □

Corolario 4.1.37. *La lógica de anuncios públicos con conocimiento común es decidible.*

Demostración. Se desprende del hecho de que tiene la propiedad de modelo finito. La demostración puede consultarse en [26]. □

Ejemplo 4.1.38. Un ejemplo simple de aplicación de esta lógica es el juego de los chicos embarrados, descripto en la Introducción.

Se trata de un padre que tiene k hijos perfectamente racionales, de los cuales $m \leq k$ tienen barro en la frente. Como ninguno puede ver su propia frente, entonces ninguno sabe si está embarrado o no. Sin embargo todos pueden ver las frentes ajenas. Todo esto es *conocimiento común*. Entonces el padre dice en voz alta que al menos uno de ellos tiene la frente embarrada. Luego ordena:

> "El que **en este momento** sepa que tiene su frente embarrada que de un paso al frente ya mismo. El resto que permanezca en su lugar."

Como los hijos son obedientes (mugrientos, pero obedientes), aquellos que sepan que tienen barro en la frente van a acatar la orden de su padre sin titubear. Si ninguno da un paso al frente el padre repite la orden.

Por un momento supongamos $m = 1$, es decir: solo uno de ellos tiene barro en la frente. Cuando su padre anuncia públicamente que hay al menos uno que tiene la frente sucia, el infante embarrado va a notar que sus $k - 1$ hermanos no tienen barro en la frente, y comprenderá que si hay al menos 1 de k sucio y los otros $k-1$ niños están limpios, entonces el sucio es él mismo. Es decir, que al oír la orden de su padre dará el paso al frente, porque en **ese momento** ya sabía que estaba embarrado. Y no solo eso sino que además, como todos estos niños son racionales, es de público conocimiento que si $m = 1$ sucedería eso.

Pero analicemos un caso más interesante. Pongámonos ahora en los zapatos de uno de los niños, quienes no conocen m.

Supongamos que vemos que sólo uno de nuestros hermanos tiene barro en la frente. Eso no nos permite decidir si nosotros tenemos barro en nuestra frente o no. Por lo tanto al momento en el que nuestro padre habló no sabemos nuestra condición, y entonces no damos un paso hacia adelante. Y es de público conocimiento que cualquier otro hermano que vea sólo un hermano sucio actuaría del mismo.

En el momento en el que nuestro padre da la orden no tenemos idea de si estamos sucios o no.

- ¿Qué pasaría si nuestra frente estuviese limpia?

Estaríamos en el caso $m = 1$, que analizamos recién y (como es de *conocimiento público*) el hermano embarrado comprendería su condición de mugriento y daría un paso al frente inmediatamente cuando nuestro padre da la orden.

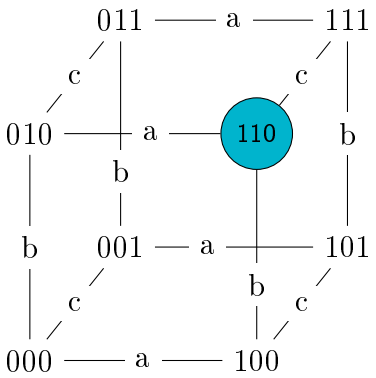
- ¿Qué pasaría si nuestra frente estuviese sucia?

Entonces nuestro hermano roñoso estaría en exactamente la misma situación que nosotros: vería una frente fraterna sucia. Y, como es de *conocimiento público*, si alguien se encuentra en esa situación, no puede concluír nada sobre el estado de su frente. Por lo tanto, el hermano sucio no daría un paso al frente al terminar de hablar nuestro padre.

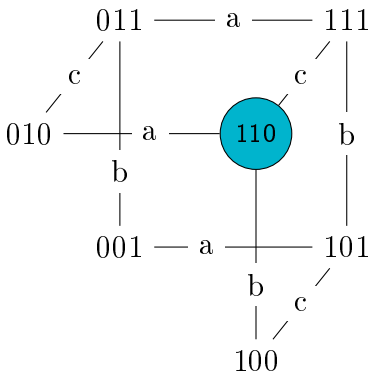
Por lo tanto, si bien mientras nuestro padre habla no tenemos idea de nuestro estado, inmediatamente después de su anuncio lo sabremos, y esto también es de *conocimiento público*.

Supongamos que el caso era que nosotros también estábamos sucios (es decir, $m = 2$). Si nuestro padre diera la orden por segunda vez, tendríamos que dar el paso al frente. Y el hermano embarrado también, porque todos estos razonamientos los puede hacer cualquiera de los niños y es de *conocimiento público* que los pueden hacer todos.

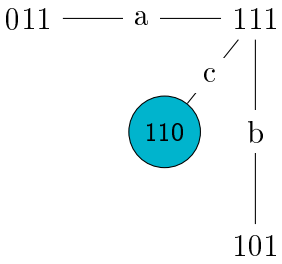
Análogamente si $m = 3$, a la primera orden del padre nadie podría decidir su estado. A la segunda tampoco. Pero si después de la segunda orden no hubo nadie que diera un paso hacia adelante, todos los que tuvieran su frente manchada comprenderían que están sucios, y al tercer anuncio avanzarían todos (Observación: los que tienen su frente limpia no saben que están limpios hasta que sus hermanos sucios avanzan). Ver Figura 4.2 para visualizar los modelos transitados en el caso $k = 3$ y $m = 2$, donde los nodos tienen una terna de 0 y 1 representando la suciedad (1) o pulcritud (0) de cada niño y las aristas representan la relación de indistinguibilidad para los niños a, b y c.



(a) Antes de que el padre hable



(b) Después del primer anuncio



(c) Después del segundo anuncio

(d) Después del tercer anuncio

Figura 4.2: El modelo inicial y su evolución al actualizar los estados epistémicos de los niños.

Sea d_i la variable proposicional para "el niño i está sucio". Entonces las siguientes abreviaciones pueden ser hechas con PAC:

$$vision := \bigwedge_{i \in N} \bigwedge_{j \neq i} ((d_j \leftrightarrow K_i d_j) \wedge (\neg d_j \leftrightarrow K_i \neg d_b))$$

$$alMenosUno := \bigvee_{i \in N} d_i$$

$$exactamenteM := \bigvee_{G \subseteq N, |G| = m} \left(\bigwedge_{i \in G} d_i \wedge \bigwedge_{i \notin G} \neg d_i \right)$$

$$nadieSabe := \bigwedge_{i \in N} (\neg K_i d_i \wedge \neg K_i \neg d_i)$$

$$losSuciosSabenQueLoEstan := \bigwedge_{i \in N} (d_i \rightarrow K_i d_i)$$

Razonando con PAC puede demostrarse que:

$$(exactamenteM \wedge Cvision) \rightarrow [!(alMenosUno)][!(nadieSabe)]^{m-1} losSuciosSabenQueLoEstan.$$

Es decir que lleva m anuncios del padre hasta que los sucios conocen su estado.

Al ser una lógica decidible este y otros juegos de información perfecta pueden ser resueltos algorítmicamente siendo modelados con PAC.

4.2. Lógica de eventos

Hasta ahora nuestros modelos de Kriple sólo capturaban situaciones epistémicas, es decir, sólo contenían información estática. En consecuencia podemos representar el resultado de los escenarios, pero no qué es lo que está sucediendo. Los escenarios epistémicos involucran varios tipos de cambios que pueden afectar el estado del conocimiento: anuncios públicos (mostrar una carta en un partido de póker), anuncios privados y legales (mirar una carta propia en un partido de póker frente a los otros jugadores, sin que nadie más la vea), anuncios privados e ilegales (mostrarle a un amigo una carta en un partido de póker sin que nadie más lo note), etcétera.

Queremos usar modelos de Kripke para representar tales tipos de eventos epistémicos en una forma similar a las representaciones que tenemos para estados: cada agente puede distinguir o no un evento de cualquier otro. Hasta ahora teníamos modelos de estados, ahora vamos a estudiar modelos de eventos. Este enfoque fue presentado en [5], donde se expuso por primera vez la lógica de eventos y se axiomatizó.

Definición 4.2.1. Un *modelo de eventos* es una terna $\mathbf{\Sigma} = \langle \Sigma, \overset{i \in N}{\rightarrow}, pre \rangle$ en la que Σ es un conjunto no vacío a lo sumo numerable, $\overset{i \in N}{\rightarrow}$ son relaciones de alternativas epistémicas para

cada agente, y *pre* es una función de precondiciones que asocia a cada $\sigma \in \Sigma$ una fórmula $pre(\sigma)$.

Llamaremos *eventos simples o acciones* a los $\sigma \in \Sigma$.

Para $\sigma \in \Sigma$ interpretamos $pre(\sigma)$ como la precondición de la acción σ: $pre(\sigma)$ será verdadera en un mundo sii σ puede ejecutarse. En algún sentido, $pre(\sigma)$ da la información implícita que trae la acción σ.

Finalmente, las relaciones de accesibilidad expresan las alternativas que los agentes consideran que pueden estar siendo ejecutadas cuando sucede una acción.

Definición 4.2.2. Llamamos *acciones determinísticas* a los eventos simples $\sigma \in \Sigma$ que son puramente epistémicos, es decir: si no cambian "la realidad" del mundo, simplemente el estado epistémico o doxástico de los agentes.

Para fijar conceptos, damos una serie de ejemplos antes de continuar.

Nota. En adelante, en todos los diagramas de modelos de eventos los eventos serán representados por nodos rectangulares, el contenido del nodo será la precondición del evento, si algún nodo está pintado será el evento real y las aristas representan las relaciones de indistinguibilidad de eventos de los agentes. Recordemos que en los diagramas de modelos epistémicos usamos nodos circulares, el contenido de los nodos es una descripción del estado y si algún nodo está pintado es el estado real.

Ejemplo 4.2.3. El modelo de eventos para el anuncio público $!\varphi$ es el siguiente:

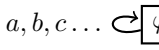


Figura 4.3: Un anuncio público modelado como evento.

Este es el modelo correcto ya que al provenir de una fuente certera, $!\varphi$ sólo puede suceder cuando vale φ. Es decir, φ es precondición de $!\varphi$. Y además al ser un anuncio **público**, todos saben que el evento está sucediendo y no consideran posible ningún otro evento.

Ejemplo 4.2.4. Otro tipo de acción epistémica es el anuncio de φ completamente privado para un grupo G de agentes, es una acción epistémica que consiste en una comunicación de manera abierta para todos los agentes de G en la que el mensaje es φ y proviene de una fuente infalible de información, sin que ningún agente que no pertenece al grupo G se percate.

Esto puede ser modelado con un modelo de dos eventos: el evento real en el que un grupo G está manteniendo una comunicación en la que se anuncia φ, y un evento en el que no sucede nada. La precondición del primer evento es que valga φ, mientras que en el segundo no hace falta que valga nada en particular, es decir, es $\top$.

Los agentes que no están en G consideran que la única alternativa posible es que no suceda nada, mientras que los agentes que están en G consideran como único evento la comunicación privada de φ. El modelo de eventos puede diagramarse del siguiente modo:

Figura 4.4: Modelo de eventos del anuncio privado $!_G\varphi$

Estos modelos son útiles para representar situaciones en las que algún agente espía o comunica secretamente algo.

Ejemplo 4.2.5. En algunos contextos existe un tipo de acción epistémica que se conoce como *anuncio privado justo*, en el cual es de conocimiento común que un agente a aprende de manera privada si φ o $\neg\varphi$. A estos eventos se los denota como $Fair_a\varphi$.

Por ejemplo, si tenemos los agentes a, b y c, donde c arroja una moneda y a y b pueden hacer apuestas hasta que c decida mostrar la moneda, dando por terminado el juego. Los jugadores a y b no pueden mirar la moneda durante el juego, sin embargo c puede mirar la moneda antes de destaparla y esto será visto públicamente.

El modelo de eventos de esa situación es este:

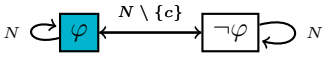


Figura 4.5: Diagrama de un anuncio privado justo

Ejemplo 4.2.6. Supongamos que tenemos otro escenario con tres agentes a, b y c, se arroja una moneda y se la tapa para que nadie vea el resultado. En este juego nadie puede mirar la moneda, pero c la mira ilegalmente sin que nadie lo note. Luego c envía un mensaje secreto a a para anunciarle que salió cara y ese mensaje es interceptado (secretamente) por b y luego llega a a. Esto es un ejemplo de anuncio privado con intercepción secreta de un grupo de outsiders. El modelo de eventos es el siguiente

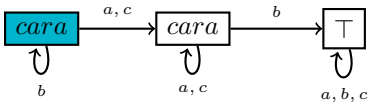


Figura 4.6: Diagrama de un anuncio privado interceptado secretamente por un grupo de outsiders.

Nota. En adelante, para diferenciar los modelos de eventos de los modelos epistémicos o doxásticos tradicionales vamos a llamar a estos últimos *modelos de estados*.

Pero un modelo de eventos en sí mismo no provee suficiente información como para modelar la información que posee un agente, porque mismos eventos producen efectos diferentes en distintos contextos. Es preciso proveer un contexto a través de un modelo de estados. En la siguiente definición vemos cómo hacerlo.

Definición 4.2.7. Sean $\mathfrak{M} = \langle W, (R_i)_{i\in N}, \mathtt{i}\rangle$ un modelo de estados y $\mathbf{\Sigma} = \langle \Sigma, (\xrightarrow{i})_{i\in N}, pre\rangle$ un modelo de eventos. Llamamos *producto-actualización* al modelo de estados $\mathfrak{M} \otimes \mathbf{\Sigma} := \langle W \otimes \Sigma, (R'_i)_{i\in N}, \mathtt{i}'\rangle$ dado por:

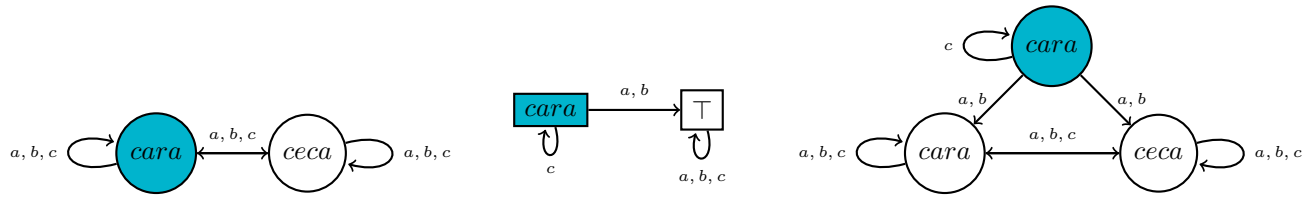
- $W \otimes \Sigma = \{(w, \sigma) \in W \times \Sigma : w \Vdash_{\mathfrak{M}} pre(\sigma)\}$
- $(w, \sigma)R'_i(w', \sigma')$ sii wR_iw' y $\sigma \xrightarrow{i} \sigma'$
- $\mathtt{i}'(p) := \{(w, \sigma) \in W \otimes \Sigma : w \in \mathtt{i}(p)\}$

Las relaciones de modelo representan los estados indintiguibles para un agente, y esto sucede si fueron resultado de eventos indistinguibles ejecutados en estados indistinguibles. Esto implica dos cosas: el conocimiento sólo puede ser obtenido a través de acciones, y una vez obtenido el conocimiento nunca se pierde.

El hecho de que la interpretación de un estado del producto es la misma que en el estado inicial nos dice que estas son acciones puramente epistémicas: no cambia el estado del mundo sino la información que se tiene sobre él.

Cuando queremos diferenciar un estado o evento del resto, consideramos los modelos puntuados. Definimos el producto actualización de $(\mathfrak{M}, w^*) \otimes (\mathbf{\Sigma}, \sigma^*)$, como un modelo punteado donde el modelo es $\mathfrak{M} \otimes \mathbf{\Sigma}$ y el estado diferenciado es (w^*, σ^*).

Ejemplo 4.2.8. Tenemos tres agentes a, b y c, y este último tira una moneda sin que ninguno de ellos pueda ver si salió cara o ceca. Esto está capturado en el modelo de estados de la Figura 4.7a.



(a) Modelo de estados luego de tirar la moneda

(b) Modelo de eventos cuando c espía

(c) Producto-actualización de ambos modelos

Figura 4.7: Ejemplo de producto-actualización.

Posteriormente, c espía la moneda (sin que nadie se percate) y observa que salió cara. Los agentes a y b no pueden distinguir este evento, del evento *nada* cuya precondición es $\top$ (es decir siempre se cumple) y que representa un evento en el cual no sucede nada, tal como se ve en la Figura 4.7b.

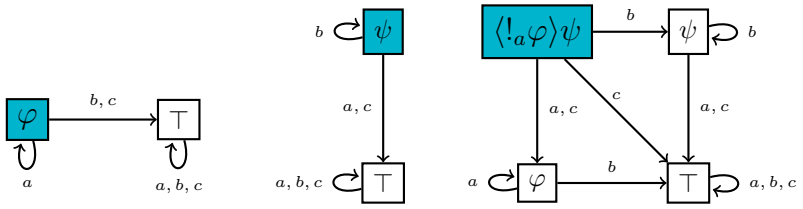
Si se realiza el producto-actualización de estos dos modelos, obtiene el modelo actualizado esperable, que es el que encontramos en la Figura 4.7c

Definición 4.2.9. Sean $\mathbf{\Sigma} = \langle \Sigma, \sim_i, pre\rangle$ y $\mathbf{\Sigma}' = \langle \Sigma', \sim'_i, pre'\rangle$ dos modelos de eventos en $\mathcal{L}_{K\otimes}$. Definimos la *composición de* $\mathbf{\Sigma}$ *y* $\mathbf{\Sigma}'$ como un tercer modelo de eventos notado por $\mathbf{\Sigma}; \mathbf{\Sigma}' = \langle \Sigma'', \sim''_i, pre''\rangle$ donde:

- $\Sigma'' = \Sigma \times \Sigma'$
- $(\sigma, \sigma') \sim''_i (\theta, \theta')$ sii $\sigma \sim_i \theta$ y $\sigma' \sim'_i \theta'$
- $pre''(\sigma, \sigma') = \langle \mathbf{\Sigma}, \sigma \rangle pre'(\sigma')$.

Si los modelos son punteados $(\mathbf{\Sigma}, \sigma^*)$ y $(\mathbf{\Sigma}', \sigma'^*)$ la composición es punteada en (σ^*, σ'^*).

Aquí tenemos un ejemplo gráfico de composición:



(a) Modelo de $!_a\varphi$ (b) Modelo de $!_b\psi$ (c) Su composición

Figura 4.8: Ejemplo de composición secuencial

A continuación vamos a introducir el lenguaje para la lógica de eventos, que incorpora el aspecto dinámico a través de dos operaciones: aplicar el producto actualización al modelo de estados contra un modelo de eventos y ejecutar al azar una de dos opciones. Además, al igual que en las lógicas de anuncios públicos, tenemos un operador para cuantificar sobre estados luego de la ejecución de acciones.

Definición 4.2.10. Dado un conjunto de agentes N, y un conjunto finito de variables proposionales Φ, definimos $\mathcal{L}_{K\otimes}(N, \Phi)$ el *lenguaje de la lógica de eventos basado en Φ para N* como la unión de las *fórmulas* $\varphi \in \mathcal{L}^{stat}_{K\otimes}(N, \Phi)$ y las *acciones* $\alpha \in \mathcal{L}^{act}_{K\otimes}(N, \Phi)$ definidas recursivamente por:

$$\varphi ::= p \mid \bot \mid \neg\varphi \mid \psi \vee \varphi \mid L_i\varphi \mid C_i\varphi \mid [\alpha]\varphi$$
$$\alpha ::= \alpha \cup \beta \mid (\mathbf{\Sigma}, \sigma)$$

donde $p \in \Phi$, $i \in N$ y $\alpha \in \mathcal{L}^{act}_{K\otimes}(N, \Phi)$ y $(\mathbf{\Sigma}, \sigma) = \langle \Sigma, \sim_i, pre, \sigma \rangle$ es un modelo de eventos punteado tal que para todo $\sigma' \in \Sigma$, $pre(\sigma') \in \mathcal{L}^{stat}_{K\otimes}(N, \Phi)$.

Además, notaremos $\langle\alpha\rangle\varphi$ para abreviar $\neg[\alpha]\neg\varphi$ y $\mathbf{\Sigma}$ para abreviar $\bigcup_{\sigma\in\Sigma}(\mathbf{\Sigma}, \sigma)$, y cuando quede claro por el contexto vamos a omitir a N y a Φ de la notación.

Podemos pensar en $\mathcal{L}^{stat}_{K\otimes}$ como las fórmulas que no modifican el modelo y a $\mathcal{L}^{act}_{K\otimes}$ como las que sí.

La operación $\alpha \cup \beta$ es la que representa la ejecución al azar alguna de las dos acciones en el modelo, y la interpretación intuitiva para $\mathbf{\Sigma}, \sigma$ es la de que se aplicará el producto-actualización contra el modelo.

Para definir la semántica del lenguaje es necesario incorporar una nueva relación intermodelos punteados para cada acción del lenguaje:

$$(\mathfrak{M}, w)[[\mathbf{\Sigma}, \sigma]](\mathfrak{M}', w') \textit{ sii } w \Vdash_{\mathfrak{M}} pre(\sigma) y (\mathfrak{M}', w') = (\mathfrak{M} \otimes \mathbf{\Sigma}, (w, \sigma)),$$

$$(\mathfrak{M}, w)[[\alpha \cup \beta]](\mathfrak{M}', w') \textit{ sii } (\mathfrak{M}, w)[[\alpha]](\mathfrak{M}', w') \; o \; (\mathfrak{M}, w)[[\beta]](\mathfrak{M}', w').$$

Esta relación representa el hecho de que se puede llegar del primer modelo de estados al segundo a través de la acción.

Dada una fórmula $\varphi \in \mathcal{L}^{stat}_{K\otimes}(N, \Phi)$, una acción $\alpha \in \mathcal{L}(N, \Phi)$, un modelo epistémico $\mathfrak{M} = \langle W, (R_i)_{i\in N}, \mathtt{i}\rangle$ y un modelo de eventos $\mathbf{\Sigma} = \langle \Sigma, (\sim_i)_{i\in N}, pre\rangle$, definimos la semántica del lenguaje del siguiente modo:

- $w \Vdash_{\mathfrak{M}} p$ sii $w \in \mathtt{i}(p)$
- $w \Vdash_{\mathfrak{M}} \neg\varphi$ sii $w \nVdash_{\mathfrak{M}} \varphi$
- $w \Vdash_{\mathfrak{M}} \varphi \wedge \psi$ sii $w \Vdash_{\mathfrak{M}} \varphi$ y $w \Vdash_{\mathfrak{M}} \psi$
- $w \Vdash_{\mathfrak{M}} K_i\varphi$ sii para todo $v \in W$ tal que wR_iv vale $v \Vdash_{\mathfrak{M}} \varphi$
- $w \Vdash_{\mathfrak{M}} [\alpha]\varphi$ sii para todo $\mathfrak{M}'$ y $w' \in \mathfrak{M}'$ tales que $(\mathfrak{M}, w)[[\alpha]](\mathfrak{M}', w')$ vale $w' \Vdash_{\mathfrak{M}'} \varphi$

Por lo tanto, al igual que en la lógica de anuncios públicos las modalidades $[\alpha]\varphi$ capturan la noción de que después de una acción α valdrá φ.

Proposición 4.2.11. *Sean $\alpha, \beta \in \mathcal{L}^{act}_{K\otimes}$. Se tiene que $\Vdash [\alpha \cup \beta]\varphi \leftrightarrow ([\alpha]\varphi \wedge [\beta]\varphi)$.*

Lo cual responde bien a nuestra intuición previa. Ahora podemos relacionar la composición sintáctica de modelos de eventos con su significado intuitivo: que después de ejecutar al azar alguna de las acciones α o β valga φ es equivalente a que después de ejecutar α valga φ y después de ejecutar β valga φ.

Proposición 4.2.12. *Sean $(\mathbf{\Sigma}, \sigma)$ y $(\mathbf{\Sigma}', \sigma') \in \mathcal{L}^{act}_{K\otimes}$ y $\varphi \in \mathcal{L}^{stat}_{K\otimes}$. Entonces $[(\mathbf{\Sigma}, \sigma); (\mathbf{\Sigma}, \sigma')]\varphi$ es equivalente a $[\mathbf{\Sigma}, \sigma][\mathbf{\Sigma}', \sigma']\varphi$. Es decir, la composición secuencial es efectivamente una composición secuencial.*

Demostración. Sea $\mathfrak{M}$ un modelo epistémico y w uno de sus mundos. Queremos ver que $w \Vdash_{\mathfrak{M}} [(\mathbf{\Sigma}, \sigma); (\mathbf{\Sigma}', \sigma')]\varphi$ sii $w \Vdash_{\mathfrak{M}} [\mathbf{\Sigma}, \sigma][\mathbf{\Sigma}', \sigma']\varphi$. Para ello alcanza con ver que existe un isomorfismo de $\mathfrak{M} \otimes (\mathbf{\Sigma}; \mathbf{\Sigma}')$ en $(\mathfrak{M} \otimes \mathbf{\Sigma}) \otimes \mathbf{\Sigma}'$, es decir una función biyectiva de un dominio de un modelo en el dominio del otro que preserva interpretación y relación. Definimos la aplicación $(w, (\sigma, \sigma')) \mapsto ((w, \sigma), \sigma')$. Veamos que está bien definida:

Sea $(w, (\sigma, \sigma')) \in Dom(\mathfrak{M} \otimes (\mathbf{\Sigma};$ mód $))$, entonces tenemos que $w \Vdash_{\mathfrak{M}} pre''(\sigma, \sigma')$, i.e. $w \Vdash_{\mathfrak{M}} \langle \mathbf{\Sigma}, \sigma\rangle pre'(\sigma')$. Pero esto es equivalente a $w \Vdash_{\mathfrak{M}} pre(\sigma) \wedge [\mathbf{\Sigma}, \sigma] pre'(\sigma')$. De $w \Vdash_{\mathfrak{M}} pre(\sigma)$ se tiene que $(w, \sigma) \in Dom(\mathfrak{M} \otimes \mathbf{\Sigma})$, y de $w \Vdash_{\mathfrak{M}} [\mathbf{\Sigma}, \sigma] pre'(\sigma')$ se tiene que $((w, \sigma), \sigma') \in Dom((\mathfrak{M} \otimes \mathbf{\Sigma}) \otimes \mathbf{\Sigma}')$. El argumento funciona en ambas direcciones.

Además $(w,(\sigma,\sigma'))R_i(v,(\theta,\theta'))$ sii wR_iv y $(\sigma,\sigma') \sim_i'' (\theta,\theta')$ sii wR_iv y $\sigma \sim_i \theta$ y $\sigma' \sim_i' \theta'$ sii $((w,\sigma),\sigma')R_i((v,\theta),\theta')$.

Y la interpretación en ambos coincide con la del modelo de estados inicial, en particular son iguales. □

COMPLETITUD, CORRECTITUD Y DECIBILIDAD

Los resultados de completitud, correctitud y decibilidad para la lógica de eventos se basan en la misma idea que los resultados de PAL: es posible escribir cualquier fórmula de $\mathcal{L}_{K\otimes}$ como una fórmula sin acciones, y luego se usan los resultados de completitud en la lógica epistémica de n agentes aislados. Damos los resultados y algunas demostraciones, mientras que otras las omitimos por ser análogas a las de PAL.

Proposición 4.2.13. *Las siguientes fórmulas son válidas en la lógica de eventos:*

1. $[\mathbf{\Sigma},\sigma]p \leftrightarrow (pre(\sigma) \to p)$
2. $[\mathbf{\Sigma},\sigma]\neg\varphi \leftrightarrow (pre(\sigma) \to \neg[\mathbf{\Sigma},\sigma]\varphi)$
3. $[\mathbf{\Sigma},\sigma](\varphi \wedge \psi) \leftrightarrow ([\mathbf{\Sigma},\sigma]\varphi \wedge [\mathbf{\Sigma},\sigma]\psi)$
4. $[\mathbf{\Sigma},\sigma]K_i\varphi \leftrightarrow (pre(\sigma) \to \bigwedge\limits_{\sigma\sim_i\theta} K_i[\mathbf{\Sigma},\theta]\varphi)$

Demostración. Los primeros 3 ítems se desprenden directamente de la definición. Vamos a demostrar el 4) que es un poco más desafiante.

Vamos a demostrar la forma dual de la proposición, reemplazando φ por $\neg\varphi$. Queremos ver que

$$\langle\mathbf{\Sigma},\sigma\rangle\hat{K}_i\varphi \leftrightarrow (pre(\sigma) \wedge \bigvee_{\sigma\sim_i\theta} \hat{K}_i\langle\mathbf{\Sigma},\theta\rangle\varphi)$$

Sea $\mathfrak{M} = \langle W,(R_i)_{i\in N},\mathtt{i}\rangle$ y $\mathbf{\Sigma} = \langle\Sigma,(\sim_i)_{i\in N},pre\rangle$.

Asumamos que $w \Vdash_{\mathfrak{M}} \langle\mathbf{\Sigma},\sigma\rangle\hat{K}_i\varphi$, por lo tanto se cumple $w \Vdash_{\mathfrak{M}} pre(\sigma)$. Pero además vale que $(w,\sigma) \Vdash_{\mathfrak{M}\otimes\mathbf{\Sigma}} \hat{K}_i\varphi$ y esto implica que existe un (v,θ) tal que wR_iv y $\sigma \sim_i \theta$ y $(v,\theta) \Vdash_{\mathfrak{M}\otimes\mathbf{\Sigma}} \varphi$. Es decir que $v \Vdash_{\mathfrak{M}} \langle\mathbf{\Sigma},\theta\rangle\varphi$, pero como wR_iv entonces vale $w \Vdash_{\mathfrak{M}} \hat{K}_i\varphi$. Como θ era uno que cumplía $\sigma \sim_i \theta$ entonces vale $w \Vdash_{\mathfrak{M}} \bigvee\limits_{\sigma\sim_i\theta} \hat{K}_i\varphi$.

Ahora asumamos que $w \Vdash_{\mathfrak{M}} pre(\sigma)$ y que existe un $\theta \in \Sigma$ tal que $\sigma \sim_i \theta$ y $w \Vdash_{\mathfrak{M}} \hat{K}_i\langle\mathbf{\Sigma},\theta\rangle\varphi$. De $w \Vdash_{\mathfrak{M}} pre(\sigma)$ se sigue que $(w,\sigma) \in \mathfrak{M}\otimes\mathbf{\Sigma}$, y de $w \Vdash_{\mathfrak{M}} \hat{K}_i\langle\mathbf{\Sigma},\theta\rangle\varphi$ se sigue que exite un $v \in W$ tal que wR_iv y $v \Vdash_{\mathfrak{M}} \langle\mathbf{\Sigma},\theta\rangle\varphi$. Esto último significa que $v \Vdash_{\mathfrak{M}} pre(\theta)$, por lo tanto $(v,\theta) \in \mathfrak{M}\otimes\mathbf{\Sigma}$ y $(v,\theta) \Vdash_{\mathfrak{M}\otimes\mathbf{\Sigma}} \varphi$. Pero como wR_iv y $\sigma \sim_i \theta$ obtenemos que $(w,\sigma) \Vdash_{\mathfrak{M}\otimes\mathbf{\Sigma}} \hat{K}_i\varphi$, y esto implica que $w \Vdash_{\mathfrak{M}} \langle\mathbf{\Sigma},\sigma\rangle\hat{K}_i\varphi$. □

Definición 4.2.14. Definimos el sistema axiomático $\mathbf{AM}^2$ dado por una axiomatización de la lógica proposicional junto con los siguientes esquemas de axiomas:

- **(K para K_i)** $K_i(\varphi \to \psi) \to (K_i\varphi \to K_i\psi)$
- **(Veracidad)** $K_i\varphi \to \varphi$
- **(Introspección positiva)** $K_i\varphi \to K_iK_i\varphi$
- **(Introspección negativa)** $\neg K_i\varphi \to K_i\neg K_i\varphi$
- **(Permanencia atómica)** $[\mathbf{\Sigma}, \sigma]p \to (pre(\sigma) \to p)$
- **(Acción-negación)** $[\mathbf{\Sigma}, \sigma]\neg\varphi \leftrightarrow \neg[\mathbf{\Sigma}, \sigma]\varphi$
- **(Acción-conjunción)** $[\mathbf{\Sigma}, \sigma](\varphi \wedge \psi) \leftrightarrow ([\mathbf{\Sigma}, \sigma]\varphi \wedge [\mathbf{\Sigma}, \sigma]\psi)$
- **(Acción-conocimiento)** $[\mathbf{\Sigma}, \sigma]K_i\varphi \leftrightarrow (pre(\sigma) \to \bigwedge_{\sigma \sim_i \theta} K_i[\mathbf{\Sigma}, \theta]\varphi)$
- **(Acción-composición)** $[\mathbf{\Sigma}, \sigma][\mathbf{\Sigma}', \sigma']\varphi \leftrightarrow [(\mathbf{\Sigma}, \sigma); (\mathbf{\Sigma}', \sigma')]\varphi$
- **(Elección no determinística)** $[\alpha \cup \beta]\varphi \leftrightarrow ([\alpha]\varphi \vee [\beta]\varphi)$

junto con las reglas de inferencia modus ponens, necesitación para K_i y necesitación para $[\mathbf{\Sigma}, \sigma]$

Definición 4.2.15. Contruimos la *traducción de* $\mathcal{L}_{K\otimes}$ como una función $t : \mathcal{L}_{K\otimes} \to \mathcal{L}_{K\otimes}$ definida recursivamente del siguiente modo:

$$\begin{aligned}
t(p) &= p \\
t(\neg\varphi) &= \neg(\varphi) \\
t(\varphi \wedge \psi) &= t(\varphi) \wedge (\psi) \\
t(K_i\varphi) &= K_i t(\varphi) \\
t([\mathbf{\Sigma}, \sigma]p) &= t(pre(\sigma) \to p) \\
t([\mathbf{\Sigma}, \sigma]\neg\varphi) &= t(pre(\sigma) \to \neg[\mathbf{\Sigma}, \sigma]\varphi) \\
t([\mathbf{\Sigma}, \sigma](\varphi \wedge \psi)) &= t([\mathbf{\Sigma}, \sigma]\varphi \wedge [\mathbf{\Sigma}, \sigma]\psi) \\
t([\mathbf{\Sigma}, \sigma]K_i\varphi) &= t(pre(\sigma) \to K_i[\mathbf{\Sigma}, \sigma]\varphi) \\
t([\mathbf{\Sigma}, \sigma][\mathbf{\Sigma}', \sigma']\varphi) &= t([(\mathbf{\Sigma}, \sigma); (\mathbf{\Sigma}', \sigma')]\varphi) \\
t([\alpha \cup \beta]\varphi) &= t([\alpha]\varphi \wedge [\beta]\varphi)
\end{aligned}$$

Para ver que esta traducción tiene su imagen en $\mathcal{L}_K$, es necesario extender la función de complejidad:

[2] AM proviene de Action models

Definición 4.2.16. Definimos la *complejidad en* $\mathcal{L}_{K\otimes}$ como $c : \mathcal{L}_{K\otimes} \to \mathbb{N}$ tal que:

$$\begin{aligned} c(p) &= 1 \\ c(\neg\varphi) &= 1 + c(\varphi) \\ c(\varphi \wedge \psi) &= 1 + max(c(\varphi), c(\psi)) \\ c(K_i\varphi) &= 1 + c(\varphi) \\ c([\alpha]\varphi) &= (1 + c(\alpha)) \cdot c(\varphi) \\ c(\mathbf{\Sigma}, \sigma) &= \text{máx}\{c(pre(\theta)) \ : \ \theta \in \mathbf{\Sigma}\} \\ c(\alpha \cup \beta) &= 1 + \text{máx}(c(\alpha), c(\beta)) \end{aligned}$$

Tenemos estos resultados que nos permitirán demostrar por inducción en la complejidad que la traducción efectivamente traduce a $\mathcal{L}_K$.

Lema 4.2.17. *Para todo φ y ψ vale:*

1. $c(\psi) \geq c(\varphi)$ *si* $\varphi \in Sub(\psi)$
2. $c([\mathbf{\Sigma}, \sigma]p) > c(pre(\sigma) \to p)$
3. $c([\mathbf{\Sigma}, \sigma]\neg\varphi) > c(pre(\sigma) \to \neg[\mathbf{\Sigma}, \sigma]\varphi)$
4. $c([\mathbf{\Sigma}, \sigma](\varphi \wedge \psi)) > c([\mathbf{\Sigma}, \sigma]\varphi \wedge [\mathbf{\Sigma}, \sigma]\psi)$
5. $c([\mathbf{\Sigma}, \sigma](K_i\varphi)) > c(pre(\sigma) \to K_i[\mathbf{\Sigma}, \sigma]\varphi)$
6. $c([\mathbf{\Sigma}, \sigma][\mathbf{\Sigma}', \sigma']\varphi) > c([(\mathbf{\Sigma}, \sigma); (\mathbf{\Sigma}', \sigma')]\varphi)$
7. $c([\alpha \cup \beta]\varphi) > c([\alpha]\varphi \wedge [\beta]\varphi)$

Este resultado nos permite probar resultados que utilicen los axiomas de reducción (permanencia atómica, acción-negación, acción-conjunción, etcétera) por induccción en $c(\varphi)$.

En particular tenemos:

Proposición 4.2.18. *Para toda $\varphi \in \mathcal{L}_{K\otimes}$ vale $\vdash \varphi \leftrightarrow t(\varphi)$.*

Y de este resultado se desprende:

Teorema 4.2.19. *Para todo $\varphi \in \mathcal{L}_{K\otimes}$ vale que $\Vdash \varphi$ sii $\vdash \varphi$.*

4.2.1. Cambio de hechos

Uno puede permitir eventos que puedan cambiar algunos hechos además de la información que posee cada agente. Formalmente esto se puede hacer agregando a la estructura de modelo de eventos una función de poscondiciones *post*, que asocia cada evento σ a un conjunto de poscondiciones que notaremos de la forma $p := post_\sigma(p)$ para cada variable proposicional $p \in \Phi$ donde $post_\sigma(p)$ es una fórmula del lenguaje.

El significado intuitivo es que p va a ser cierto después de ejecutar σ sii $post_\sigma(p)$ era verdad antes del evento.

Ejemplo 4.2.20. Situémonos en el escenario de los tres agentes a, b, c, en el que c tira una moneda. La atrapa, la da vuelta frente a todos y la destapa frente a todos mostrando la moneda. Es decir: es de público conocimiento que si había salido cara mostrará ceca y si había salido ceca mostrará cara. Este sería el modelo de eventos:

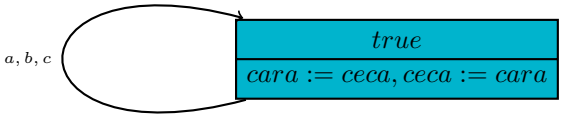


Figura 4.9: Las precondiciones están indicadas arriba y las poscondiciones abajo

Definición 4.2.21. Definimos el *producto-actualización con modelo de eventos cambiantes* del mismo modo que en la Definición 4.2.7 salvo que ahora

$$\mathtt{i}(p)_{\mathfrak{M}\otimes\mathbf{\Sigma}} = \{(w, \sigma) \in \mathfrak{M} \otimes \mathbf{\Sigma} : w \in \mathtt{i}(post_\sigma(p))_{\mathfrak{M}}\}$$

Este producto funciona bien siempre que las creencias falsas no sean contradichas por la nueva información. En tal caso, si un agente a es confrontado con una contradicción entre sus creencias previas y la información nueva, no quedan relaciones a saliendo desde el mundo real. Es decir: comienza a creer todo.

Ejemplo 4.2.22. Supongamos el siguiente escenario: hay tres agentes a, b y c, este último tira una moneda y al caer la tapa antes de que nadie pueda verla. Luego, c espía la moneda, sin que nadie se percate, y observa que salió cara. Para finalizar le manda un mensaje secreto a a que dice "Sé que salió cara."

De acuerdo a nuestra intuición esto debería resultar en un modelo de estado como el siguiente:

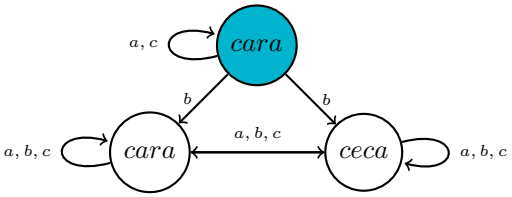
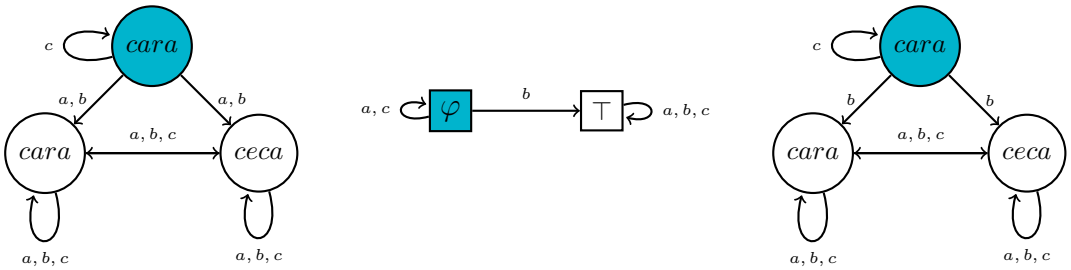


Figura 4.10: Modelo intuitivo resultante de la actualización

Sin embargo, al realizar el producto del modelo de estados de la Figura 4.7c con el modelo de eventos para el anuncio privado de c a a lo que sucede es esto:



(a) Después de espiar

(b) Modelo de $!_{\{a,c\}}K_c cara$

(c) Al actualizar queda distinto a la Figura 4.10

Figura 4.11: Falencia de la lógica de eventos para modelar incorporación de información inconsistente.

Para arreglar este problema es necesario modificar el producto-actualización incorporando ideas de Teoría de Revisión, como veremos en el siguiente capítulo.

Capítulo 5

Revisión y modelos de plausibilidad

> El sabio puede cambiar de opinión. El necio, nunca.
>
> Immanuel Kant.

En el capítulo anterior estudiaremos la manera en la que cambia el estado epistémico de un agente con la incorporación de información. Esta información se interpreta como verídica y proviniente de una fuente infalible de información. Dado que la forma de conocimiento modelada en los capítulos anteriores es veraz, no puede suceder que se anuncie φ en un mundo en el que el agente sabe que vale $\neg\varphi$ ya que esto contradice el hecho de que tanto la información sabida por el agente como la información obtenida por la fuente infalible son veraces. De esta manera un agente nunca se encuentra recibiendo un mensaje que contradice sus conocimientos. En la teoría esto queda capturado en la existencia de precondiciones para la incorporación de nueva información.

En un contexto doxástico, donde no necesariamente lo que se cree es veraz, la imposición de restricciones sobre la información a ser incorporada no es algo propio del problema a modelar. Sin embargo, ¿qué sucedería si un agente aprende un hecho φ que contradice sus creencias anteriores? Recordemos que asumimos que las creencias no necesariamente son veraces, pero sí son consistentes, por lo tanto si acepta que vale φ tiene que descartar parte de lo que cree. ¿Qué parte? ¿Todo lo que contradice φ? No, la idea detrás de todas las teorías de revisión es descartar la menor cantidad posible de información. Hay distintos enfoques y maneras de enfrentar este problema. Vamos a presentar dos de ellos.

5.1. Teoría AGM

El problema de revisión es un problema difícil y no hay un única solución para todos los contextos. Para reducir la dificultad del problema, la teoría AGM toma un enfoque en el cual se revisan creencias sobre fórmulas de lógica proposicional básica. Esto significa que no vamos a revisar creencias sobre creencias, ni creencias sobre creencias sobre creencias, etcétera. Además en esta teoría la revisión es interna a un agente, por esto no se consideran

múltiples agentes. En caso de haberlos, la manera en la que cada uno de ellos revisa sus creencias no va a estar afectada por la presencia de otros agentes.

En esta sección no vamos a hacer hincapié en la parte técnica ni profundizar más allá de los conceptos elementales. Tomamos esta sección como una nota de interés histórico en la evolución de la teoría de revisión. Todos los resultados mencionados pueden encontrarse en [2] o [26].

La teoría de AGM[1] comienza con el paper [2] donde se definen dos tipos elementales de cambio de información: expansión (añadir creencias) y contracción (remover creencias). Con estos operadores se construye un operador de revisión $*$ que toma un conjunto de creencias T y una fórmula a incorporar φ y devuelve las creencias revisadas después de la incorporación de φ.

Definamos algunos conceptos útiles.

Definición 5.1.1. Definimos $Cn(\cdot)$, el *operador de consecuencia clásica* tal que $Cn(\Sigma) = \{\sigma \in \mathcal{L}_p \ : \ \Sigma \vdash \sigma\}$ (donde $\mathcal{L}_p$ denota el lenguaje proposicional).

Definición 5.1.2. Decimos que un conjunto T de fórmulas proposicionales es un *conjunto de creencias* si es no vacío y cerrado bajo consecuencia clásica. Es decir, si $Cn(T) = T \neq \emptyset$.

Definición 5.1.3. Llamaremos *expansión de T con φ* al conjunto $T + \varphi = Cn(T \cup \{\varphi\})$.

Esta operación representa la incorporación de φ, lo cual puede llevar a inconsistencias en $T+\varphi$. Esto nos impide adoptar a la expansión como un operador de revisión, ya que podría violar la consistencia de las creencias de los agentes. Para que esto no suceda quisieramos que al momento de incorporar φ no haya nada que la contradiga. Es decir, queremos quitar las fórmulas que contradicen a φ del conjunto de creencias. Para ello tenemos que definir un nuevo operador de contracción, que vamos a notar con $-$, que tome un conjunto de creencias T y una fórmula ψ y devuelva un conjunto de creencias en el cuál ψ no esté.

Lamentablemente este operador no es tan simple de definir como lo fue el operador de expansión. Una primera aproximación podría ser considerar quitar la información no deseada y clausurar con el operador de consecuencia clásica. Pero esto no es suficiente ya que si tenemos $T = Cn(\{p, q\})$ y queremos renunciar a p, también sucede $p \in Cn(T \setminus \{p\})$ (ya que q y $q \to p \in T \setminus \{p\}$). Otra dificultad notoria es si quisiésemos renunciar a $p \wedge q$, en cuyo caso hay que elegir a cuál de las variables renunciar.

Las siguientes son condiciones deseables para el operador de contracción que estamos buscando.

1. (clausura $-$) $T - \varphi$ es un conjunto de creencias
2. (contracción) $T - \varphi \subseteq T$
3. (mínima acción) Si $\varphi \notin T$, entonces $T - \varphi = T$

[1] AGM son las iniciales de Alchourrón, Gärdenfors y Makinson, quienes dieron comienzo a la teoría de revisión en 1985 a través del paper [2] en el cual introdujeron los operadores de expansión, contracción total y contracción parcial y con ellos el primer operador de revisión.

4. (éxito $-$) Si $\nvdash \varphi$, entonces $\varphi \notin T - \varphi$
5. (recuperación) Si $\varphi \in T$, entonces $T \subseteq (T - \varphi) + \varphi$
6. (extensionalidad $-$) Si $\vdash \varphi \leftrightarrow \psi$, entonces $T - \varphi = T - \psi$
7. (min-conjunción $-$) $(T - \varphi) \cap (T - \psi) \subseteq T - (\varphi \wedge \psi)$
8. (max-conjunción $-$) Si $\varphi \notin T - (\varphi \wedge \psi)$, entonces $T - (\varphi \wedge \psi) \subseteq T - \varphi$

Definición 5.1.4. Llamamos *operador de contracción* a un operador que toma un conjunto de creencias y una fórmula, devuelve otro conjunto de creencias y cumple 1-8.

Las condiciones min-conjunción y max-conjunción establecen relaciones entre la conjunción y la contracción. Esto no determina unívocamente la contracción de una conjunción, lo cual es intuitivamente correcto. Si un agente tiene que renunciar a creer $\varphi \wedge \psi$, la decisión de renunciar a φ, a ψ o a ambos depende del "arraigo" que el agente tenga a cada una de sus creencias.

Por ejemplo, hay dos alumnos en un curso de lógica epistémica uno de ellos es bueno y el otro es excepcionalmente talentoso. Un agente epistémico cree que ambos deben haber aprobado el examen tomado la semana anterior, pero el profesor le informa que al menos uno de ellos no aprobó. Es natural esperar que el agente epistémico renuncie a creer que el alumno bueno aprobó el examen y que conserve la creencia de que el alumno talentoso aprobó.

Una pregunta de interés es si existe otra alternativa diferente de renunciar a uno, otro o ambos conyuntos. El siguiente teorema nos responde.

Teorema 5.1.5. *Sea $-$ un operador que cumple las condiciones 1-6. Entonces $-$ satisface 7 y 8 sii sucede una de las siguientes:*

- $T - (\varphi \wedge \psi) = T - \varphi$
- $T - (\varphi \wedge \psi) = T - \psi$
- $T - (\varphi \wedge \psi) = (T - \varphi) \cap (T - \psi)$

Habiendo definido la expansión y contracción, ahora estamos en condiciones de avanzar hacia un operador de revisión. Queremos definir un operador $*$ de revisión tal que después de revisar con una fórmula φ: a) las creencias sigan siendo consistentes, b) el agente crea φ, c) se pierda la menor cantidad posible de creencias y d) adopte la menor cantidad posible de nuevas creencias. Habiendo definido la expansión y contracción, ahora estamos en condiciones de avanzar hacia un operador de revisión. Queremos definir un operador $*$ de revisión tal que después de revisar con una fórmula φ: a) las creencias sigan siendo consistentes, b) el agente crea φ, c) se pierda la menor cantidad posible de creencias y adopte la menor cantidad posible de nuevas creencias.

Definición 5.1.6. Llamamos *operador de revisión* a un un operador $*$ que toma un conjunto de creencias y una fórmula proposicional y cumple las siguientes:

1. (clausura $*$) $T * \varphi$ es un conjunto de creencias
2. (éxito $*$) $\varphi \in T * \varphi$
3. (cota superior) $T * \varphi \subseteq T + \varphi$
4. (cota inferior) Si $\neg\varphi \notin T$, entonces $T + \varphi \subseteq T * \varphi$
5. (trivialización) $T * \varphi = \mathcal{L}_p$ sii $\vdash \neg\varphi$
6. (extensionalidad $*$) Si $\vdash \varphi \leftrightarrow \psi$, entonces $T * \varphi = T * \psi$
7. (min-conjunción $*$) Si $\neg\psi \in T * (\varphi \wedge \psi)$, entonces $(T * \varphi) + \psi \subseteq T * (\varphi \wedge \psi)$
8. (max-conjunción $*$) Si $T * (\varphi \wedge \psi) \subseteq (T * \varphi) + \psi$

Definición 5.1.7. Sea $-$ un operador de contracción. Llamamos *operador de Levi* a un operador $*$ definido por $T * \varphi = (T - \neg\varphi) + \varphi$.

Proposición 5.1.8. *Un operador de Levi es un operador de revisión.*

Demostración. Se desprende directamente de aplicar las definiciones. □

Si tuviésemos una función de contracción definida, entonces tendríamos un operador de revisión. Ahora vamos a avanzar en esa dirección: sean $\varphi \in \mathcal{L}_p$ y T un conjunto de creencias, queremos entender qué forma tiene que tener $T - \varphi$. Si $\varphi \notin T$, no hay nada que decir porque el axioma de mínima acción establece cómo es la contracción con φ. Sea $\varphi \in T$, para una contracción $T - \varphi$ estamos interesados en los conjuntos $T' \subseteq T$ tales que no contienen a φ y que son resultado de remover lo mínimo posible para que no implique φ. Esto queda capturado en la siguiente definición.

Definición 5.1.9. Sean T, T' conjuntos de creencias tales que $T' \subseteq T$. Se dice que T' es *un maximal de T que no implica φ* si:

- $\varphi \notin T'$
- Si $\psi \in T$ pero $\psi \notin T'$, entonces $\psi \to \varphi \in T'$

El conjunto de todos los maximales de T que no implican φ lo notaremos $T \perp \varphi$. Además notaremos $M(T) = \{T \perp \varphi \ : \ \varphi \in \mathcal{L}_p\}$.

Los elementos T' de $T \perp \varphi$ son los candidatos a $T - \varphi$ por definición cumplen clausura $-$, contracción y éxito $-$, pero además cumplen todas las demás condiciones (para una demostración de esto consultar el paper iniciador [2], o el libro [26]). Esto nos permite construír una función de contracción.

Definición 5.1.10. Una *función de selección* es una función S tal que dado un $T\perp\varphi \in M(T)$ devuelve algunos de los conjuntos en $T\perp\varphi$ cuando este es no vacío, y en caso contrario devuelve T.

La idea detrás de esta definición es elegir la información que definitivamente se quiere conservar, y en base a esta elección construír una contracción.

Definición 5.1.11. Sea S una función de selección, *una contracción partial meet basada en* S es una función $-_{pm}$ tal que

$$T -_{pm} \varphi = \begin{cases} \bigcap_{C\in S(T\perp\varphi)} C & \text{si } T\perp\varphi \neq \emptyset \\ T & \text{en caso contrario} \end{cases}$$

Una manera de discriminar entre los distintos elementos de $T\perp\varphi$ es vía una relación transitiva $\leq$ en $M(T)$ (de esta manera no depende del φ elegido para contraer T).

Definición 5.1.12. Sea S una función de selección. Decimos que S *está inducida por* $\leq$ si para toda φ vale $S(T\perp\varphi) = \{T' \in T\perp\varphi \ : \ T'' \leq T' \text{ para todo } T'' \in T\perp\varphi\}$.

Esto da lugar al siguiente teorema.

Teorema 5.1.13. *Una función $-$ es una contracción sii es una contracción parcial basada en una función de selección S inducida por una relación $\leq$ transitiva.*

Puede interpretarse esta relación de $\leq$ como una preferencia del agente entre los distintos conjuntos de creencias posibles luego de remover una fórmula.

Las relaciones de preferencia son un leitmotiv de la Teoría de Revisión. Además de las recién vistas, las de mayor importancia son las que encontramos en los modelos de plausibilidad.

5.2. Modelos de plausibilidad

Estamos interesados en saber cómo reacciona un agente a la incorporación de información que contradice sus creencias. A tal efecto, en esta sección vamos a desarrollar la teoría de modelos de plausibilidad presentada por primera vez en [6]. Estos modelos permiten representar los "planes de contingencia" de un agente ante nueva información y las preferencias de un agente entre las distintas alternativas doxásticas. Primero vamos a estudiar el caso de agente único para luego estudiar las interacciones entre múltiples agentes.

Comenzamos por recordar algunas definiciones básicas.

Definición 5.2.1. Un *preorden en* W es una relación transitiva y reflexiva en $W \times W$.

Definición 5.2.2. Sea $\leq$ un preorden en W, decimos que es *total* si para todo $v, w \in W$ vale $v \leq w$ o $w \leq v$.

Definición 5.2.3. Sea $\leq$ un preorden en W y $P \subseteq W$, w se dice maximal de P respecto de $\leq$ si no existe $v \in P$ tal que $w \leq v$ pero no $w \geq v$.

Definición 5.2.4. Un *frame de plausibilidad* es un par $\mathfrak{F} = \langle W, \leq \rangle$ donde W es un conjunto de mundos posibles y $\leq$ es un preorden total en W tal que todo subconjunto no vacío tiene elementos maximales.

Dado un $P \subseteq W$ no vacío, notaremos $maxP$ al conjunto de los elementos maximales de P. Además vamos a notar $w' < w$ si $w' \leq w$ pero $w \not\leq w'$, y $w \cong w'$ si $w \leq w'$ y $w' \leq w$.

La lectura intuitiva que hacemos de $w' \leq w$ es "el agente considera a w al menos tan plausible como w", mientras que $w' < w$ representa que "el agente considera w es estrictamente más plausible que w'", y finalmente $w \cong w'$ puede leerse como que "el agente considera w y w' igualmente plausibles".

Definición 5.2.5. Dado un frame de plausibilidad $\mathfrak{F} = \langle W, \leq \rangle$, *una proposición-*$\mathfrak{F}$ es cualquier conjunto $P \subseteq W$.

Además decimos que *un mundo* w *satisface* P si $w \in P$ y que P es una *proposición-*W si $P \subseteq W$.

Observemos que un frame de plausibilidad es un caso particular de un frame de Kripke. La definición de modelo de plausibilidad es inmediata.

Definición 5.2.6. Un *modelo de plausibilidad* es un modelo de Kripke basado en un frame de plausibilidad. Es decir, es una terna $\mathfrak{M} = \langle W, \leq, \mathtt{i} \rangle$ en la que $\langle W, \leq \rangle$ es un frame de plausibilidad, Φ un conjunto de variables proposicionales e $\mathtt{i} : \Phi \to \mathcal{P}(W)$. Un *modelo de plausibilidad con mundo real indicado* es un par $\mathfrak{M}^* = \langle \mathfrak{M}, w^* \rangle$ donde $\mathfrak{M}$ es un modelo de plausibilidad y w^* es un mundo de $\mathfrak{M}$. Cuando no haya ambigüedad, nos referiremos a estos modelos también como modelos de plausibilidad.

Diremos que P es una *proposición-*$\mathfrak{M}$ si es una proposición-W.

Observación 5.2.7. En un modelo podemos identificar las variables proposicionales como proposiciones-$\mathfrak{M}$ vía la función $\mathtt{i}$. ¿Podemos hacer lo mismo con cualquier fórmula?

Dado un modelo de plausibilidad $\mathfrak{M} = \langle W, \leq, \mathtt{i} \rangle$ definimos operaciones booleanas para las proposiciones-$\mathfrak{M}$:

$$\neg P := W \setminus P$$

$$P \vee Q = P \cup Q.$$

Además definimos las constantes $\top_{\mathfrak{M}} := W$ y $\bot_{\mathfrak{M}} := \emptyset$. Además, toda relación $R \subseteq W \times W$, induce una modalidad de Kripke

$$[R]Q := \{w \in W \ : \text{para todo } wRv \text{ vale } v \in Q\}.$$

Con esta última incorporación vamos a lograr traducir todas las fórmulas a proposiciones-$\mathfrak{M}$. La motivación de esto es establecer un puente entre el lenguaje y la semántica que simplifica ciertas demostraciones técnicas. La mayoría de ellas no serán presentadas en esta tesis, sin embargo se decidió conservar este enfoque para el contenido técnico de este capítulo con el objetivo de facilitarle al lector interesado la consulta de bibliografía.

Por otro lado, para ejemplos introductorios no vamos a utilizar proposiciones-$\mathfrak{M}$ sino fórmulas por su caracter intuitivo.

Hasta ahora nuestra noción de proposición es local: sólo tenemos proposiciones-$\mathfrak{M}$ para cada modelo $\mathfrak{M}$. Pero estamos interesados en estudiar los cambios de modelos por lo tanto necesitamos una noción de proposición que no esté asociada a un modelo en particular. Esto motiva la siguiente definición.

Definición 5.2.8. Una *proposición doxástica* es una función $\mathbf{P}$ que asigna a cada modelo de plausibilidad $\mathfrak{M}$ una proposición-$\mathfrak{M}$ $\mathbf{P}_{\mathfrak{M}}$. Diremos que $\mathbf{P}$ es verdadera en w si $w \in \mathbf{P}_{\mathfrak{M}}$. Notaremos $w \Vdash_{\mathfrak{M}} \mathbf{P}$, y cuando el modelo se deduce del contexto $w \Vdash \mathbf{P}$.

Denotamos `Prop` el conjunto de proposiciones doxásticas, y definimos las operaciones booleanas en `Prop` de manera estándar: $(\mathbf{P} \wedge \mathbf{Q})_{\mathfrak{M}} := \mathbf{P}_{\mathfrak{M}} \cap \mathbf{Q}_{\mathfrak{M}}$, $(\neg \mathbf{P})_{\mathfrak{M}} := W \setminus \mathbf{P}_{\mathfrak{M}}$ y las operaciones definidas en base a estos. También tenemos las constantes $\top$ y $\bot$ definidas según las constantes $\top_{\mathfrak{M}}$ y $\bot_{\mathfrak{M}}$ de cada modelo. Y similarmente, se pueden definir las modalidades asociadas a una relación R: $([R]\mathbf{P})_{\mathfrak{M}} = [R]\mathbf{P}_{\mathfrak{M}}$.

Además, definimos consecuencia semántica del siguiente modo: $\mathbf{P} \Vdash \mathbf{Q}$ sii $\mathbf{P}_{\mathfrak{M}} \subseteq \mathbf{Q}_{\mathfrak{M}}$ para todo modelo $\mathfrak{M}$. También notaremos esto como $\Vdash \mathbf{P} \to \mathbf{Q}$.

Ahora introducimos una relación de accesibilidad doxástica $\to$ para estos modelos definida por: Ahora introducimos una relación de accesibilidad doxástica $\to$ para estos modelos definida por:

$$w \to v \text{ sii } v \in maxW$$

Es decir, en un mundo w el agente cree que puede estar cualquiera de los mundos en $maxW$. Esto intuitivamente representa el hecho de que el agente cree estar en alguno de los mundos que maximizan la plausibilidad.

Pero además podemos definir una relación de accesibilidad doxástica condicional:

$$w \to^P v \text{ sii } v \in maxP$$

Y una lectura intuitiva de esto "si en un mundo w el agente obtiene la información de que P es verdadero creerá que está en alguno de los mundos que maximizan la plausibilidad cumpliendo P".

Finalmente, definimos la relación de posibilidad epistémica como la relación universal:

$$w \sim v \text{ sii } w, v \in W.$$

Es decir, todos los estados en W se asumen posibles. La única manera de que un agente sepa algo es que sea verdadero en todos los mundos del modelo.

En este contexto, podemos introducir los operadores de conocimiento, creencia y creencia condicional como:

$$KP := [\sim]P$$
$$BP := [\rightarrow]P$$
$$B^Q P := [\rightarrow^Q]P$$

Leemos KP como "el agente sabe P". Este conocimiento responde perfectamente a la definición de conocimiento de Leibnitz: algo es sabido, si es verdadero en todos los mundos. Similarmente leemos BP como "el agente cree P".

Pero la diferencia respecto de modelos anteriores es aportada por las fórmulas $B^Q P$, que leeremos como "Si el agente recibe la información Q, creerá que P era verdadero (antes de recibir la información)". A ese operador binario lo llamaremos *operador de creencia condicional*.

Es decir que tenemos en el operador de creencia condicional una operación de revisión que captura los potenciales cambios de creencias del agente en caso de recibir nueva información. Y a diferencia del operador de revisión propuesto por la teoría AGM, éste permite revisar con cualquier tipo de fórmulas. En particular, fórmulas de órdenes superiores.

Ejemplo 5.2.9. Juan Pérez siente que es un genio. Él sabe que hay sólo dos explicaciones posibles para ello: es un genio o está borracho. Él no se siente borracho, así que cree que es un genio (sobrio). Sin embargo, si llegase a enterarse que está borracho, va a pensar que su sentimiento de genialidad se debía a los efectos del alcohol. Es decir: si se entera que está borracho, creerá que es sólo un borracho no-genio.

Pero en realidad es un genio y está borracho.

El modelo de plausibilidad queda representado en el siguiente gráfico, donde los rombos representan los mundos posibles, las flechas reprensentan el preorden, el rombo pintado representa el mundo interior y en el interior de los nodos se encuentran las variables proposicionales válidas en él. En este caso b representa la variable proposicional para "Juan está borracho" y g para "Juan es un genio".

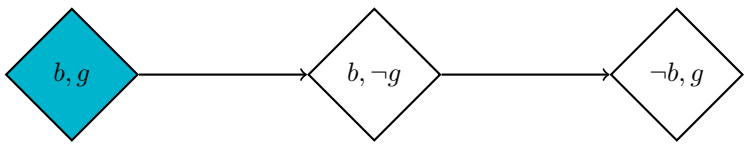


Figura 5.1: En los diagramas de modelos de plausibilidad representamos el preorden con flechas, y omitimos las flechas reflexivas y transitivas para mayor claridad visual.

En el diagrama no representamos el mundo $\neg b, \neg g$ porque Juan sabe (K) que está borracho o es un genio $(b \vee g)$.

Juan Pérez no sabe (K) que es un genio, porque hay mundos posibles en los que no lo es: $(b, \neg g)$. Sin embargo el cree que es un genio, y esto es verdadero.

En este ejemplo vemos que este modelo respeta la diferencia entre conocimiento y creencia verdadera, discutido por Sócrates y citado por Platón en Menón.

Múltiples agentes

Podemos generalizar las nociones de frame y modelo de plausibilidad a múltiples agentes teniendo unas precauciones. Las discutimos brevemente para motivar la definición posterior.

En el caso de un único agente los estados que el agente sabe imposibles no forman parte del modelo, esto queda en evidencia en el hecho de que la relación de indistinguibilidad epistémica es la relación universal que considera a cualquier mundo del modelo como una alternativa epistémica de cualquier otro.

Sin embargo en el caso multiagente no podemos excluír del modelo los estados que un agente i sabe imposibles, porque podría suceder que para otro agente j sean perfectamente plausibles. Más aún: a pesar de que el agente i considere imposible un mundo no podemos descartarlo siquiera de su estructura de plausibilidad, porque podría ser de importancia para las creencias de otros agentes acerca de las creencias de i. Entonces es necesario definir relaciones $\sim_i$ de indistinguibilidad, con la misma motivación que en el Capítulo 3. Por lo tanto a nuestros frames tendríamos que definirlos como ternas $\langle W, (\leq_i)_{i\in N}, (\sim_i)_{i\in N}\rangle$ donde $\sim_i$ y $\leq_i$ cumplan algunas condiciones.

Para empezar, si dos estados son distinguibles por un agente ($w \not\sim_i v$) significa que el agente i no los considera posibles simultáneamente. Por lo tanto, no tiene sentido preguntarse cuál es más plausible de ambos, ya que pertenecen a realidades disjuntas para el agente i. Esto significa que deberíamos tener que $w \leq_i v$ implica $w \sim_i v$. Pero además queremos que el modelo restringido a alguna clase de equivalencia de $\sim_i$ y considerando sólo el agente i, sea como un modelo de plausibilidad de un sólo agente. Esto significa que queremos que $\leq_i$ tenga maximales en los subconjuntos no vacíos de la clase de equivalencia.

Definición 5.2.10. Dado un conjunto finito de agentes N, un *frame de plausibilidad para* N es una terna $\mathfrak{F} = \langle W, (\leq_i)_{i\in N}, (\sim_i)_{i\in N}\rangle$, tal que para todo $i \in N$ vale que $\leq_i$ es un preorden (no necesariamente total) y para cada w todo subconjunto no vacío de $\{v : w \sim_i v\}$ tiene maximales. Si P es uno de tales subconjuntos, al conjunto de elementos maximales los notamos $max_i P$.

Llamamos $w(i) := \{v : w \leq_i v \text{ o } v \leq_i w\}$ *clase de comparabilidad de* w *para* i *o celda de información de* w *para* i.

Observación 5.2.11. Es fácil ver que bajo estas restricciones $\sim_i = \leq_i \cup \geq_i$. Por lo tanto en adelante vamos a omitir a las relaciones de indistinguibilidad de la notación al momento de definir los frames de plausibilidad.

Definición 5.2.12. Dado un conjunto finito de agentes N, un *modelo de plausibilidad para* N es una terna $\mathfrak{M} = \langle W, (\leq_i)_{i\in N}, \mathtt{i}\rangle$, donde $\langle W, (\leq_i)_{i\in N}\rangle$ es un frame de plausibilidad para N e $\mathtt{i} : \Phi \to \mathcal{P}(W)$, donde Φ es el conjunto de variables proposicionales del lenguaje.

La definición de la relación de accesibilidad doxástica para cada agente es la misma que antes:

$$w \to_i v \text{ sii } v \in max_i w(i)$$

y la relación de accesibilidad doxástica condicional:

$$w \to_i^P v \text{ sii } v \in max_i\,(w(i) \cap P)$$

De este modo, estamos en condiciones de definir las modalidades de interés de manera análoga al caso de un único agente:

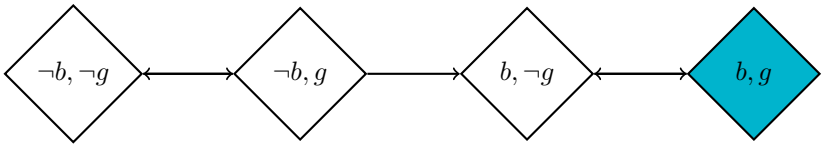
$$K_i P := [\sim_i] P$$
$$B_i P := [\to_i] P$$
$$B_i^Q P := [\to_i^Q] P$$

Y dado que tenemos múltiples agentes también incorporamos el operador C de conocimiento común con su semántica tradicional:

$$CP := \left[\left(\bigcup_{i\in N} \sim_i\right)^*\right] P$$

Observación 5.2.13. Ahora, a diferencia del caso de un único agente, que un agente sepa algo depende del mundo en el que nos paremos porque la cuantificación no es sobre todos los mundos sino únicamente sobre su celda de información.

Ejemplo 5.2.14. Agregamos un agente más a la historia de Juan Pérez del Ejemplo 5.2.9: su mejor amigo Carlos Fernández. Carlos está bastante seguro de que Juan está borracho: lo conoce hace bastante tiempo, aunque tiene en cuenta la posibilidad de estar equivocado. Sin embargo no tiene idea de si Juan es un genio o no, considera ambas cosas igualmente plausibles. El modelo para Carlos solo sería:



Para poner el orden de Juan junto con el orden Carlos necesitamos saber qué saben el uno del otro.

Supongamos que todas las premisas que asumimos acerca de Juan y acerca de Carlos son conocimiento común, salvo cuál es el mundo real y que Juan se siente un genio. Más precisamente todas las opiniones (conocimientos, creencias y creencias condicionales) de Carlos son conocimiento común, y también lo es que:

- Si Juan está borracho, sentirá que es un genio
- Si Juan es genio, sentirá que es un genio
- No hay otros motivos por los cuales Juan se pueda sentir genio
- Juan sabe lo que siente (¡acerca de ser genio o no!)
- Si Juan se entera que está borracho, creerá que no es un genio
- Juan cree que es un genio pero que no está borracho.

Entonces el modelo multiagente de plausibilidad está dado por el siguiente diagrama, donde las flechas etiquetadas con j es el preorden de Juan y las etiquetadas con c el preorden de Carlos.

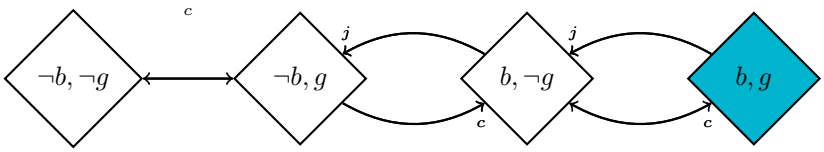


Figura 5.2: Modelo multiagente de plausibilidad

Ahora relajemos un poco el contexto: Supongamos que las opiniones de Carlos no son conocimiento común y mantengamos todas las demás hipótesis. Además ahora es conocimiento común que Carlos no tiene opinión acerca de si Juan es genio o no (que considera ambas opciones igualmente plausibles), pero no es conocimiento común que tiene una opinión acerca de si está borracho o no (cree que está borracho o cree que no lo está, pero su opinión es desconocida para Juan que considera ambas opiniones de Carlos igualmente plausibles). Este es el modelo resultante:

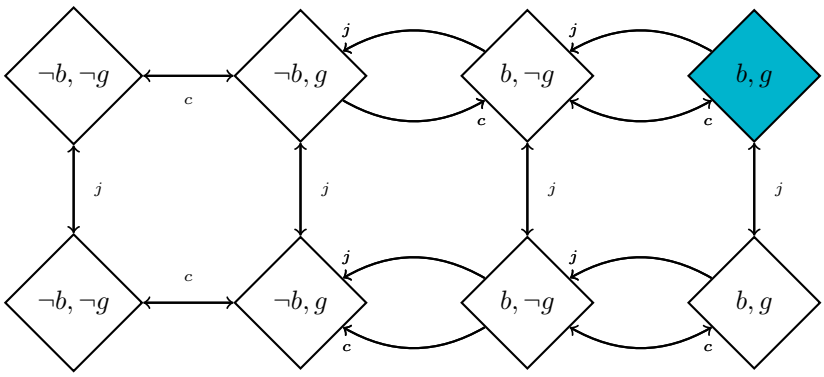


Figura 5.3: Otro modelo multiagente de plausibilidad

Interacción de conocimiento, creencia y revisión

Es comúnmente aceptado por epistemólogos que el ingrediente faltante a la definición de conocimiento como creencia verdadera justificada dada Platón en Menón (que fue refutada por Gettier en [16]) es la propiedad de estabilidad bajo revisiones con información verdadera: el conocimiento no puede ser refutable por ninguna información real que no se posea. Esta extensión de la definición platónica es propuesta por numerosos autores, entre ellos Hintikka en [18] y Stalnaker en [25], y ampliamente aceptada por la comunidad filosófica.

Sin embargo, a continuación veremos una caracterización del conocimiento representado por K como una "creencia absoluta", invariante bajo cualquier revisión (sin importar si se revisa con información verdadera o no). Es decir, si sigue siendo creída bajo cualquier condición.

Proposición 5.2.15. *Para todo modelo $\mathfrak{M}$ y mundo w vale: $w \Vdash K_i Q$ sii $w \Vdash \bigwedge_{P \in Prop} B_i^P Q$*

Demostración. Se desprende directamente de las definiciones. □

En este sentido, la noción de conocimiento capturada (con esta semántica) por K_i es un poco más fuerte que la propuesta por Stalnaker y Hintikka. A pesar de esto no deja de ser una noción útil ya que cumple **S5**, axiomas aceptados para modelar el conocimiento, principalmente en aplicaciones para Ciencias de la Computación.

Proposición 5.2.16. *En la clase de modelos de plausibilidad valen los axiomas de **S5** para los operadores K_i.*

En adelante vamos a referirnos a K_i como *conocimiento certero*, y leeremos $K_i P$ como "el agente sabe certeramente P".

Ahora vamos a introducir nuevas modalidades $\Box_i$ para capturar una forma conocimiento más débil, con la siguiente semántica:

$$\Box_i P := [\leq_i]P = \{w \in W \ : \ v \in P \text{ para todo } w \leq_i v\}.$$

Nos referiremos a $\Box_i$ como *conocimiento irrefutable* y leeremos $\Box_i \mathbf{P}$ como "el agente i sabe P" o "el agente i sabe P irrefutablemente".

Por supuesto, el operador $\Box$ induce un operador análogo en las proposiciones doxásticas de la manera estándar: $(\Box_i \mathbf{P})_{\mathfrak{M}} := \Box_i \mathbf{P}_{\mathfrak{M}}$.

Observación 5.2.17. Dado que $\leq_i$ es reflexiva y transitiva tenemos que $\Box_i$ valida los axiomas de **S4**: cumple veracidad e introspección positiva, pero no necesariamente es negativamente introspectivo.

¿Cómo se relacionan, en este contexto, K_i, $\Box_i$ y B_i?

Proposición 5.2.18. *Valen las siguientes:*

$$\Vdash K_i \boldsymbol{P} \rightarrow \Box_i \boldsymbol{P}$$

$$\Vdash \Box_i \boldsymbol{P} \rightarrow B_i \boldsymbol{P}$$

Entonces, efectivamente, $\Box_i$ representa una noción de conocimiento más débil que K_i (pero que sigue siendo más fuerte que B_i). Pero esto no es una caracterización. El siguiente resultado caracteriza un poco mejor qué clase de conocimiento captura $\Box_i$.

Proposición 5.2.19. *El conocimiento representado por $\Box_i$ resiste revisiones con cualquier información verdadera. Esto significa que:*

$$w \Vdash \Box_i \boldsymbol{Q} \textit{ es equivalente a}$$

$$w \Vdash B_i^{\boldsymbol{P}} \boldsymbol{Q} \textit{ para todo } \boldsymbol{P} \textit{ tal que } w \Vdash \boldsymbol{P}$$

Esta proposición afirma que $\Box_i$ representa el conocimiento definido por Hintikka: es conocimiento irrefutable por información verdadera. Pero además de representar esta noción de conocimiento filosóficamente útil, la modalidad $\Box_i$ cumple otras propiedades interesantes:

Proposición 5.2.20. *Son válidas:*

1. $K_i \Box_i \boldsymbol{P}$ *es equivalente a* $K_i \boldsymbol{P}$
2. $B_i \boldsymbol{P}$ *es equivalente a* $B_i \Box_i \boldsymbol{P}$

El ítem 1 asevera que saber certeramente que se sabe algo es lo mismo que saberlo certeramente, mientras que el ítem 2 afirma que creer algo es creer que se sabe. Esta última propiedad es deseable en los sistemas epistémicos-doxásticos en muchos contextos y, como vimos en el Capítulo 3, no es simple de incorporar para la noción de conocimiento certero ya que conlleva al colapso de conocimiento certero y creencia.

De las definiciones se desprende el siguiente resultado, que nos premite prescindir de la modalidad de creencia condicional para tomarla como una abreviación:

Proposición 5.2.21. $B_i^{\boldsymbol{P}} \boldsymbol{Q}$ *es equivalente a* $\neg K_i \neg \boldsymbol{P} \rightarrow \neg K_i \neg (\boldsymbol{P} \wedge \Box_i (\boldsymbol{P} \rightarrow \boldsymbol{Q}))$

Finalmente vamos a incorporar un último operador Sb_i que es una variación de la noción doxástica representada por B_i:

$$Sb_i \mathbf{P} := B_i \mathbf{P} \wedge K_i (\mathbf{P} \rightarrow \Box_i \mathbf{P})$$

que en términos del orden de plausibilidad establece que todos los mundos en los que vale $\mathbf{P}$ son más plausibles que todos los mundos en los que vale $\neg \mathbf{P}$.

Llamamos a Sb_i *modalidad de creencia fuerte para el agente* i y leemos $Sb_i\mathbf{P}$ como "el agente i cree fuertemente $\mathbf{P}$".

De la definición se desprende de manera automática que Sb_i representa una forma de creencia más fuerte que B_i. Pero no sólo eso, sino que además algo es fuertemente creído sii sigue siendo creído dada cualquier nueva información que no lo contradiga. Esto queda formalizado en la siguiente Proposición:

Proposición 5.2.22. *Son equivalentes:*

$$w \Vdash Sb_i \boldsymbol{Q},\ y$$

$$w \Vdash B_i^{\boldsymbol{P}} \boldsymbol{Q} \text{ para todo } \boldsymbol{P} \text{ tal que } w \not\Vdash K_i(\boldsymbol{P} \to \neg \boldsymbol{Q})$$

Quizás el más conocido ejemplo de creencia fuerte se encuentra en el derecho, la justicia cree fuertemente en la inocencia de los acusados: "inocente hasta que se demuestre lo contrario".

Ejemplo 5.2.23. En el Ejemplo 5.2.14, Juan Pérez cree fuertemente que $\neg b$, y además cree fuertemente que $g \wedge \neg b$. Sin embargo, no cree fuertemente g. Es decir:

$$Sb(\varphi \wedge \psi) \not\to Sb\varphi \wedge Sb\psi$$

Esto demuestra que a diferencia del conocimiento y las creencias simples, las creencias fuertes no son cerradas por inferencia lógica.

Se puede definir una lógica con operadores de conocimiento certero y conocimiento irrefutable, que es completa y decidible respecto de los modelos de plausibilidad.

Definición 5.2.24. Sea N un conjunto finito de agentes y Φ un conjunto finito de variables proposicionales. Definimos el lenguaje $\mathcal{L}_{K\Box}(N, \Phi)$ definido recursivamente por:

$$\varphi ::= p \mid \bot \mid \neg\varphi \mid \psi \wedge \varphi \mid K_i\varphi \mid \Box_i\varphi$$

donde p es alguna variable proposicional (i.e.: $p \in \Phi$) e $i \in N$.

Además definimos las abreviaciones $B_i^{\varphi}\psi := \neg K_i\neg\varphi \to \neg K_i\neg(\varphi \wedge \Box_i(\varphi \to \psi))$ (inspirada por la Proposición 5.2.21) y $B_i\varphi := B_i^{\top}\varphi$ con $\top := \neg(p \wedge \neg p)$.

Le damos a los operadores K_i y $\Box_i$ la semántica de cuantificadores modales universales para $\sim_i$ y para $\leq_i$ respectivamente.

Definición 5.2.25. Definimos el sistema axiomático $\mathbf{K}\Box$ dado por una axiomatización de la lógica proposicional junto con:

- Axiomas de **S5** para cada K_i
- (introspección positiva para $\Box_i$) $\Box_i P \to \Box_i\Box_i P$

- (veracidad para $\Box_i$) $\Box_i P \to P$
- $K_i P \to \Box_i P$
- $K_i(P \vee \Box_i Q) \wedge K_i(Q \vee \Box_i P) \to K_i P \vee K_i Q$

y reglas de necesitación para K_i y $\Box_i$.

La correctitud del sistema es rutina, y sin escalas nos dirigimos al resultado de completitud y decibilidad.

Teorema 5.2.26. *La lógica $\boldsymbol{K}\Box$ es completa respecto de los modelos de plausibilidad. Además es decidible y tiene la propiedad de modelo finito.*

Idea de la demostración. Cada modelo de plausibilidad $\mathfrak{M} = \langle W, (\leq_i)_{i\in N}, \mathtt{i}\rangle$ induce un modelo de Kripke, $\hat{\mathfrak{M}} = \langle W, (\leq_i)_{i\in N}, (\sim_i)_{i\in N}, \mathtt{i}\rangle$ donde $\leq_i$ es un preorden en el que todo subconjunto no vacío de su dominio tiene elementos maximales, $\sim_i$ es de equivalencia y vale que $\leq_i \subseteq \sim_i$ y la restricción de $\leq_i$ a $\sim_i$ es un preorden total sin cadenas. Análogamente si tenemos un modelo de Kripke que cumple estas condiciones, induce un modelo de plausibilidad.

Teniendo esto en consideración la demostración de este teorema es una réplica de la demostración del Teorema 3.0.22 y el Corolario 3.0.35.

□

5.2.1. Revisión dinámica

En la sección anterior presentamos el operador de creencia condicional para estudiar revisiones de creencias. Pero la revisión capturada por este operador es estática, de carácter hipotético. Uno puede querer revisar las creencias por dos motivos: cambió la información que se tiene (por ejemplo: un agente se enteró de algo que descarta un mundo posible y ejecuta un plan de contingencia), o cambiaron las plausibilidades (por ejemplo: un agente cambió de opinión acerca de cuáles son los mundos más plausibles). La diferencia entre estos tipos de cambios es que la segunda implica un cambio en la situación mucho mayor: no sólo cambia lo que el agente cree, sino que también cambian sus planes de contingencia.

El operador de creencia condicional no revisa las creencias sobre la situación después de la revisión (con los planes de contingencia después de revisar), sino las de la situación antes de la revisión (con los planes de contingencia antes de revisar). Esto significa que si esos planes de contingencia cambiaron al recibir cierta información, este tipo de revisión no será efectiva.

Desde un punto de vista semántico, una revisión de creencias de órdenes superiores es, de algún modo, revisar la estructura relacional: cambiar la relación de plausibilidad y/o su dominio, y esto no es capturado de manera efectiva por el operador de creencia condicional ya que opera sobre una estructura de plausibilidad no revisada.

Para ejemplificar, consideremos el modelo de Juan Pérez del Ejemplo 5.2.9. Llamemos $\varphi := b \wedge \neg Bb$ (llamamos a este tipo de fórmulas, *fórmulas de Moore*). Se puede verificar que vale $B^{\varphi}\neg Bb$. ¿Contradice esto el hecho de que la información incorporada anuncia que b es verdadero? No, porque las creencias condicionales son la revisión de las creencias en el modelo **antes de que sea anunciada** φ, y en ese modelo el agente aún no creía que estaba borracho.

La aparente contradicción proviene de que el anuncio de φ conlleva necesariamente a un cambio de su valor de verdad por un cambio en la estructura de plausibilidad del agente.

Lo que le falta capturar a los modelos de plausibilidad es la noción de acción o evento que vimos en el Capítulo 4 que encierra el aspecto dinámico que le falta a las creencias condicionales. Esto motiva la siguiente definición.

Definición 5.2.27. Un *modelo de plausibilidad de eventos o modelo de eventos plausibles* es una terna $\mathbf{\Sigma} = \langle \Sigma, (\leq_i)_{i\in N}, pre\rangle$ donde $\langle \Sigma, (\leq_i)_{i\in N}\rangle$ es un frame de plausibilidad y $pre : \Sigma \to \texttt{Prop}$ es un mapa de precondiciones que asigna a cada $\sigma \in \Sigma$ una proposición doxástica pre_σ.

Llamaremos *eventos* a los elementos $\sigma \in \Sigma$ y *precondición de* σ a pre_σ.

Definición 5.2.28. Un *programa doxástico sobre un modelo de plausibilidad de eventos* $\mathbf{\Sigma}$ *o programa-*$\mathbf{\Sigma}$ es un subconjunto $\Gamma \subseteq \Sigma$.

Podemos pensar los programas-$\mathbf{\Sigma}$ como eventos no determinísticos donde se ejecuta uno de los eventos $\gamma \in \Gamma$.

Observación 5.2.29. Los programas-Σ son los análogos dinámicos de las proposiciones-$\mathfrak{M}$

Ahora que ya tenemos modelos de plausibilidad de eventos, quisiéramos ver cómo actúan sobre los modelos de plausibilidad.

Definición 5.2.30. Sean $\mathfrak{M} = \langle W, (\leq_i)_{i\in N}, \texttt{i}\rangle$ un modelo de plausibilidad multiagente y $\mathbf{\Sigma} = \langle \Sigma, (\leq_i)_{i\in N}, pre\rangle$, un modelo de plausibilidad de eventos. Definimos el *producto actualización preferencial de* $\mathfrak{M}$ *y* $\mathbf{\Sigma}$ como $\mathfrak{M} \otimes \mathbf{\Sigma} = \langle W \otimes \Sigma, \leq_i, \hat{\texttt{i}}\rangle$ donde:

- $W \otimes \Sigma = \{(w, \sigma) \ : \ w \Vdash pre_\sigma\}$
- $\hat{\texttt{i}}(p) = \texttt{i}(p)$
- $(w, \sigma) \leq_i (w', \sigma')$ sii $\sigma <_i \sigma'$ y $w \sim_i w'$ o $\sigma \cong_i \sigma'$ y $w \leq_i w'$.

Observación 5.2.31. Del producto actualización preferencial resulta un modelo de plausibilidad.

Ejemplo 5.2.32. Supongamos que tenemos el siguiente modelo de plausibilidades:

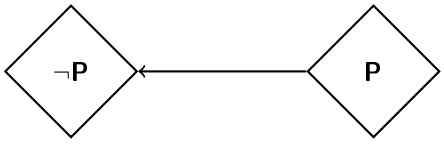


Figura 5.4: Es más plausible un mundo en el que vale $\neg\mathbf{P}$ que uno en el que vale $\mathbf{P}$.

Supongamos que sucede un evento de un modelo de plausibilidad de eventos $\mathbf{\Sigma} = \langle \Sigma, \leq, pre \rangle$ donde sólo hay 2 eventos σ y σ' con precondiciones $\neg\mathbf{P}$ y $\mathbf{P}$ respectivamente tales que $\sigma < \sigma'$. Representamos este modelo en el siguiente diagrama, donde las elipses representan eventos, su contenido las precondiciones para ejecutarlo y las flechas la relación de plausibilidad:

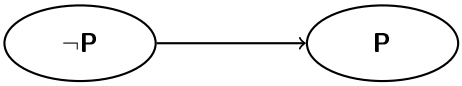


Figura 5.5: El agente considera más plausible que suceda σ' (cuya precondición es $\mathbf{P}$) a que suceda σ (cuya precondición es $\neg\mathbf{P}$).

Uno espera que si un agente piensa que es más plausible un evento que otro (y suceda uno de ellos) entonces considere más plausible la precondición del evento más plausible que la del otro. Esa es la motivación detrás de la manera en la que queda definida $\leq_i$ para el producto actualización preferencial. El producto actualización preferencial es estos dos modelos resulta en:

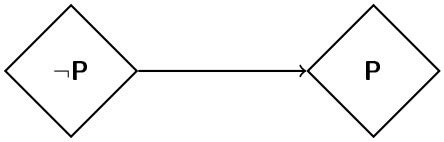


Figura 5.6: Luego de actualizar el agente considera más plausible un mundo en el que vale $\mathbf{P}$ que uno en el que vale $\neg\mathbf{P}$.

Observación 5.2.33. Con esta definición de producto actualización preferencial tenemos como caso particular el producto actualización definido en el Capítulo 4.

Definición 5.2.34. Una *actualización* es una transformación de modelos α tal que a un modelo de plausibilidad $\mathfrak{M} = \langle W, (\leq_i)_{i \in N}, \mathtt{i} \rangle$ le asigna un nuevo modelo de plausibilidad $\alpha(\mathfrak{M}) = \langle W', \leq', \mathtt{i}' \rangle$ donde $W' \subseteq W$, $\mathtt{i}' := \mathtt{i}|_{W'}$.

En este contexto llamaremos dominio de α a W' y lo notamos $Dom_{\mathfrak{M}}(\alpha)$. Son los estados en los que puede ser ejecutado la actualización.

Del mismo modo que lo hicimos con las acciones en el capítulo anterior, podemos agregar al lenguaje un operador dinámico $[\alpha]\mathbf{P}$ para cada actualización α, para expresar que $\mathbf{P}$ va a ser verdadera en el nuevo modelo después ejecutar el actualización α. Esta sería la semántica:

$$w \Vdash_{\mathfrak{M}} [\alpha]\mathbf{P} \text{ sii } w \in Dom_{\mathfrak{M}}(\alpha) \text{ implica } w \Vdash_{\alpha(\mathfrak{M})} \mathbf{P}$$

Vamos a definir tres actualizaciones que capturan distintas políticas de revisión.

Definición 5.2.35. Definimos la operación $!\mathbf{P}$ de *actualización infalible con* $\boldsymbol{P}$ que toma un modelo de plausibilidad $\mathfrak{M} = \langle W, (\leq_i)_{i\in N}, \mathtt{i}\rangle$ y devuelve un modelo restringido sin los estados en los que no vale $\mathbf{P}$: $\mathfrak{M}' = \langle \mathbf{P}_{\mathfrak{M}}, (\leq_i)_{i\in N}, \mathtt{i}\rangle$.

Este tipo de actualización se corresponde con anuncios públicos provenientes de fuentes de información infalibles, que jamás anuncian algo erróneo. Por lo tanto, en caso de que haya un mundo real indicado, para poder ejecutar una actualización infalible con $\mathbf{P}$, tiene que ser verdadera en el mundo real.

Esto representa conceptualmente lo mismo que el anuncio público definido en el Capítulo 4: un anuncio público que todos los agentes oyen, y todos los agentes saben que todos los agentes oyen, y todos los agentes saben que todos los agentes oyen, etcétera. Semánticamente se corresponde (al igual que antes) con borrar todos los mundos en los que no vale lo que se anuncia.

Esta operación tiene un modelo de plausibilidad de eventos que la representa que está dada por $!\mathbf{P} := \langle \Sigma, (\leq_i)_{i\in N}, pre\rangle$ con $\Sigma = \{\sigma\}$, y $pre_\sigma = \mathbf{P}$.

Ejemplo 5.2.36. Retomemos el ejemplo de Juan Pérez, cuyo modelo inicial está diagramado en la Figura 5.1. Supongamos que una fuente infalible (su esposa) le dice "estás borracho pero no lo crees". Esto induce una actualización $!(b \wedge \neg Bb)$, del que resulta este modelo:

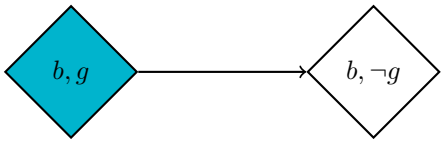


Figura 5.7: Resultado de una actualización infalible con una fórmula de Moore.

Vemos que se logra modelar correctamente el cambio de creencias al revisar con una fórmula de Moore, a diferencia de cuando intentamos hacerlo usando creencias condicionales en la introducción de esta sección.

Las fórmulas Moore son ejemplos interesantes de por qué no necesariamente tiene que valer la condición de Éxito pedida por los operadores de revisión de AGM.

Definición 5.2.37. Definimos la operación $\Uparrow \mathbf{P}$ de *actualización radical* como aquella que transforma un modelo de plausibilidad $\mathfrak{M} = \langle W, (\leq_i)_{i\in N}, \mathtt{i}\rangle$ en otro modelo $\mathfrak{M}' = \langle W, (\leq'_i)_{i\in N}, \mathtt{i}\rangle$ donde $\leq'_i$ está dada por convertir todos los mundos de $\mathfrak{M}$ en los que vale $\mathbf{P}$ en estrictamente más plausibles que todos los mundos de $\mathfrak{M}$ en los que no, y conservar el orden en el resto. Es decir:

$$\text{Si } v \Vdash \mathbf{P} \text{ y } w \Vdash \neg\mathbf{P} \text{ entonces } w \leq'_i v$$

$$\text{En caso contrario } \leq'_i = \leq_i$$

Las actualizaciones radicales modelan anuncios públicos provinientes de fuentes de información altamente confiables (o al menos persuasivas), pero falibles.

El modelo de plausibilidad de eventos para una actualización radical consiste en la terna $\Uparrow \mathbf{P} = \langle \Sigma, (\leq_i)_{i \in N}, pre \rangle$ donde $\Sigma = \{\sigma, \sigma'\}$, $\sigma' \leq_i \sigma$, $pre_\sigma = \mathbf{P}$ y $pre_{\sigma'} = \neg\mathbf{P}$.

Queda representado en el siguiente diagrama:

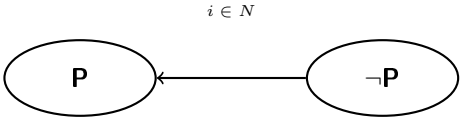


Figura 5.8: Actualización radical en un modelo de plausibilidad.

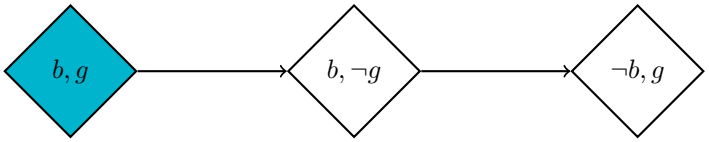
Finalmente, vamos a definir una última política de revisión más débil.

Definición 5.2.38. La operación $\uparrow \mathbf{P}$ de *actualización conservadora con* $\boldsymbol{P}$ como aquella que redefine los órdenes $\leq_i$ colocando al más plausible (para el agente i) de los mundos en los que vale $\mathbf{P}$ como máximo, y en el resto preserva el orden anterior.

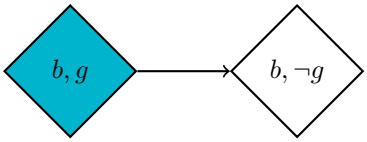
Esta operación modela el efecto producido por un anuncio público de una fuente ligeramente confiable y falible. La información es adoptada pero puede ser renunciada con facilidad.

Ejemplo 5.2.39. Retomando el Ejemplo 5.2.9 de Juan Pérez: supongamos que su vecino le dice que está borracho.

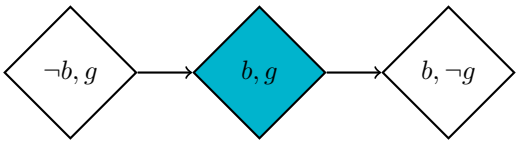
Depende de cuánto confíe Juan Pérez en su vecino va a actualizar su estado epistémico de distintos modos.



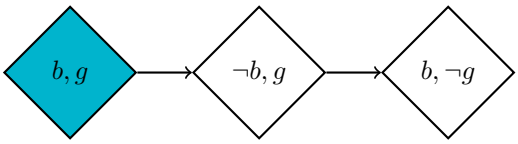
(a) Estado inicial de Juan Pérez



(b) Si su vecino es infalible

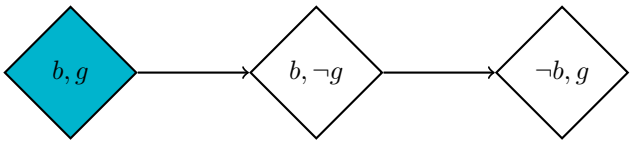


(c) Si su vecino es falible pero confía plenamente en su opinión

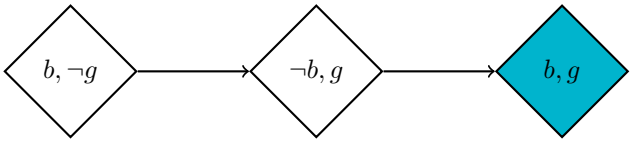


(d) Si confía en su vecino, pero no plenamente

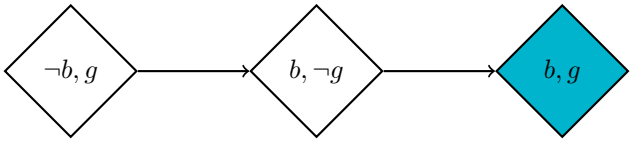
Ejemplo 5.2.40. Supongamos que Juan Pérez confía plenamente en su vecino. Él le dice primero que es un genio y está borracho, y luego que está un borracho. Tenemos los siguientes modelos:



(a) Estado inicial de Juan Pérez



(b) Luego de $\Uparrow (b \wedge g)$



(c) Luego de $\Uparrow b$

Obsérvese que este último modelo no puede ser obtenido del modelo inicial con una única actualización radical. Esta situación nos demuestra que, si bien las actualizaciones infalibles son cerradas por composición, las actualizaciones radicales no.

5.2.2. Actualizaciones iteradas

Es natural preguntarse qué sucede con la concatenación de actualizaciones: ¿Ante una secuencia infinita de actualizaciones arbitrarias el conocimiento se estabilizará? ¿Las crencias? ¿Las creencias fuertes? ¿Se converge al mundo real? En esta sección formalicemos un poco estos conceptos, que comenzaron a ser estudiados recientemente en [7].

Definición 5.2.41. Sea un modelo con mundo real indicado $\langle \mathfrak{M}, w^* \rangle$ y una actualización conservadora, radical o infalible con información $\mathbf{P}$. La actualización se dice *veraz respecto del modelo* si $w^* \Vdash \mathbf{P}$.

Definición 5.2.42. Llamaremos iteración de actualizaciones a una secuencia infinita de actualizaciones $\alpha_0, \alpha_1, \ldots, \alpha_n, \ldots$.

Un modelo finito $\mathfrak{M}_0 = \langle W_0, \leq_0, \mathtt{i}_0, w_0^* \rangle$ y una iteración de actualizaciones inducen una sucesión infinita de modelos $\mathfrak{M}_0, \mathfrak{M}_1, \ldots, \mathfrak{M}_n, \ldots$ definida por $\mathfrak{M}_{k+1} = \alpha_k(\mathfrak{M})$.

Definición 5.2.43. Diremos que una iteración de actualizaciones se estabiliza para el modelo $\mathfrak{M}$, si existe un n tal que $\mathfrak{M}_m = \mathfrak{M}_n$ para todo $m \geq n$.

Proposición 5.2.44. *Toda iteración de actualizaciones infalibles se estabiliza para todo modelo finito.*

Esto sucede simplemente porque en cada actualización pueden descartarse mundos pero no agregarse.

Pero no sucede lo mismo para iteraciones de actualizaciones arbitrarias. Por ejemplo, es claro que la iteración:

$$\Uparrow p, \Uparrow \neg p, \Uparrow p, \Uparrow \neg p, \ldots$$

no se estabiliza. Pero esto está dentro de lo correcto, ya que las revisiones que se hacen son mutuamente inconsistentes, por lo tanto no representa aprendizaje real.

Sin embargo, puede suceder que aunque las actualizaciones sean siempre veraces, una iteración nunca se estabilice. Peor aún: esto también sucede aún cuando revisamos con la misma información una y otra vez.

Ejemplo 5.2.45. En las elecciones de un país hay dos partidos 1 y 2. Un encuestador cree que un votante dado va a votar al partido 1, pero cree que en caso de que no suceda eso votará en blanco. El votante no puede faltar a la votación.

Supongamos que el votante va a votar al partido 2. Tenemos este diagrama:

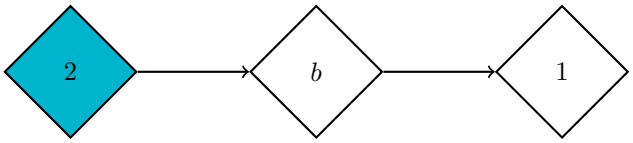


Figura 5.9: Ejemplo del votante para iteraciones infinitas

Una fuente confiable, le informa al encuestador: "El votante votará 2 o lo que vos creés acerca de si votará a 1 es incorrecto", que podríamos escribirlo como $\mathbf{P} = 2 \vee (1 \wedge B\neg 1) \vee (\neg 1 \wedge B1)$. En este modelo $\mathbf{P}$ sólo vale en los mundos 2 y b, pero no en 1. Supongamos que el encuestador hace una actualización conservadora con esta información: como el mundo más plausible satisfaciendo $\mathbf{P}$ es b el nuevo modelo es:

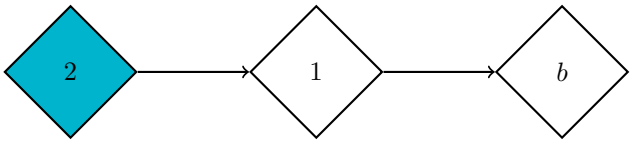


Figura 5.10: Luego de la primera actualización

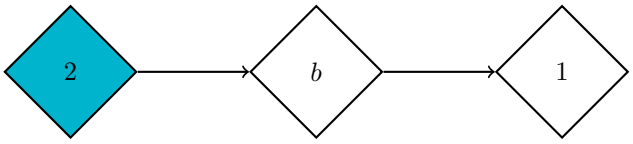
Observemos que $\mathbf{P}$ sigue siendo verdadera en el mundo real, entonces esta información puede ser (verazmente) anunciada de nuevo por la misma fuente. Pero cuando el encuestador vuelva a hacer la actualización de su modelo volverá al modelo inicial.

Para esta iteración de actualizaciones conservadoras, el orden de plausibilidad quedará oscilando infinitamente (y las creencias del encuestador también) aún a pesar de que todos los anuncios sean con la misma información.

En este caso una iteración $\uparrow \mathbf{P}, \uparrow \mathbf{P}, \ldots$, donde $\mathbf{P}$ es siempre verdadera en el mundo real no estabilizó al modelo inicial. Uno podría caer en la tentación de pensar que si para actualizaciones infalibles hay estabilizaciones (incorporación de información fuerte), pero para

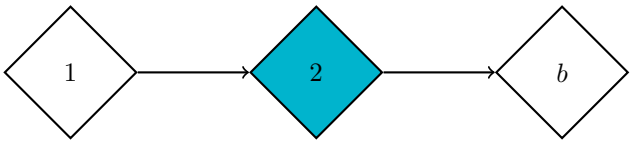
actualizaciones conservadoras (incorporación de información débil) no vale, quizás para actualizaciones "en el medio" (como actualizaciones radicales) puede llegar a valer. Bueno, es mejor no caer en esa tentación, porque tampoco vale para actualizaciones radicales:

Ejemplo 5.2.46. Partiendo del mismo modelo inicial

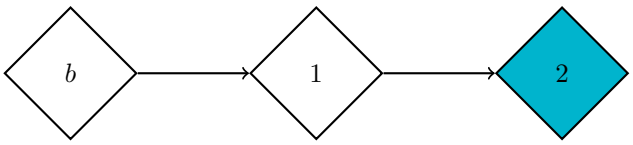


ahora consideremos que al encuestador una fuente fuertemente confiable le dice "Si supieras fehacientemente que el votante no votará a 2, entonces tu creencia acerca de si vota 1 o no será incorrecta". Este anuncio es capturado por la fórmula $\psi = [!\neg 2]((1 \wedge \neg B1) \vee (\neg 1 \wedge B1))$ que es equivalente a $2 \vee (1 \wedge \neg B^{\neg 2}1) \vee (\neg 1 \wedge B^{\neg 2}1)$.

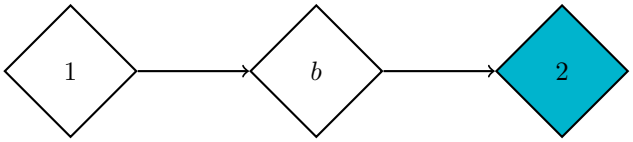
Esta fórmula es verdadera en 2 y en b, pero no en 1, entonces la actualización radical $\Uparrow \psi$ resulta en:



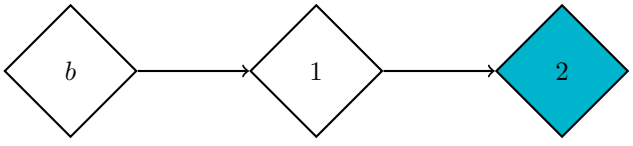
Una nueva ejecución de la actualización $\Uparrow \psi$ resulta en:



Una tercera vez, realizamos $\Uparrow \psi$:



Y hacerlo una vez más nos devuelve al modelo resultante de la primera actualización:



Por lo tanto, que la información nueva sea veraz no alcanza para que los modelos se estabilicen. Pero a diferencia del caso de actualizaciones conservadoras, en este ejemplo las

creencias se estabilizan: desde cierto momento en adelante el encuestador cree (correctamente) que el mundo real es 2.

Esto es un caso particular de un hecho más general, presentado en [9]:

Teorema 5.2.47. *Toda secuencia infinita de actualizaciones radicales veraces* $\{\Uparrow \boldsymbol{P}_k\}_k$*, aplicadas a un modelo inicial finito* $\mathfrak{M}_0$*, el conjunto de estados más plausibles eventualmente se estabiliza, después de finitas actualizaciones. Es decir que las creencias permanecen sin cambiar (a pesar de las posibles oscilaciones del orden de plausibilidad).*

Para demostrar esto vamos a precisar una serie de lemas previos:

Nota. Dada una secuencia de actualizaciones radicales veraces $\{\Uparrow \mathbf{P}_k\}_k$ aplicadas a un modelo inicial finito $\mathfrak{M}_0 = \langle W_0, (\leq_i^0)_{i\in N}, \mathtt{i}\rangle$, y $\mathfrak{M}_k = \langle W_k, (\leq_i^k)_{i\in N}, \mathtt{i}\rangle$ los modelos inducidos por la secuencia. En adelante vamos a denotar $max_i^k P = max_{\leq_i^k} P$.

Lema 5.2.48. *Sea una secuencia de actualizaciones radicales veraces* $\{\Uparrow \boldsymbol{P}_k\}_k$ *aplicadas a un modelo inicial finito* $\mathfrak{M}_0 = \langle W_0, (\leq_i^0)_{i\in N}, \mathtt{i}, w_0\rangle$*, y* $\mathfrak{M}_k = \langle W_k, (\leq_i^k)_{i\in N}, \mathtt{i}, w_0\rangle$ *los modelos inducidos por la secuencia. Vale que:*

$$\textit{Para todo } n \in \mathbb{N},\ w \in w_0(i) \cap \bigcap_{k<n} (\boldsymbol{P}_k)_{\mathfrak{M}_k} \textit{ y } v \in w_0(i) \setminus \left(\bigcap_{k<n} (\boldsymbol{P}_k)_{\mathfrak{M}_k}\right) \textit{ se tiene } v <_i^k w.$$

Demostración. La demostración es por inducción en n. Cuando $n = 0$, es trivial ya que $\bigcap_{k<n} (\mathbf{P}_k)_{\mathfrak{M}_k} = \emptyset$. Supongamos ahora que es verdadera para n, después de aplicar $\Uparrow \mathbf{P}_n$ a $\mathfrak{M}_n$ tenemos que $t <_i^{n+1} w$ para todo $w \in w_0(i) \cap (\mathbf{P}_n)_{\mathfrak{M}_n}$ y todo $t \in w_0(i) \setminus (\mathbf{P}_n)_{\mathfrak{M}_n}$ (ya que todos los mundos $\mathbf{P}_n$ son movidos por encima de los mundos $\neg\mathbf{P}_n$).

Además, como dentro de $w_0(i)\cap\mathbf{P}_n$ se preserva el orden $<_i^n$ y vale la hipótesis inductiva, tenemos que:

$$v <_i^{n+1} \text{ para todo } w \in w_0(i) \cap (\mathbf{P}_n)_{\mathfrak{M}_n} \bigcap_{k<n} (\mathbf{P}_k)_{\mathfrak{M}_k} \text{ y } v \in w_0(i) \cap (\mathbf{P}_n)_{\mathfrak{M}_n} \setminus \left(\bigcap_{k<n} (\mathbf{P}_k)_{\mathfrak{M}_k}\right)$$

De estos dos hechos y por transitividad de $<_i^{n+1}$, queda demostrado. □

Lema 5.2.49. *Bajo las mismas hipótesis del lema anterior vale que para todo* $n \in \mathbb{N}$*:*

$$max_i^n(w_0(i)) \subseteq \bigcap_{k<n} (\boldsymbol{P}_k)_{\mathfrak{M}_k}$$

Demostración. Supongamos que no, entonces existe un n tal que para algún estado $w \in max_i^n(w_0(i))$ pero $w \notin \bigcap_{k<n} (\mathbf{P}_k)_{\mathfrak{M}_k}$. Pero entonces $w \in w_0(i) \setminus \left(\bigcap_{k<n} (\mathbf{P}_k)_{\mathfrak{M}_k}\right)$ pero como estamos asumiendo que los anuncios son veraces se tiene que $w_0 \in \mathbf{P}_k$ para todo k, y en consecuencia $w_0 \in w_0(i) \cap \left(\bigcap_{k<n} (\mathbf{P}_k)_{\mathfrak{M}_k}\right)$. Por el Lema 5.2.48, tenemos que $w <_i^n w_0$, lo que contradice la hipótesis de que $w \in max_i^n(w_0(i))$.

□

Lema 5.2.50. *Bajo las hipótesis del lema anterior vale que para todo $n \in \mathbb{N}$:*

$$max_i^n(w_0(i)) = max_i^n\left(w_0(i) \cap \bigcap_{k<n} (\boldsymbol{P}_k)_{\mathfrak{M}_k}\right)$$

Demostración. Llamemos $Q = \bigcap_{k<n} (\mathbf{P}_k)_{\mathfrak{M}_k}$

Por el Lema 5.2.49 se tiene la inclusión $\subseteq$. Para ver $\supseteq$ tomemos un $w \in max_i^n(w_0(i) \cap Q)$ y sea $v \in w_0(i)$. Por la existencia de maximales en $w_0(i)$ (de la definición de modelo de plausibilidad), existe un $t \in max_i^n w_0(i)$ tal que $v \leq_i^n w$. Nuevamente por el Lema 5.2.49 tenemos que $t \in Q$, pero además $t \in w_0(i)$. Por lo tanto, como $t \in w_0(i) \cap Q$ y $w \in max_i^n(w_0(i) \cap Q)$ tenemos que $t \leq_i^n w$. Por transitividad $v \leq_i^n w$, y como $v \in w_0(i)$ era arbitrario tenemos que $w \in max(w_0(i))$. □

Lema 5.2.51. *Bajo las hipótesis del lema anterior, existe un n_0 tal que*

$$w_0(i) \cap \bigcap_{k<n_0} (\boldsymbol{P}_k)_{\mathfrak{M}_k} = w_0(i) \cap \bigcap_{k<m} (\boldsymbol{P}_k)_{\mathfrak{M}_k} \text{ para todo } m \geq n_0$$

.

Demostración. La cadena $w_0(i) \supseteq w_0(i) \cap (\mathbf{P}_0)_{\mathfrak{M}_0} \supseteq \ldots \supseteq w_0(i) \cap \bigcap_{k<n} (\mathbf{P}_k)_{\mathfrak{M}_k} \supseteq \ldots$ es una cadena descendiente comenzando en un conjunto finito, por lo tanto se estabiliza.

□

Finalmente estamos en condiciones de dar la demostración del Teorema 5.2.47:

Demostración. Por definición de $\Uparrow$ sabemos que para todo m, la actualización $\Uparrow \mathbf{P}_m$ deja intacto el orden dentro de $w_0(i) \cap (\mathbf{P}_m)_{\mathfrak{M}_m}$.

Sea un n_0 como el del Lema 5.2.51. Entonces tenemos que $w_0(i) \cap \bigcap_{k<n_0} (\mathbf{P}_k)_{\mathfrak{M}_k} \subseteq w_0(i) \cap (\mathbf{P}_m)_{\mathfrak{M}_K}$ para todo $m \geq n_0$.

Por lo tanto el orden dentro de $w_0(i) \cap \bigcap_{k<n_0} (\mathbf{P}_k)_{\mathfrak{M}_k}$ no va a cambiar con las posteriores actualizaciones $\Uparrow \mathbf{P}_k$ con $m \geq n_0$. En particular:

$$max_i^{n_0}\left(w_0(i) \cap \bigcap_{k<n_0} (\mathbf{P}_k)_{\mathfrak{M}_k}\right) = max_i^m\left(w_0(i) \cap \bigcap_{k<n_0} (\mathbf{P}_k)_{\mathfrak{M}_k}\right).$$

Esto junto con el Lema 5.2.50 y el hecho de que n_0 es el del Lema 5.2.51 nos da la siguiente cadena de igualdades:

$$max_i^{n_0}(w_0(i)) = max_i^{n_0}\left(w_0(i) \cap \bigcap_{k<n_0} (\mathbf{P}_k)_{\mathfrak{M}_k}\right) = max_i^m\left(w_0(i) \cap \bigcap_{k<n_0} (\mathbf{P}_k)_{\mathfrak{M}_k}\right)$$

$$= max_i^m\left(w_0(i) \cap \bigcap_{k<m} (\mathbf{P}_k)_{\mathfrak{M}_k}\right) = max_i^m(w_0(i))$$

. □

5.2.3. Confianza, sinceridad, honestidad y persuasión

Un anuncio con la misma información afecta de manera distinta a distintos agentes. Esto se debe a la predisposición del receptor a incorporar información proviniente del emisor. Según esta predisposición, el receptor actualizará su estructura de plausibilidades de distintas maneras. Es decir, ejecutará distintas actualizaciones. Esto motiva la siguiente definición:

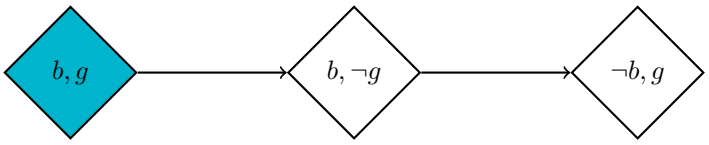
Definición 5.2.52. Llamamos *actitud doxástica dinámica (o simplemente actitud dinámica) de un agente hacia una fuente de información* a la forma de revisión de creencias que adopta el agente respecto a información dada por la fuente. Si llamamos τ a la actitud doxástica, al recibir la información $\mathbf{P}$ de la fuente se aplica una actualización $\tau\mathbf{P} : \mathfrak{M} \to \mathfrak{M}'$ que toma un modelo de plausibilidad y devuelve uno nuevo.

El concepto que se intenta capturar con esta definición es cómo recibe un agente la información proviniente de una fuente, y de qué manera revisa sus creencias con la información emitida por la fuente.

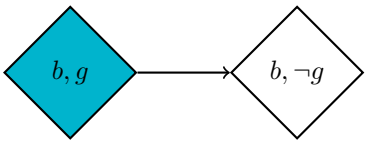
Definición 5.2.53. En lo que resta de la tesis vamos a trabajar únicamente con las siguientes actitudes doxásticas:

- Llamamos *confianza ciega* (y notamos !) a la actitud doxástica que a la proposición $\mathbf{P}$ le asigna la actualización $!\mathbf{P}$.
- Llamamos *confianza fuerte* (y notamos $\Uparrow$) a la actitud doxástica que a la proposición $\mathbf{P}$ le asigna la actualización $\Uparrow \mathbf{P}$.
- Llamamos *confianza* (y notamos $\uparrow$) a la actitud doxástica que a la proposición $\mathbf{P}$ le asigna la actualización $\uparrow \mathbf{P}$.
- Llamamos *indiferencia* (y notamos *id*) a la actitud doxástica que a la proposición $\mathbf{P}$ le asigna la actualización identidad.
- Llamamos *desconfianza* (y notamos $\uparrow^-$) a la actitud doxástica que a la proposición $\mathbf{P}$ le asigna la actualización $\uparrow \neg\mathbf{P}$.
- Llamamos *desconfianza fuerte* (y notamos $\Uparrow^-$) a la actitud doxástica que a la proposición $\mathbf{P}$ le asigna la actualización $\Uparrow \neg\mathbf{P}$.
- Llamamos *desconfianza ciega* (y notamos $!^-$) a la actitud doxástica que a la proposición $\mathbf{P}$ le asigna la actualización $!\neg\mathbf{P}$.

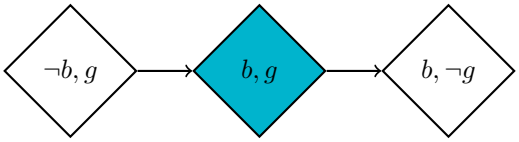
Ejemplo 5.2.54. Basándonos en el Ejemplo 5.2.39. Llamaremos τ a la actitud doxástica dinámica de Juan Pérez hacia su vecino. A continuación tenemos cómo actualiza Juan Pérez su estructura de plausibilidades (de acuerdo a distintas actitudes doxásticas) cuando vecino le dice que está borracho



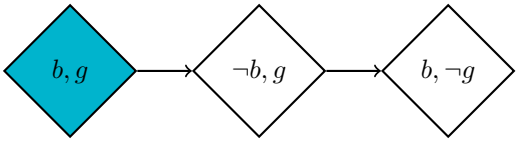
(a) Estado inicial de Juan Pérez



(b) Si $\tau =!$, ejecutará $!b$



(c) Si $\tau =\Uparrow$, ejecutará $\Uparrow b$



(d) Si $\tau =\uparrow$, ejecutará $\uparrow b$

Quisiésemos definir la noción de actitud doxástica positiva, es decir determinar cuándo un agente está bien predispuesto a recibir información de una fuente a partir de su actitud doxástica respecto a ella. Por ejemplo, las actualizaciones infalibles, radicales y conservadoras capturan una predisposición del agente a recibir la información, de hecho la reciben: después de las actualizaciones cree (o sabe) la información recibida.

¿Es la incorporación de la información lo que define la predisposición del agente?

Ejemplo 5.2.55. Supongamos que Ana y Blas son muy buenos amigos, motivo por el cual Ana en general está predispuesta a creer que todo lo que Blas dice es verdadero. Ana sabe hablar perfectamente inglés, pero no sabe si Blas sabe inglés o no. En este contexto llega Blas y dice "I can't pronounce any English word" en perfecto inglés.

Esta afirmación es consistente con el conocimiento de Ana, por lo que Ana puede ejecutar la actualización perfectamente. Por supuesto, después de oír a Blas, Ana no puede creer en esa afirmación porque el sólo hecho de que la haya pronunciado demuestra que es falsa.

Es importante observar que el anuncio no cambió el valor de verdad del mensaje, pero cuando Ana recibe esta información de Blas, ella instantáneamente incorpora la información opuesta a la que el mensaje transmitió.

Sin embargo, la actitud de Ana hacia Blas es positiva. Por esta clase de situaciones, lo que define una actitud doxástica positiva no es que después de recibir $\mathbf{P}$ lo crea, sino que después de recibir $\mathbf{P}$ lo crea salvo que contradiga su conocimiento. En otras palabras (u otros símbolos): $[\tau\mathbf{P}]\neg K\neg\mathbf{P} \to [\tau\mathbf{P}]B\mathbf{P}$.

Definición 5.2.56. Una actitud doxástica τ se dice *positiva* si para todo modelo vale $\Vdash [\tau\mathbf{P}]\neg K\neg\mathbf{P} \to [\tau\mathbf{P}]B\mathbf{P}$.

Se dice *negativa* si para todo modelo vale $\Vdash [\tau\mathbf{P}]\neg K\mathbf{P} \to [\tau\mathbf{P}]B\neg\mathbf{P}$

De las actualizaciones doxásticas que hemos vimos hasta ahora, la confianza ciega, confianza fuerte y confianza son actitudes doxásticas positivas; y la desconfianza ciega, desconfianza fuerte y desconfianza son negativas.

Como dijimos al comienzo de la sección, cada agente puede tener distintas actitudes doxásticas hacia las fuentes de información, aplicando cada uno una actualización diferente

ante un mismo anuncio: un agente puede considerar a una fuente fuertemente confiable, y mientras que otro no.

En un entorno comunicativo, las fuentes de información son agentes. Vamos a denotar τ^{ji} a la actitud del agente j como receptor de información proviniente del agente i. Formalmente τ^{ji} son transformaciones epistémicas-doxásticas (actualizaciones) que toman una proposición doxástica $\mathbf{P}$ y un modelo de plausibilidad uniagente $\mathfrak{M} = \langle W, \leq \rangle$ (de algún agente anónimo) y devuelve un nuevo modelo de plausibilidad $\mathfrak{M} = \langle W, \leq' \rangle$ (con el mismo conjunto de estados).

En adelante vamos a suponer que estas actitudes son conocimiento común. También vamos a asumir que todas las comunicaciones son públicas (los agentes sólo hacen anuncios públicos al grupo completo) y que las actitudes son estables bajo actualizaciones (los agentes no cambian de opinión acerca de la confiabilidad de la información dada por otros agentes).

Vamos a expandir el lenguaje para añadir las modalidades dinámicas $[i : \mathbf{P}]$, que representarán que el agente i anuncia públicamente $\mathbf{P}$. También añadiremos proposiciones atómicas τ_{ji}, (que representan que el agente j tiene la actitud τ hacia la información proviniente del agente i donde τ es un símbolo que representa una actitud doxástica dinámica ($!$, $\Uparrow$, $\uparrow$, etcétera): la actitud del agente j hacia la información proveniente del agente i.

Vamos a poner unas restricciones intuitivas a la interpretación del modelo para las proposiciones atómicas τ_{ji}:

- Para todo $w \in W$ existe una única actitud τ tal que $w \Vdash \tau_{ji}$

Es decir: en todos los mundos posibles, todos los agentes tienen una única actitud hacia el resto de los agentes.

- Si $w \overset{j}{\sim} v$ entonces $w \Vdash \tau_{ji} \leftrightarrow v \Vdash \tau_{ji}$

Esta restricción es para imponer introspección de las actitudes: cada agente conoce sus propias actitudes doxásticas.

Nuestra restricción a las actitudes estables y de conocimiento común queda fijada agregando esta otra restricción semántica:

- Para todo $w, v \in W$ vale $w \Vdash \tau_{ji} \leftrightarrow v \Vdash \tau_{ji}$

En muchos contextos se asume esta restricción para simplificar la estructura agregada al modelo. Bajo esta suposición, la extensión de la interpretación puede ser reemplazada por un grafo, conocido como grafo de confianza mutua, que tiene como nodos los agentes y las aristas etiquetados con la actitud doxástica correspondiente.

Sin embargo, no todos los grafos son consistentes con la suposición de que las actitudes doxásticas dinámicas son conocimiento común, tiene que cumplirse $!_{ji} \rightarrow !_{ki}$ y $!^{-}_{ji} \rightarrow !^{-}_{ki}$ para todo $j, k \neq i$, porque estas actitudes doxásticas se interpretan como "no opinables": para

ejecutar una actualización $!\mathbf{P}$ tiene que provenir de una fuente infalible que es públicamente sabida infalible (y análogamente para $!^-$).

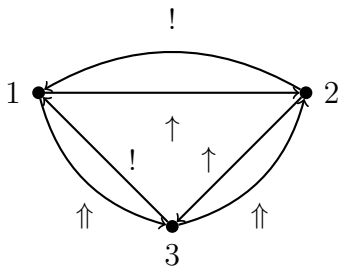


Figura 5.11: Grafo de confianza mutua

La actualización $i : \mathbf{P}$ está dada por una transformación que toma un modelo de plausibilidad $\mathfrak{M} = \langle W, \leq_j, \mathtt{i} \rangle$ y devuelve un nuevo modelo $i : \mathbf{P}(\mathfrak{M}) := \langle W', \leq'_j, \mathtt{i}' \rangle$, donde: los preordenes de los agentes receptores ($j \neq i$) están dados por aplicar dentro de cada celda de información de j la transformación τ, tal que τ es la única actitud cumpliendo τ_{ji} en toda la celda. Por otro lado, el preorden del emisor ($\leq_i$) es igual. El nuevo conjunto de mundos W' es la unión de todas las nuevas celdas de información y la nueva interpretación es la restricción $\mathtt{i}'(p) := \mathtt{i}(p) \cap W'$.

La modalidad dinámica se define del modo usual:

$$w \Vdash_{\mathfrak{M}} [i : \mathbf{P}]\mathbf{Q} \text{ sii } w \Vdash_{i:\mathbf{P}(\mathfrak{M})} \mathbf{Q}$$

Y su lectura es "si el agente i anuncia públicamente $\mathbf{P}$, luego de eso valdrá $\mathbf{Q}$".

Definición 5.2.57. Un acto comunicativo $i : \mathbf{P}$ es *sincero en un mundo w de un modelo $\mathfrak{M}$* si el emisor cree su propio anuncio, ie: si vale $w \Vdash_{\mathfrak{M}} B_i\mathbf{P}$. Si el modelo $\mathfrak{M}$ tiene mundo real indicado w^* decimos que el anuncio es *sincero en* $\mathfrak{M}$ si es sincero en w^*.

En situaciones cooperativas, es natural asumir el conocimiento común de la sinceridad. Esto puede ser formalizado modificando la semántica de $i : \mathbf{P}$, restringiendo la aplicabilidad a mundos en los que vale $B_i\mathbf{P}$. Llamamos a estas comunicaciones *inherentemente sinceras*. En tal caso hay que cambiar las reglas de reducción agregando la precondición $B_i\mathbf{P}$.

Observación 5.2.58. En general, sinceridad no implica veracidad: "sos el amor de mi vida" es un ejemplo clásico de una mentira sincera, popular entre los adolescentes.

Esta observación da lugar a una actitud doxástica dinámica mixta $!_i \Uparrow$ con una interpretación muy natural: el receptor cree fuertemente que el emisor i dice la verdad, pero además lo considera infalible cuando anuncia algo que puede saber por introspección ("Creo que...", "Sé que..."). Esta actitud es frecuente cuando se asume el conocimiento común de la sinceridad: cuando se anuncian fórmulas con propiedades introspectivas, sinceridad es

lo mismo que infalibilidad, en los demás casos la sinceridad al menos es una garantía de probabilidad alta.

En [26], tenemos un resultado que afirma que todas las sentencias $\mathbf{P}$ introspectivas para i, son equivalentes a $K_i\mathbf{P}$. Por lo tanto la actitud $!\Uparrow_{ji}$ puede ser entendida así: si el anuncio i es (equivalente a) un anuncio de la forma $K_i\mathbf{P}$, entonces j ejecuta una actualización infalible $!\mathbf{P}$. En caso contrario, ejecuta una actualización radical $\Uparrow \mathbf{P}$. Llamamos a esta actitud *confianza super-fuerte*.

Para no dejar lugar a dudas vamos a definirla en términos semánticos: si vale $K_j(K_i\mathbf{P} \vee K_i\neg\mathbf{P})$ (el agente j sabe que el agente i sabe si es verdad o no lo que dice) se ejecuta una actualización infalible $!\mathbf{P}$, en caso contrario ejecuta una actualización radical.

Definición 5.2.59. A cada actitud dinámica doxástica dada por una transformación τ, se le puede asociar una *actitud estática* $\bar{\tau}$. Decimos que el agente i tiene la actitud $\bar{\tau}$ hacia $\mathbf{P}$, si la estructura de plausibilidad de i es un punto fijo de la transformación $\tau\mathbf{P}$. Formalmente: $w \Vdash_{\mathfrak{M}} \bar{\tau}_i\mathbf{P}$ sii $\tau\mathbf{P}(W, (\leq_i)_{i\in N}, w^*) = (W, (\leq_i)_{i\in N}, w^*)$

Ejemplo 5.2.60. El punto fijo de las actualizaciones infalibles es el conocimiento: $\bar{!}_j\mathbf{P} \leftrightarrow K_j\mathbf{P}$, mientras que el de las actualizaciones radicales es la creencia fuerte: $\bar{\Uparrow}_j\mathbf{P} \leftrightarrow Sb_j\mathbf{P}$ y el de las actualizaciones conservadoras es la creencia: $\bar{\uparrow}_j\mathbf{P} \leftrightarrow B_j\mathbf{P}$

La importancia de los puntos fijos $\bar{\tau}$ es que capturan la actitud hacia una información que es inducida en un agente después de recibirla de una fuente hacia la que tiene una actitud τ. Expresan el alcance de la incorporación de una información en la estructura epistémica-doxástica: el hecho de que una actitud estática sea inducida en el receptor por un anuncio, significa que repetir el anuncio no aporta más información, es redundante.

La siguiente definición y ejemplo son esclarecedores al respecto del rol de las actitudes estáticas:

Definición 5.2.61. Dado un modelo $\mathfrak{M}$ y un mundo w decimos que un anuncio $i : \mathbf{P}$ es *honesto con respecto al receptor j en el mundo w* (y notamos: $H(i : \mathbf{P} \to j)$) si el emisor i tiene la misma actitud hacia $\mathbf{P}$ que la que (i cree que) será inducida en el receptor j en el mundo w. Es decir $w \Vdash H(i : \mathbf{P} \to j) = \bigwedge_{\tau}(B_i\tau_{ji} \to \bar{\tau}_i\mathbf{P})$.

Si el modelo posee mundo real w^*, decimos que el anuncio es *honesto con respecto al recepto j* si lo es en w^*. Y decimos que es *honesto* si es honesto con respecto a todos los receptores.

Nosotros estamos asumiendo que las actitudes τ_{ji} son conocimiento común, pero si no lo fueran tendrían creencias acerca de las distintas posibilidades y la honestidad quedaría representada en la siguiente fórmula: $H(i : \mathbf{P} \to j) = \tau_{ji} \to \bar{\tau}_i\mathbf{P}$.

Ejemplo 5.2.62. Asumamos que es conocimiento común que el emisor i es infalible (ie: vale $!_{ji}$ para todo $j \neq i$). Entonces el anuncio $i : \mathbf{P}$ es honesto sii i sabe que es verdad: si vale $K_i\mathbf{P}$. Esta misma condición asegura que $i : K_i\mathbf{P}$ es honesto.

Si en lugar de asumir que es infalible, asumimos que es conocimiento común que i es fuertemente confiable pero no infalible (ie: vale $\Uparrow_{ji}$ para todo $j \neq i$), se tiene que el anuncio $i : \mathbf{P}$ es honesto sii vale $Sb_i\mathbf{P}$.

Finalmente, si asumimos que es conocimiento común que i es confiable (ie: vale $\uparrow_{ji} \varphi$ para todo $j \neq i$), entonces el anuncio $i : \varphi$ es honesto sii vale $B_i\varphi$.

¿Hay relación entre sinceridad y honestidad? ¿Ser sincero implica ser honesto?

Ejemplo 5.2.63. Diálogo tomado de la película Piratas del caribe en [3]:

Mullroy: What's your pupose in Port Royal, Mr Smith?
Murtogg: Yeah, and no lies.
Jack Sparrow: Well, then, I confess, it is my intention to commandeer one of these ships, pick up a crew in Tortuga, raid, pillage, plunder and otherwise pilfer my weasely black guts out.
Murtogg: I said no lies.
Mullroy: I think he's telling the truth.
Murtogg: Don't be stupid: if he were telling the truth, he wouldn't have told it us.
Jack Sparrow: Unless, of course, he knew you wouldn't believe the truth even if he told it to you.

En este ejemplo hay un agente i (Jack Sparrow) que cree fuertemente $\mathbf{P}$, de hecho: sabe $\mathbf{P}$ por lo tanto $\mathbf{P}$ es verdadera. En este caso $\mathbf{P}$ es que va a tomar una nave, levantar una tripulación y cometer sus fechorías.

Supongamos que i sabe que él es fuertemente desconfiado por su audiencia j, es decir tenemos $\Uparrow^{-}_{ji}$ y esto es de público conocimiento. Entonces, el anuncio $i : \mathbf{P}$ es sincero (incluso veraz) pero deshonesto.

Sinceridad no implica honestidad, pero ¿honestidad implica sinceridad? No. La sinceridad está relacionada con la actitud doxástica del emisor hacia el anuncio, y la honestidad con la actitud doxástica del receptor hacia el anuncio, y estas actitudes no necesariamente están relacionadas.

De hecho para un agente i, que es fuertemente desconfiable (y además esto conocimiento común), sólo puede hacer un anuncio es honesto sii i cree fuertemente que es falso.

Sin embargo, en contextos donde los agentes tienen actitud positiva hacia los otros agentes, la honestidad implica sinceridad.

Definición 5.2.64. Decimos que un anuncio $i : \mathbf{P}$ es *persuasivo hacia un agente j en el mundo w respecto de un asunto $\boldsymbol{Q}$ para el que i tiene una actitud doxástica* ϵ (y escribimos $P(i : \mathbf{P} \to j; \mathbf{Q})$) si el efecto del anuncio es que el receptor j convierte su actitud respecto de $\mathbf{Q}$ a la misma del emisor i. Es decir: $w \Vdash \tau_{ji} \to [i : \mathbf{P}]\epsilon_j\mathbf{Q}$.

En un modelo con mundo real indicado decimos que el mismo anuncio *es persuasivo hacia un agente j respecto de un asunto $\boldsymbol{Q}$ para el cual tiene una actitud* ϵ (sin especificar

cuál es el mundo), entonces nos referimos a que el anuncio es persuasivo hacia j respecto de $\mathbf{Q}$ en el mundo real.

Ejemplo 5.2.65. Supongamos que el agente i es fuertemente confiable, y quiere ser honesto pero persuasivo con respecto a un asunto $\mathbf{Q}$ en el cual cree débilmente. Es decir, hablamos de un contexto comunicativo en el que vale $\Uparrow_{ji}$ para todo $j \neq i$ y $B_i\mathbf{Q}$ (pero no vale $Sb_i\mathbf{P}$). ¿Qué puede anunciar i para ser honesto y persuasivo? ¿Cómo puede el agente i convertir a los otros agentes a las creencias propias, manteniendo la honestidad y sinceridad?

- ¿Sirve anunciar $\mathbf{Q}$?

El anuncio $i : \mathbf{Q}$ es sincero ya que vale $B_i\mathbf{Q}$. También es un anuncio persuasivo, porque el agente tiene la actitud doxástica B hacia el asunto $\mathbf{Q}$. Sin embargo es un anuncio deshonesto, porque no vale $Sb_i\mathbf{Q}$ (y el punto fijo de $\Uparrow$ es Sb).

- ¿Y si se anuncia $K_i\mathbf{Q}$?

El anuncio $i : K_i\mathbf{Q}$ ni siquiera es sincero, ya que no es verdad que $B_iK_i\mathbf{Q}$: Si en el más plausible de los mundos de $w^*(i)$ vale $K_i\mathbf{Q}$, por definición de K_i se tiene que $w^* \Vdash K_i\mathbf{Q}$ también. Pero dijimos que el agente no cree fuertemente en $\mathbf{Q}$, por lo tanto no puede saberlo.

- Entonces que anuncie $B_i\mathbf{Q}$...

Ejecutar $i : B_i\mathbf{Q}$ sería una comunicación sincera, porque $w^* \Vdash B_iB_i\mathbf{Q}$. Además es honesta porque el punto fijo de $\Uparrow$ es Sb y $w^* \Vdash Sb_iB_i\mathbf{Q}$ (ya que no hay mundos en $w^*(i)$ en los que no valga B_i). Pero no es persuasiva: no cambia las creencias de los agentes acerca de $\mathbf{Q}$, sino acerca de las creencias de i sobre $\mathbf{Q}$.

- Bueno... no sé... que se anuncie $\Box_i\mathbf{Q}$

$i : \Box_i\mathbf{Q}$ es un anuncio sincero, ya que por la Proposición 5.2.20 vale $w^* \Vdash B_i\Box_i\mathbf{Q}$. Además es un anuncio honesto, ya que $w^* \Vdash Sb_i\Box_i\mathbf{Q}$. En efecto, sean $w \Vdash \Box_i\mathbf{Q}$ y $v \nVdash \Box_i\mathbf{Q}$, entonces existe un $t \geq v$ tal que $t \nVdash \mathbf{Q}$. En consecuencia necesariamente vale $t \leq w$ (porque $w \Vdash \Box_i\mathbf{Q}$) y, como $\leq$ es transitiva, $v \leq w$.

Y también es un anuncio persuasivo porque $w \Vdash [i : \Box_i\mathbf{Q}]B_j\mathbf{Q}$. En efecto, sea $j \neq i$ y $w \in max_{\leq'_j} w^*(j)$ donde $\leq'_j$ representa el orden en $\mathfrak{M}'$, el modelo que resulta de ejecutar la acción $i : \Box_i\mathbf{Q}$. Queremos ver que $w \Vdash \mathbf{Q}$. Supongamos que no vale, entonces existe un $w \in max_{\leq'_j} w^*(j)$ tal que $w \nVdash_{\mathfrak{M}'} \mathbf{Q}$ y, por lo tanto, $w \nVdash_{\mathfrak{M}'} \Box_i\mathbf{Q}$. Pero, como $w^* \Vdash B_i\Box_i\mathbf{Q}$ (por la Proposición 5.2.20), entonces para todo $v \in max_i w^*(i)$ vale $v \Vdash \Box_i\mathbf{Q}$. Dado que la actualización con $\Uparrow$ (para la estructura de plausibilidad del agente j) coloca a todos los mundos en los que vale $\Box_i\mathbf{Q}$ estrictamente encima de aquellos en los que no vale, tenemos que $w <'_j v$, y esto niega la hipótesis de que $w \in max_{\leq'_j} w^*(j)$.

Por lo tanto tenemos que $i : \Box_i \mathbf{Q}$ es un anuncio sincero, honesto y persuasivo. A este tipo de anuncios se los llama *exageración honesta*.

Moraleja (conocida por la mayoría de los políticos): Todo lo que se necesita para convertir honestamente a otros: un público que crea fuertemente en lo que oye y anunciar que uno "sabe" cosas (aún cuando uno sólo las cree débilmente).

> "We know that Saddam Hussein has acquired weapons of mass destruction..."
> - George W. Bush, 2002.

5.2.4. Unión de información

Un aspecto interesante de la revisión dinámica en modelos multiagente consiste en la posibilidad estudiar el comportamiento de un grupo a partir de la estructura de plausibilidad de cada agente. El problema central en la teoría de elección colectiva es ese mismo: ¿Cuál es la información que tiene un grupo (como unidad) a partir de la que tiene cada uno de los individuos? Si todos los agentes comparten su información ¿qué sabrían, qué creerían? (en los distintos sentidos posibles: K, $\Box$, B, Sb, etcétera)

Definición 5.2.66. Una *fusión de relaciones de un grupo* G es una función $\bigodot$ que toma relaciones $(\mathbb{R}_i)_{i\in G}$ y devuelve una relación $\bigodot_{i\in G}$, en el mismo conjunto de mundos posibles.

Dependiendo del contexto, uno podría querer fusionar las creencias de los agentes (dada por las relaciones $\rightarrow_i$) o sus conocimientos (dado por $\overset{i}{\sim}$), u otro tipo de actitudes doxásticas.

Vamos a presentar los dos tipos de fusiones más importantes.

Definición 5.2.67. La *fusión paralela (o por intersección)* consiste en $\bigodot_{i\in G} R_i := \bigcap_{i\in G} R_i$.

Este tipo de fusión funciona bien para fusionar conocimiento certero (K) ya que con este tipo de conocimiento no hay riesgo de obtener una relación vacía, porque el conocimiento certero de los distintos agentes no puede ser inconsistente. De hecho, este tipo de fusión sirve para representar el conocimiento (certero) distribuído.

Para información más "débil" vamos a precisar otro tipo de fusión, porque son potencialmente inconsistentes, conduciendo a una intersección $\bigcap_{i\in G} R_i$ vacía.

Definición 5.2.68. La *fusión prioritaria de preordenes según* $a_1, \ldots, a_n$ está dada por el orden lexicográfico $(\leq_{a_1}, \ldots, \leq_{a_n})$.

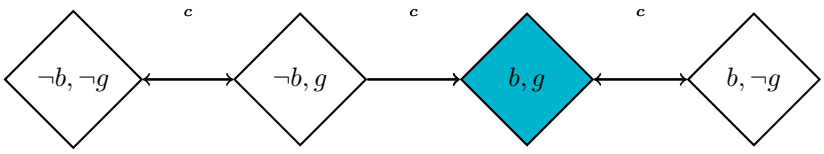
Sin embargo este tipo de fusión no es 100 % satisfactoria si queremos fusionar las estructuras de plausibilidad completas (con su información fuerte y su información débil). En particular si un agente sabe que un mundo es imposible pero uno de mayor prioridad no, esta información no trasciende al grupo. Así nace un tercer tipo de fusión:

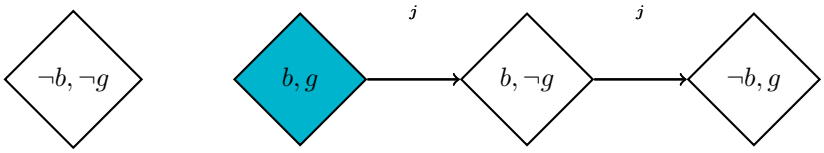
Definición 5.2.69. La *fusión prioritaria relativa* es una combinación entre una fusión paralela y una fusión prioritaria.

Se establece un orden de prioridad de los agentes $a_1, \ldots, a_n$ (n = cantidad de agentes en G) y la relación $\leq_G := \bigodot_{i \in G} \leq_i$ queda definida en el dominio de $\bigcap_{i \in G} \overset{i}{\sim}_i$ por el orden lexicográfico dado por $\leq_{a_1}, \leq_{a_2}, \ldots, \leq_{a_n}$ (donde $\sim_i = \leq_i \cup \geq_i$)[2].

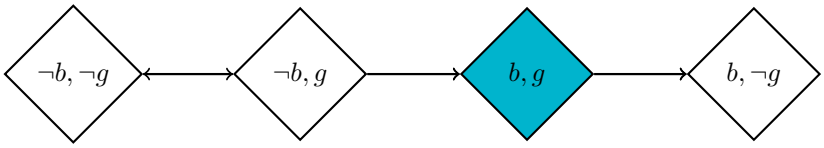
En palabras: esta relación representa un orden de plausibilidad determinado por los agentes en un cierto orden, en el conjunto de mundos plausibles para todos. En este contexto dos mundos son igualmente plausibles si lo son para todos los agentes.

Ejemplo 5.2.70. Supongamos que queremos fusionar las estructuras de plausibilidad completas de Juan Pérez y Carlos Fernández del Ejemplo 5.2.14, donde la prioridad la tiene Carlos (que no es un borracho). Los modelos serían:

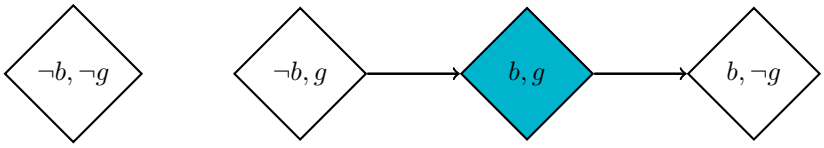




Luego de la fusión prioritaria se obtiene:



Mientras que si elegimos fusionar con fusión prioritaria relativa tenemos:



[2]En realidad pueden ser relaciones arbitrarias, pero por motivos de claridad lo expresamos directamente en términos de $\leq_i$ y $\sim_i$.

Pero una fusión de información en un grupo es un proceso que lleva varios pasos hasta que finalmente todos los miembros del grupo llegan a un acuerdo. Por lo tanto, queremos estudiar las fusiones de un modo dinámico: ¿cómo alcanzar una fusión de las plausibilidades a través de una comunicación de agentes que comparten información? El proceso será distinto dependiendo de si los agentes quieren llegar a un acuerdo acerca de lo que saben, de lo que creen, de lo que creen fuertemente, etcétera. Aspiramos a modelar un proceso en el que a medida que los agentes van compartiendo distinta información todo el grupo va *revisando* su información, hasta llegar a un acuerdo total en el que todos los agentes tienen la misma información y estructura de plausibilidades.

Como esto corresponde a contextos colaborativos, vamos a asumir que los agentes tienen todos la misma actitud los unos hacia los otros, que esas actitudes son conocimiento común y que los anuncios son sinceros, honestos, persuasivos y públicos. Cuando un agente hace un anuncio sincero, honesto, persuasivo y público vamos a decir que está *compartiendo información.*

Queremos definir un proceso que cumpla que cada vez que un agente comparte información, todos los agentes llegan a un acuerdo parcial acerca de esa información, y que el proceso termine cuando se alcanza un acuerdo total, i.e.: cuando las estructuras doxásticas de todos los agentes son iguales. El problema de fusión dinámica de creencias en contextos colaborativo, fue estudiado por primera vez en [8] y profundizado en [9], de donde tomamos las siguientes definiciones y resultados.

Definición 5.2.71. Llamamos *realización de la fusión* $\bigodot$ a una secuencia π de anuncios en los que los agentes comparten información de manera ordenada, transformando un modelo inicial $\mathfrak{M} = \langle W, R_i\rangle_{i\in G}$ en un modelo $\mathfrak{M}' = \langle W, R'_i\rangle_{i\in G}$ donde $R'_j = \bigodot_{i\in G} R_i$

Ejemplo 5.2.72. Supongamos que tenemos un modelo de dos agentes a y b con mundo real w^*, en el que es conocimiento común la infalibilidad de ambos. Consideramos la secuencia $\pi := \rho_1;\rho_2$, donde $\rho_i := (i : w^*_{i-1}(i))$ y $w^*_j(i)$ es la clase de comparabilidad de i luego de ser ejecutados los primeros j anuncios. Resulta una realización de la fusión paralela de las relaciones $\sim_i$.

Observación 5.2.73. Como supusimos infalibilidad de los agentes, los anuncios resultan en una actualización infalible de los órdenes de plausibilidad cuyo único efecto es descartar los mundos en los que no vale el anuncio.

Veamos que es así: tenemos un modelo $\mathfrak{M} = \langle W, (\leq_i)_{i\in N}, \mathtt{i}, w^*\rangle$. Si resulta que $\pi\mathfrak{M} = \langle W', (\leq'_i)_{i\in N}, \mathtt{i}, w^*\rangle$, queremos ver que la relación $\sim'_i := \leq'_i \cup \geq'_i$ cumple $w^* \sim'_i w$ sii $w^* \bigcap_{i\in G} \sim_i w$.

$\Rightarrow$) Si $w^* \sim'_i w$ entonces w no fue descartado con el anuncio $(1 : w^*_0(1))$ por lo tanto $w \in w^*_0(1)$, es decir: $w^* \sim_1 w$. Además si $w^* \sim'_i w$, entonces w no fue descartado con el anuncio $\rho_2 = (2 : w^*_1(2))$, entonces $w \in w^*_1(2) \subseteq w_0(2)$ (esto último porque ρ_1 puede descartar mundos pero no agregar) y por lo tanto $w^* \sim_2 w$.

$\Leftarrow$) Supongamos que $w^* \sim_1 w$ y $w^* \sim_2 w$, queremos ver que w no se descarta con los anuncios ρ_1 y ρ_2 porque, como las relaciones resultantes son simples reestricciones, ver que

no se descarta w es lo único que necesitamos para ver que $w^* \sim'_i w$. Como $w^* \sim_1 w$, se tiene $w \in w^*(1) = w^*_0(1)$ y no se descarta al ejecutarse ρ_1. Como $w^* \sim_2 w$ y al ejecutarse ρ_1 no se descartó w, tenemos que $w \in w^*_1(2)$ y, en consecuencia, tampoco se descarta al ejecutar ρ_2.

Por lo tanto π es una realización de la fusión paralela de $\sim_1$ y $\sim_2$.

Esto se puede generalizar inmediatamente por inducción:

Proposición 5.2.74. *Si es conocimiento común que las actitudes doxásticas de todos los agentes son* !*, entonces la secuencia* $\pi := \rho_1; \ldots; \rho_n$ *con* $\rho_i := (i : w^*_{i-1}(i))$ *y* $w^*_j(i)$ *la clase de comparabilidad de* i *luego de ser ejecutados los primeros* j *anuncios, es una realización de la fusión paralela de las relaciones* $\sim_i$*.*

Es decir que si cada agente anuncia lo que sabe, se realiza la fusión paralela.

Con razonamientos parecidos obtenemos:

Proposición 5.2.75. *Si es conocimiento común que la actitud doxástica de todos los agentes es* $\Uparrow$*, entonces la secuencia* $\pi := \rho'_{a_1}; \ldots; \rho'_{a_n}$*, donde* $\rho'_i := (i : \Box_i P)$ *con* $P = max_i(w^*_{i-1}(i))$ *y* $w^*_j(i)$ *la clase de comparabilidad de* i *luego de ser ejecutados los primeros* j *anuncios, es una realización de la fusión prioritaria de las relaciones* $\leq_i$ *según el orden* $a_1, \ldots, a_n$*.*

Esta propiedad indica que la fusión prioritaria se puede realizar si, respetando el orden de prioridad de los agentes, cada agente anuncia lo que cree que sabe. Recordemos que estos son anuncios sinceros, honestos y persuasivos.

Problema 5.2.76. *Ni la fusión paralela ni la fusión prioritaria preserva la condición de maximales en subconjuntos no vacíos cuando el modelo es infinito. Por lo tanto este tipo de fusiones no son compatibles con nuestros modelos. El problema de fusiones de creencias sigue siendo un problema abierto para el caso infinito.*

Capítulo 6

Conclusiones

A lo largo de esta tesis hemos presentado numerosos modelos para representar el conocimiento y las creencias de uno o más agentes. Los modelos epistémico y doxástico más clásicos son los vistos en el Capítulo 3, que responden a las axiomatizaciones **S5** y **KD45** respectivamente. Para estos presentamos lógicas y probamos resultados de completitud, correctitud y decibilidad. Estas lógicas han sido especialmente bien recibidas para modelar agentes con alto poder de cómputo y juegos con jugadores racionales, tales como el juego de los niños embarrados presentado en la Introducción.

Además hemos discutido las falencias de estas representaciones y presentado modelos alternativos, útiles en otros contextos, principalmente aquellos que pretenden modelar con más detalle el conocimiento humano.

En el Capítulo 4 comenzamos a introducir aspectos dinámicos y estudiar los cambios en las estructuras de conocimientos. En el Capítulo 5 extendimos este estudio a creencias, siendo estos últimos los más complicados ya que nos enfrentamos con potenciales inconsistencias. A partir de los modelos dinámicos estudiados pudimos capturar los conceptos, propuestos en la Introducción, de sinceridad, confianza, honestidad y persuación. Y basándonos en ellos vimos cómo se pueden fusionar de manera estática y dinámica las creencias de un grupo en un contexto colaborativo de modelos finitos. Proponemos como trabajo futuro, el estudio de fusiones de creencias para múltiples agentes para modelos infinitos.

Por motivos de síntesis, no fue posible presentar otros numerosos aspectos interesantes de la lógica epistémica pero proponemos al lector interesado sumergirse en la bibliografía para descubrir un sinfín de estudios filosóficos y matemáticos inmensamente cautivadores.

Bibliografía

[1] Marco Aiello, Ian Pratt-Hartmann, y Johan van Benthem, editors. *Handbook of Spatial Logics*. Springer, 2007.

[2] Carlos Alchourròn, Peter Gärdenfors, y David Makinson. On the logic of theory change: Partial meet contraction and revision functions. *Journal of Symbolic Logic*, 50:510–530, 1985.

[3] Alexandru Baltag. Topics in dynamic epistemic logic.

[4] Alexandru Baltag, Lawrence Moss, y Slawomir Solecki. The logic of public announcements, common knowledge, and private suspicions. In *Proceedings of the 7th Conference on Theoretical Aspects of Rationality and Knowledge*, TARK '98, pages 43–56, San Francisco, CA, USA, 1998. Morgan Kaufmann Publishers Inc.

[5] Alexandru Baltag y Lawrence S. Moss. Logics for epistemic programs. *Synthese: Knowledge, Rationality and Action*, 139(2):165–224, 2004.

[6] Alexandru Baltag y Sonja Smets. A qualitative theory of dynamic interactive belief revision. In G Bonanno, W van der Hoek, y M Wooldridge, editors, *Proceedings of LOFT 2007*, volume 3 of *Texts in Logic and Games*, pages 9–58. Amsterdam University Press, 2008.

[7] Alexandru Baltag y Sonja Smets. Group belief dynamics under iterated revision: fixed points and cycles of joint upgrades. In *Proceedings of the 12th Conference on Theoretical Aspects of Rationality and Knowledge (TARK-2009), Stanford, CA, USA, July 6-8, 2009*, pages 41–50, 2009.

[8] Alexandru Baltag y Sonja Smets. Talking your way into agreement: Belief merge by persuasive communication. In Matteo Baldoni, Cristina Baroglio, Jamal Bentahar, Guido Boella, Massimo Cossentino, Mehdi Dastani, Barbara Dunin-Keplicz, Giancarlo Fortino, Marie Pierre Gleizes, João Leite, Viviana Mascardi, Julian Padget, Juan Pavón, Axel Polleres, Amal El Fallah-Seghrouchni, Paolo Torroni, y Rineke Verbrugge, editors, *MALLOW*, volume 494 of *CEUR Workshop Proceedings*. CEUR-WS.org, 2009.

[9] Alexandru Baltag y Sonja Smets. Protocols for belief merge: Reaching agreement via communication. *Logic Journal of the IGPL*, 21(3):468–487, 2013.

[10] Alexandru Baltag, Hans van Ditmarsch, y Lawrence Moss. Epistemic Logic and Information Update. In Pieter Adriaans y Johan van Benthem, editors, *Handbook on the Philosophy of Information*, pages 369–463. Elsevier Science, 2008.

[11] J. C. Beall y G. Restall. Logical consequence. In Edward N. Zalta, editor, *The Stanford Encyclopedia of Philosophy*. Center for the Study of Language and Information, Stanford University, fall 2014 edition, 2014.

[12] Patrick Blackburn, de Maarten Rijke, y Yde Venema. *Modal Logic*. Cambridge University Press, New York, NY, USA, 2001.

[13] Patrick Blackburn, Johan van Benthem, y Frank Wolter. *Handbook of Modal Logic, Volume 3 (Studies in Logic and Practical Reasoning)*. Elsevier Science Inc., New York, NY, USA, 2006.

[14] Ronald Fagin y Joseph Y. Halpern. Belief, awareness, and limited reasoning. *Artificial intelligence*, 34:39–76, 1988.

[15] Dov Gabbay y John Woods. Logic and modalities in the twentieh century. In Dov Gabbay Christoph Benzmüller y John Woods, editors, *Handbook of the History of Logic*, volume 7. Elsevier, 2014.

[16] Edmund Gettier. Is justified true belief knowledge? *Analysis*, 23:121–123, 1963.

[17] Patrick Girard, Olivier Roy, y Mathieu Marion. *Dynamic Formal Epistemology*. Springer, 2011.

[18] Jaakko Hintikka. *Knowledge and Belief*. Ithaca, N.Y.,Cornell University Press, 1962.

[19] Geoffrey Hunter. *Metalogic: An Introduction to the Metatheory of Standard First Order Logic*. Berkeley, University of California Press, 1971.

[20] John-Jules Meyer y Wiebe van der Hoek. *Epistemic Logic for AI and Computer Science*. Cambridge University Press, New York, NY, USA, 1995.

[21] Frank R. Palmer. *Mood and Modality*. Cambridge University Press, 2001.

[22] J. A. Plaza. Logics of public communications. In M. L. Emrich, M. S. Pfeifer, M. Hadzikadic, y Z. W. Ras, editors, *Proceedings of the Fourth International Symposium on Methodologies for Intelligent Systems*, pages 201–216. Oak Ridge National Laboratory, 1989.

[23] Henrik Sahlqvist. Completeness and correspondence in the first and second order semantics for modal logic. In Kanger Steig, editor, *Proceedings of the Third Scandinavian Logic Symposium, Uppsala, 1973*. North-Holland Publishing Company, Amsterdam, 1975.

[24] Frederick Schmitt. Knowledge and Belief. In *Problems of Philosopy*. Routledge London, New York, 1992.

[25] Robert Stalnaker. On logics of knowledge and belief. *Philosophical Studies*, 128(1):169–199, 2006.

[26] Hans van Ditmarsch, Wiebe van der Hoek, y Barteld Kooi. *Dynamic Epistemic Logic*. Springer Publishing Company, Incorporated, 1st edition, 2007.

[27] Wiebe vander Hoek. Systems for knowledge and belief. *Journal of Logic and Computation*, 3(2):173–195, 1993.