2 INTENCIONALIDAD

La noción de **agencia débil** [187], adoptada en el capítulo anterior, define un aparato conceptual entorno a la autonomía, la iniciativa y la sociabilidad, suficiente para caracterizar los atributos ineludibles en el comportamiento de un agente –Nos permite diferenciar lo que es un agente, de lo que no lo es. Sin embargo, y esto es particularmente cierto en el contexto de la IA, solemos estar interesados en descripciones de los agentes más cercanas a las que usamos al hablar del comportamiento de los seres humanos como agentes racionales con creencias, deseos, intenciones y otros atributos más allá de la agencia débil. Siendo ese nuestro interés, es necesario abordar una **noción fuerte de agencia** que provea los argumentos a favor de tal postura.

Agencia débil

Agencia fuerte

Así como la noción débil de agencia parece tener su piedra angular en el concepto de **acción** flexible y autónoma, la piedra angular en la noción fuerte de agencia es la **Intencionalidad**[1], entendida como la propiedad de algunos estados mentales de ser acerca de algo [120]. En este capítulo se abordaran tres dimensiones de la Intencionalidad de nuestro interés:

Intencionalidad

- La **postura Intencional**, propuesta por Dennett [48]. Se trata de una estrategia para interpretar el comportamiento de los otros, al adscribirles creencias, deseos, e intenciones; y asumir que se trata de agentes racionales. ¿Cómo es que el uso de **representaciones** Intencionales hace posible tal interpretación?

  Representación

- El **razonamiento práctico** en la teoría sobre los planes, las intenciones y la razón práctica de Bratman [22], que provee una definición funcional de **racionalidad**. ¿Cómo formamos y revisamos nuestras intenciones? ¿Qué estructura tiene una intención? ¿Qué rol juegan las intenciones en el razonamiento práctico?

  Racionalidad

- Los **actos de habla**, propuestos por Searle [164], como una teoría de **comunicación** de nuestros estados Intencionales. ¿Qué intención tiene un agente racional al comunicarse? ¿Qué relación hay entre su estado mental y sus mensajes? ¿Cómo se debe responder a un acto de habla?

  Comunicación

Evidentemente, el orden de la presentación no es cronológico y obedece a que no todos los sistemas Intencionales son capaces de comunicarse, y no todos los sistemas Intencionales son capaces de planear al estilo del razonamiento práctico. En cambio, la postura Intencional, asume alguna forma de racionalidad, como lo puede ser el razonamiento práctico. Y que por supuesto, algunos de entre ellos pueden comunicar.

Finalmente, revisaremos algunos **argumentos computacionales** [125, 168,

Intencionalidad y computación

1 Para distinguir este uso técnico del término Intencionalidad, se le denotará con una mayúscula inicial, mientras que intención; Con minúscula inicial, denotará el sentido común del término, como en "Tiene la intención de ganarse una beca" ó en "Disparo intencionalmente".

176, 60] a favor de modelar, describir, implementar y razonar acerca de nuestros sistemas de cómputo, como si fuesen agentes racionales que implementan alguna forma de razonamiento práctico y comunicación basada en actos de habla, es decir, como Sistemas Intencionales.

2.1 INTENCIONALIDAD

Muchos de nuestros estados mentales están en cierto sentido dirigidos a objetos o asuntos del mundo. Si tengo una creencia, debe ser una creencia que tal y tal es el caso; si deseo algo debe ser el deseo de hacer algo, o que algo ocurra, o que sea el caso; si tengo una intención, debe ser la intención de hacer algo; etc. Es esta característica **direccional** de algunos de nuestros estados mentales, es lo que muchos filósofos han etiquetado como Intencionalidad [165].

Dirección e Intención

Lo que es relevante en esta caracterización, es que los estados mentales Intencionales parecen tener una estructura o prototipo que consiste en una **actitud**, como creer, desear, intentar, etc., que opera sobre el **contenido** del estado, que a su vez está relacionado con algo más allá de si mismo, el objeto hacía el cual apunta. En este sentido, los estados Intencionales son **representaciones de segundo orden**, es decir, representaciones de representaciones. Si además, el contenido de un estado Intencional se puede expresar en forma proposicional, hablamos de una **actitud proposicional**. La Figura 2.1 ilustra el orden superior de las actitudes proposicionales, en tanto que representaciones.

Actitud

Representación de segundo orden

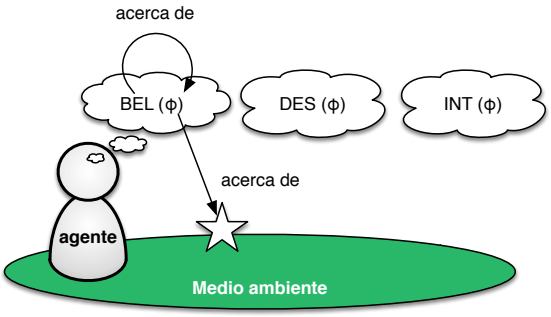

Figura 2.1: Las actitudes proposicionales son representaciones de segundo orden. La proposición ϕ puede ser "la estrella es blanca" y es acerca de la estrella blanca en el medio ambiente. El agente puede tener como actitud creer (BEL) que la estrella es blanca, o desearlo (DES), o intentarlo (INT). Las actitudes son acerca de la proposición, no acerca de la estrella en el medio ambiente.

El estudio de la Intencionalidad tiene su origen en las discusiones filosóficas medievales sobre la diferencia entre la existencia natural de las cosas, o *esse naturae*, y la existencia mental o intencional de las cosas, o *esse intentionale*, que deriva del latín *intentio* y significa dirigir la atención del pensamiento hacía algo, o simplemente apuntar hacía un objetivo, o **ser acerca de** [120]. La doctrina escolástica afirma que todos los hechos de conciencia poseen y manifiestan una dirección u orientación hacia un objeto. Esta orientación, que se afirma de todo pensamiento, volición, deseo o representación, en general

Intentio

consiste en la presencia o existencia mental del objeto que se conoce, quiere o desea; y en la referencia de este hecho a un objeto real [61]. Pero fue Brentano [24] en el siglo XIX, quien desarrolló la idea de que la Intencionalidad es la característica propia de todos los fenómenos mentales.

2.2 SISTEMAS INTENCIONALES

De acuerdo con Dennett [47, 48], los sistemas Intencionales son por definición, todas y sólo aquellas entidades cuyo comportamiento puede ser explicado o predicho, al menos algunas veces, asumiendo una **postura Intencional**. Tal postura consiste en la interpretación del comportamiento de la entidad en cuestión (persona, animal o artefacto) asumiendo que se trata de un **agente racional** que gobierna su selección de acción considerando sus **actitudes proposicionales**: creencias, deseos, intenciones, etc. La postura Intencional puede verse como una **estrategia**, de entre otras posibles, para explicar o predecir el comportamiento de un agente. Dennett considera además, las estrategias físicas y de diseño.

Postura Intencional

Racionalidad

Estrategias

2.2.1 Posturas física, de diseño e Intencional

Consideren el siguiente ejemplo, donde el comportamiento del sistema propuesto puede aproximarse desde las tres posturas propuestas por Dennett.

Ejemplo 2.1. *Una computadora que juega al **ajedrez**, de forma que nosotros como oponentes, debemos interpretar su comportamiento para poder ganarle una partida.*

- **Estrategia física.** Desde esta postura, nuestras predicciones se basan en el estado físico de la entidad que queremos predecir; y se logran aplicando nuestros conocimientos, los que sean, de las **leyes naturales**. Es la más simple de las posturas propuestas por Dennett –No requieren más supuestos de nuestra parte acerca de la entidad interpretada.

  Leyes naturales

  Ejemplo 2.2. *Abordar el problema de anticipar la próxima jugada de la computadora que juega ajedrez desde esta postura sería absurdo, pero en principio posible. Lo que no resulta absurdo es nuestro convencimiento de que ninguna pieza del tablero se movera por si sola.*

- **Estrategia de Diseño.** Desde esta postura, nuestras predicciones se basan en la funcionalidad que ofrece la entidad siendo interpretada; y asumiendo funcionará tal y como fue diseñado. Diferentes estrategias de interpretación basadas en el diseño pueden discernirse, pero todas descansan en la noción de **función**, donde el diseño del sistema se descompone en partes funcionales más grandes o más pequeñas, y la predicción descansa en el supuesto de que todas estas partes funcionarán apropiadamente.

  Función

  Ejemplo 2.3. *Si uno conoce exactamente cómo es nuestra computadora que juega ajedrez y su programa, es posible predecir su respuesta diseñada para cualquier movimiento, siguiendo las instrucciones computacionales de su*

programa. Nuestra predicción será verdadera si la computadora funciona tal y como fue diseñada. Es posible hacer predicciones a diferente nivel de abstracción, por ejemplo: al nivel de procedimiento generador de estrategias del programa; o al del procedimiento análisis de consecuencias del programa; o a nivel de los transistores y puertos del hardware, aunque en esta dirección nos acercamos cada vez más a la estrategia física.

Artefactos y objetos naturales

Además de adoptar la postura de diseño para predecir **artefactos**, como nuestro oponente artificial en el ajedrez, también la usamos para aproximarnos a **objetos naturales**, por ejemplo: es racional sembrar trigo antes de lluvias y esperar una buena cosecha.

- **Estrategia Intencional.** Las mejores computadoras que juegan ajedrez se nos han vuelto prácticamente inaccesibles a la predicción desde las estrategias física y de diseño. Se han vuelto **complejas** a tal grado, que resultan inaccesibles desde las estrategias anteriores aún para sus propios diseñadores. Al asumir esta postura, no solo asumimos que (1) la máquina funcionará como fue **diseñada**; sino que (2) el diseño es **óptimo** y que la computadora elegirá el movimiento más **racional**.

  Complejidad

  Ejemplo 2.4. *Es racional de mi parte **creer** que la computadora movera su alfil porque **desea** amenazar a mi torre, ya que **intenta** darme jaque mate.*

2.2.2 Racionalidad y postura Intencional

Racionalidad y postura Intencional

Las predicciones que provee la postura Intencional, descansan en el supuesto de que el agente que está siendo interpretado es racional. Ahora bien, **racionalidad** solo significa aquí, tal y como lo hemos discutido [2], **diseño óptimo** con respecto a una meta, o una jerarquía ponderada de metas (jaque mate, ganar piezas, defensa, etc., en el caso del ajedrez); y a un conjunto de restricciones (las reglas y la posición inicial). Se trata de una noción de **racionalidad acotada**, donde las predicciones toman la siguiente forma: ¿Cual es la decisión más racional para el agente, dadas las metas $X, Y, Z, \ldots$, las restricciones $A, B, C, \ldots$ y la información (incluyendo la desinformación, si es el caso) sobre el estado de las cosas $P, Q, R, \ldots$?

Racionalidad acotada

Predicciones precarias

Observen que las predicciones Intencionales son en extremo **precarias**. No solamente son relativas a postulados sobre metas, restricciones e información que posee el sistema interpretado; y asumen optimización bajo criterios que no siempre son accesibles; sino que además son vulnerables a falsificaciones indetectables desde la estrategia Intencional. La estrategia física nos permite establecer, por si misma, el mal funcionamiento de los sistemas interpretados. La estrategia de diseño nos permite identificar un mal diseño, pero es vulnerable al mal funcionamiento que no puede interpretar. De la misma forma, la estrategia intencional es vulnerable al mal diseño que le es inabordable. El éxito en el uso de la estrategia intencional se basa exclusivamente en que sus predicciones sean verdaderas con suficiente **regularidad**, como para lograr que el método sea práctico.

Éxito y regularidad

2 Ver la discusión sobre hacer lo correcto y las medidas de desempeño, en el capítulo anterior, pág. 5.

Ejemplo 2.5. *Mi predicción de que la computadora me dará jaque mate moviendo su alfil tendrá éxito, en la medida en que regularmente ese sea el caso. De lo contrario, terminaré por asumir que la computadora no juega bien al ajedrez (no tiene un diseño óptimo); y en el caso extremo, que no es racional.*

La cuestión de si los sistemas interpretados tienen o no creencias, deseos e intenciones, no es la adecuada, puesto que la estrategia intencional descansa solo en nuestra **adscripción** de estas actitudes al sistema que está siendo interpretado. Sin embargo, hay que cuidar que los estados Intencionales sean consistentes con sus referentes epistemológicos, por ejemplo:

Adscripción consistente de actitudes

Ejemplo 2.6. *Las creencias no son equiparables al simple almacenamiento en la computadora. En todo caso, tener creencias es equiparable a poseer información en el sentido que esa información se usa para decidir qué hacer. Una pieza de información es una creencia por su función en el razonamiento del agente, no por su forma o implementación.*

Resulta interesante que las predicciones Intencionales, a pesar de ser precarias, funcionen cuando ninguna otra estrategia de interpretación está disponible. Adoptarla es una decisión **pragmática**, en el sentido de que tal decisión no es intrínsecamente correcta o incorrecta. Regresando al ejemplo de la computadora que juega ajedrez, Dennett [47] ejemplifica la adopción de diferentes estrategias de interpretación con base en nuestro rol ante la máquina como lo muestra el Cuadro 2.1

Pragmatismo de la postura Intencional

Rol	Estrategia	Supuestos
Oponente	Intencional	Racionalidad
Programador	Diseño	Funcional
Mantenimiento	Física	Leyes Naturales

Cuadro 2.1: Estrategias de interpretación ante la computadora que juega ajedrez en función de nuestro rol ante ella y los supuestos en que descansa la adopción de cada estrategia. Adaptada de Dennett [47].

La adopción de la postura Intencional no implica que consideremos equivalente nuestras interacciones con la computadora que juega ajedrez y aquellas que tenemos con nuestros perros; o nuestros colegas. En todo caso, puede parecer que su uso conlleva a una **antropomorfismo** inevitable. En ese sentido, Dennett [47] argumenta que es un antropomorfismo conceptualmente inocente, que no impone creencias comunes con los sistemas interpretados, sino categorías como **racionalidad**, **percepción** y **acción**. En todo caso, en virtud de la racionalidad, algunos estados Intencionales parecerían compartidos con nosotros, por ejemplo, nuestra creencia en las verdades lógicas y la normal ausencia de un deseo por auto-destruirse.

Antropomorfismo

2.2.3 Escala de Intencionalidad

Dennett [48] identifica una escala de Intencionalidad, cuyo orden provee una **escala de inteligencia**:

- **Sistemas Intencionales de primer orden**. Sistemas Intencionales con creencias, deseos y otras actitudes proposicionales pero sin creencias ni deseos acerca de sus propias creencias y deseos (sin actitudes proposicionales anidadas).
- **Sistemas Intencionales de segundo orden.** Sistemas Intencionales con creencias, deseos y otras actitudes proposicionales, más creencias y deseos acerca de sus propias creencias y deseos (con actitudes proposicionales anidadas).
- **Sistemas Intencionales de orden** $n > 2$**.** La jerarquía de intencionalidad puede extenderse tanto como sea necesario.

Evidentemente, si estamos interesados en aspectos sociales del comportamiento, es necesario abordar eventualmente una Intencionalidad de orden mayor a uno. Aunque se trata de una escala ascendente de inteligencia, observen que para $n = 3$ el asunto comienza a complicarse aún para nosotros:

Ejemplo 2.7. *Yo creo que Dennett cree que ustedes no creen que él cree lo que yo creo.*

2.2.4 Teorías Intencionales del Comportamiento

Teoría del comportamiento

Las explicaciones Intencionales constituyen una **teoría del comportamiento** del sistema interpretado, y es necesario comparar esta teoría con otras posibles teorías del comportamiento. Primero, es fácil reconocer que nuestras explicaciones y predicciones sobre el comportamiento de animales y humanos, basadas en el **sentido común**, son Intencionales –Asumen racionalidad.

Sentido común

Ejemplo 2.8. *No esperamos que nuestros nuevos conocidos se comporten irracionalmente, y si lo hacen y nuestras predicciones resultan falsas, antes de cuestionar el principio de racionalidad, revisamos la información que poseía el sistema interpretado (no sabía que..., no hablaba español...) o las metas del mismo (el partido de fútbol era más importante que el seminario y menos que la novia...).*

Revisión de estrategia

En casos extremos, cuando el agente parece impredecible desde la estrategia Intencional, entonces la **abandonamos** a favor de la estrategia de diseño o la física. Observen que al abandonar la postura Intencional, también abandonamos el supuesto de que la entidad que está siendo interpretada es un agente racional.

Tautologías

Antes mencionamos que algunos estados Intencionales parecerían compartidos por todos los sistemas Intencionales, por ejemplo, nuestra creencia en las **verdades lógicas**. La extraña imagen de un ratón computando una lista de tautologías, puede evitarse. Asumir que algo es un sistema Intencional es asumir que es racional. Esto es, no se llega a ninguna parte asumiendo que un agente tiene las creencias $p, q, r, \ldots$ al menos que también supongamos que el agente cree lo que se sigue de $p, q, r, \ldots$; de otra forma no habremos ganado ningún poder predictivo. Así que, digamos o no que un animal cree las verdades lógicas, suponemos que sigue las reglas de la lógica.

Ejemplo 2.9. *Seguro que un ratón sigue o cree en el* modus ponens *dado que si le adscribimos las creencias: (1) hay un gato a la izquierda y (2) si hay un gato a la*

izquierda, mejor no voy por la izquierda; entonces podemos predecir basados que el ratón no ira por la izquierda, y de otra forma el ratón no sería "tan racional" como habíamos asumido.

La anterior situación nos lleva a las siguientes preguntas ¿Todas las tautologías lógicas y sus consecuencias son creídas por los sistemas Intencionales? ¿Son los sistemas Intencionales **omniscientes**? Si el sistema fuese perfectamente racional, todas las verdades lógicas aparecerían, pero todo sistema Intencional está racionalmente acotado y por tanto es imperfecto; de forma que no es omnisciente. Peor aún, como lo muestra el ejemplo de ratón (Ejemplo 2.9), no todas las reglas de inferencia del sistema Intencional son válidas. Si descubrimos que el sistema que intentamos interpretar no sigue al pie de la letra el *modus ponens* terminaremos por excluir esta regla de inferencia y eventualmente, en ese proceso, abandonar la estrategia intencional a favor de la de diseño, y por tanto abandonar el supuesto de racionalidad.

Omnisciencia

Esto implica que nuestras teorías Intencionales del comportamiento son inherentemente representaciones **incompletas**. Dado que la postura Intencional asume la racionalidad, cada que mantenemos nuestras explicaciones a ese nivel, dejamos sin explicar algún aspecto de la racionalidad. La Intencionalidad **abstrae** detalles de la racionalidad asumida, innecesarios en la construcción de estas explicaciones. He aquí la fuente de tal incompletez. Por ello, Skinner, citado por Dennett [47], afirma con razón que la Intencionalidad no puede ser la base de una psicología y sugiere buscar **regularidades** puramente mecánicas en las actividades de los sujetos observados. Dennett argumenta que hay poca evidencia a favor de que tales regularidades sean observables en la superficie del comportamiento observado, como no sea bajo condiciones artificiales de laboratorio; y sugiere que buscarlas en el funcionamiento interno de los sistemas observados, cuyo diseño es una aproximación a lo óptimo (en relación con ciertos fines). La táctica más adecuada en esta tarea es:

Incompletez de las teorías Intencionales

Regularidades mecánicas

1. Asumir racionalidad, como un préstamo;
2. Atribuir contenidos a los eventos internos y periféricos del comportamiento observado, asumiendo actitudes proposicionales;
3. Buscar los mecanismos que funcionarían apropiadamente bajo los supuestos anteriores, de forma que el préstamos de racionalidad pueda pagarse al explicar los mecanismos citados.

Este es, en cierta forma, el ***modus operandi*** de la *nouvelle* IA: a partir de agentes caracterizados Intencionalmente, diseñar algoritmos de racionalidad que aproximan alguna forma de comportamiento óptimo. Esto es, la postura Intencional justifica que representemos a nuestros agentes artificiales en términos de creencias, deseos e intenciones; a condición de que formulemos de manera precisa como es que esta representación es usada para decidir qué hacer. Debemos diseñar algoritmos de razonamiento práctico que operen sobre las representaciones Intencionales.

Modus operandi *de la IA*

Dos comentarios finales sobre la incompletez de las teorías Intencionales, ésta no significa que la Intencionalidad sea vacua desde cualquier perspectiva. Por ejemplo, la Teoría de Juegos es intrínsecamente Intencional, pero

como se trata de una teoría formal **normativa** y no de una psicología, no hay nada fuera de lugar en ella. Las predicciones de la Teoría de Juegos aplicadas a humanos son precisas gracias a la garantía evolutiva de que el ser humano es un buen jugador, una forma especial de racionalidad. Un razonamiento similar puede aplicarse a la economía. Las Neurociencias tienen un *modus operandi* similar al de la IA, aunque no buscan construir agentes, sino explicarlos. Sus resultados son explicaciones de la racionalidad de los agentes naturales, no la construcción de agentes artificiales racionales.

Caracter normativo de la Economía y la Teoría de Juegos

2.3 RAZONAMIENTO PRÁCTICO

Cuando pensamos en la racionalidad, solemos hacerlo en términos del **razonamiento teórico** que solemos usar para resolver preguntas, encontrar explicaciones o hacer predicciones. La base de este razonamiento son hechos que ha sucedido y nuestro conocimiento acerca de ellos. De alguna forma este tipo de razonamiento tiene que ver con las normas que regulan las **creencias** de los agentes. Sin embargo, los agentes racionales deben además confrontar el problema de ¿Qué hacer? en un momento dado. Se trata de razonar sobre **acciones** que no han ocurrido, su valor y de qué tan deseables son. El **razonamiento práctico** tiene que ver con las normas que regulan las acciones de los agentes. Aunque si queremos ser precisos, el objeto del razonamiento práctico son nuestras acciones intencionales, así que mientras el razonamiento práctico está orientado a las creencias, el práctico lo está hacía las **intenciones**.

Razonamiento teórico

Razonamiento práctico

Son cuatro los supuestos básicos de la teoría de razonamiento práctico propuesta por Bratman [22], a saber:

- Las intenciones están ligadas a los planes, de hecho son agregados de planes parciales y jerárquicos.
- Somos agentes que planean para contender con nuestra racionalidad acotada y poder decidir ahora que haremos en el futuro.
- Somos agentes racionales acotados, de forma que nuestros planes y su ejecución dependen de cierta deliberación.
- Una intención no es igual que un deseo, aunque ambos tienen roles motivadores, en realidad una intención conlleva compromiso, mientras que un deseo no.

De los cuatro principios, el segundo resulta problemático y justifica el trabajo de Bratman: las intenciones futuras y el compromiso nos llevan a un trilema difícil de resolver. Esto se puede apreciar en el siguiente ejemplo: Asumamos que un agente α tiene la intención i de ir de compras el próximo fin de semana. Entonces α tiene una intención futura. Pero esta no es una suposición inocente, una vez formada esta intención surgen tres problemas:

Trilema de la intención futura

- Objeción metafísica. Cuando i se forma, no controla todas nuestras acciones futuras, pues de ser así i implicaría acción a distancia: una cosa es compromiso y otra cosa es acción a distancia.

- Objeción racional. Una vez que i se forma, no es preciso ni se sigue que i no sea irrevocable. El mundo es dinámico y los agentes no siempre anticipan el futuro del mundo.
- Objeción pragmática. Dadas las dos objeciones anteriores, tenemos que i debería formarse sólo si es racional para α formar i, pero eso es inútil: si ese fuera el caso no tendría porque haber planes a futuro, pero los hay.

Tesis tradicionales

El enfoque tradicional, no Bratmaniano, para resolver este trilema se basa en cuatro tesis principales:

- Tesis metodológica. La prioridad metodológica de la intención de actuar. Esto es, considerar siempre la intención presente sin considerar la intención futura.
- Tesis creencia-deseo. Las acciones intencionales son aquellas que se definen como compatibles con las creencias y deseos del agente.
- Tesis de extensión. Asumiendo que las dos tesis anteriores pueden explicar la intención presente, es posible extenderlas para explicar intenciones futuras.
- Tesis de reducción. Por tanto es posible reducir las intenciones a una combinación adecuada de creencias y deseos.

Bratman argumenta que estas tesis no son suficientes para tratar con las intenciones futuras, y por lo tanto, no resuelven las objeciones del trilema. Asumiendo la tesis de reducción, el modelo BD (Creencias-Deseos) tiene dos aspectos: Uno normativo que dictamina que hace que una intención sea racional; y uno descriptivo, concerniente a qué hace que una acción sea racional. Normativamente, el modelo BD parece adecuado para explicar el papel de las creencias y los deseos en la formación de intenciones, sin embargo un punto importante aquí es la diferencia entre las intenciones futuras y los deseos. Los deseos como las intenciones son pro-actitudes, pero mientras los primeros son potenciales influencias de la conducta, las intenciones conducen la conducta. Esto es, los deseos no implican compromiso, ni explican la planeación a futuro. Las intenciones si.

Es posible extender el modelo BD en el sentido descriptivo, manteniendo su sentido normativo. Dado el supuesto de que los agentes planean, el comportamiento de estos está delineado por un proceso de razonamiento práctico. Este tipo de razonamiento está enfocado a realizar acciones basadas en lo que el agente cree y desea, y tiene dos características importantes:

- La deliberación que consiste en la adopción de intenciones; y
- El razonamiento medios-fines que consiste en la determinación de los medios para satisfacer las intenciones.

Bajo estos dos procesos, las intenciones pueden adaptarse con sus características:

- Pro-actividad. Las intenciones pueden motivar el cumplimiento de metas, son controladoras de la conducta.
- Inercia. Las intenciones persisten, es decir, una vez adoptadas se resisten a ser revocadas. Sin embargo, no son irrevocables. Si la razón por la cual se creo la intención desaparece, entonces es racional abandonar la intención.
- Intenciones futuras. Una vez adoptada una intención, ésta restringirá los futuros razonamientos prácticos, en particular el agente no considerará adoptar intenciones incompatibles con la intención previamente adoptada. Es por ello que las intenciones proveen un filtro de admisibilidad para las posibles intenciones que un agente puede considerar.

2.3.1 Planes

Los planes en tanto que cursos de acción, son intenciones, y en ese sentido comparten las propiedades de estas: poseen inercia, son controladores de la conducta del agente, y sirven como futuras entradas para próximos razonamientos prácticos. Sin embargo, los planes también poseen otras características distintivas.

- Los planes son parciales, no son estructuras completas y estáticas.
- Los planes tienen una estructura jerárquica, contienen razones medios-fines y estas razones tienen un procedimiento ordenado.
- Los planes poseen consistencia interna en el sentido de poder ser ejecutables.
- Los planes son fuertemente consistentes con las creencias del agente.
- Los planes poseen coherencia medios-fines en el sentido de que los sub-planes de un plan son coherentes con los fines del plan.

2.3.2 La tesis de asimetría

Una de las exigencias a los planes, la consistencia fuerte, nos muestra que las creencias y las intenciones mantienen ciertas relaciones. Bratman considera estas relaciones como principios de racionalidad. Quizás la más conocida de estas relaciones es la expresada en la tesis de asimetría:

- Inconsistencia intención-creencia. Es irracional para un agente intentar ϕ y creer al mismo tiempo que no hará ϕ.
- Incompletitud intención-creencia. Es racional para un agente intentar ϕ pero no creer que logrará ϕ.

2.4 ACTOS DE HABLA

Actos de habla

La referencia obligada con relación a los agentes Intencionales comunicativos es la **Teoría de los actos de habla** o actos Comunicativos, propuesta

originalmente por el filósofo del lenguaje, Austin [6], quien observó que no todos los enunciados tienen un significado determinado exclusivamente por su valor de verdad lógico; algunos no pueden clasificarse como verdaderos o falsos, ya que su enunciación constituye la ejecución de una **acción** y por tanto su significado tiene que ver con el resultado de esa acción. Por ejemplo (adaptado de Perrault y Allen [144]), dados los siguientes enunciados:

1. Pásame la sal.
2. ¿Tienes la sal?
3. ¿Te queda la sal cerca?
4. Quiero sal.
5. ¿Me puedes pasar la sal?
6. Juan me pidió que te pidiera pasar la sal.

observen que su significado nada tiene que ver con su valor de verdad; y si está en relación con el efecto que consiguen en el medio ambiente y en los otros agentes: Conseguí la sal o no. Por lo anterior, Austin los llamó enunciados **performativos** (del inglés *performatives*). Estos enunciados tienen una Intencionalidad asociada: todas pueden interpretarse como peticiones de la sal; aunque (3) podría en extremo interpretarse como una solicitud de información, dando lugar al mal chiste –Si, me queda cerca.

Enunciados performativos

Austin identificó tres **clases** de Actos de Habla:

Clases de Actos de Habla

- **Locución.** Es el acto de decir algo, por ejemplo, al pronunciar una secuencia de palabras de un vocabulario en un lenguaje dado, conforme a su gramática.
- **Elocución.** Es el acto que se lleva a cabo al decir algo: promesas, advertencias, informes, solicitudes, son algunas elocuciones. Una enunciación tiene una fuerza elocutoria F si el agente que la enuncia intenta llevar a cabo la elocución F con ese acto. Los verbos que nombran a las elocuciones son llamados **verbos performativos**. De aquí en adelante asumiremos que los actos de habla se corresponden con las elocuciones.
- **Perlocución.** Es el acto que se lleva a cabo por decir algo. En (6) Juan puede, vía su petición, convencer a los otros agentes de que le pasen la sal, y hacerse finalmente con ella. El éxito de una perlocución está fuera del alcance del agente emisor, por ejemplo, nada garantiza que los otros le pasen la sal a Juan.

Searle [166] [3] argumenta que las elocuciones puede definirse mediante las **condiciones de necesidad** y suficiencia para la ejecución exitosa del acto de habla. En particular, señala que el agente emisor ejecuta la elocución si y sólo si intenta que el receptor reconozca su intención en ese acto, al reconocer la fuerza elocutoria del mismo. Esto relaciona el trabajo de Austin y Searle con el de Grice [77] sobre comunicación y reconocimiento de Intenciones.

Condiciones de necesidad

3 Paradoja: ¡el autor de la célebre Paradoja del Cuarto Chino!

2.4.1 Actos de habla y estados Intencionales

Searle [165] replantea la cuestión medieval sobre la Intencionalidad en los siguientes términos ¿Cual es la relación exacta entre los estados Intencionales y los objetos o asuntos a los que apuntan o son acerca de? Esta relación no es tan simple. Consideren el siguiente ejemplo: Puedo creer que el rey de Francia es un enano, aún cuando, sin que yo lo sepa, no hay monarquía en ese país. Es decir, puedo tener un estado Intencional sobre un contenido para el cual no hay referente, e incluso es inexistente. La respuesta de Searle es que los estados Intencionales representan objetos y asuntos del mundo, en el mismo sentido que los actos de habla los representan. Los puntos de **similitud** entre estos conceptos pueden resumirse como sigue:

Similitudes

- Los Actos de Habla y los Estados Intencionales tienen una **estructura** similar, resaltando la distinción entre, por una parte, la **fuerza elocutoria** de los Actos de Habla y la **actitud** de los Estados Intencionales; y por otra, el **contenido proposicional** de ambos.
- Las distinciones en la **dirección de ajuste** entre palabra y realidad (Ver Figura 2.2), familiares en los actos de habla, también aplican en el caso de los estados Intencionales. Se espera que los **actos asertivos** (afirmaciones, descripciones, aserciones) de alguna forma tengan **correspondencia con el mundo** cuya existencia es independiente. En la media que lo logran son o no verdaderos. Pero los **actos directivos** (ordenes, comandos, peticiones) y los **comisorios** (promesas, ruegos, juramentos) no se supone que tengan correspondencia con la realidad independientemente existente, sino por el contrario causan que el mundo corresponda con lo expresado. Y en ese sentido, no son verdaderos, ni falsos; sino que se son exitosos o fallan, se mantienen o se rompen, etc. Y aún más, puede haber casos de no direccionalidad, por ejemplo cuando felicito a alguien.

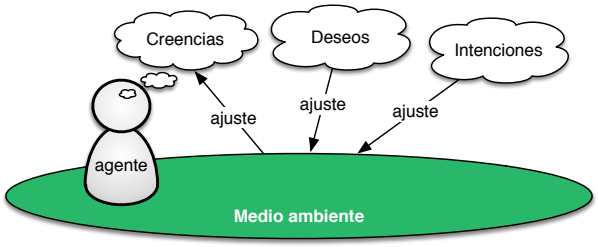

Figura 2.2: Direccionalidad en el ajuste entre los actos de habla y el medio ambiente del agente: el agente mantiene sus aserciones, afirmaciones, etc., consistentes con el medio ambiente; pero sus ordenes, peticiones, etc., y sus promesas, juramentos, etc., buscan que el medio ambiente sea consistente con sus contenidos proposicionales.

Lo mismo sucede en los estados Intencionales (Ver Figura 2.3). Si mis creencias fallan, es culpa de mis creencias, no del mundo y de hecho puedo corregir la situación en un proceso de **mantenimiento de Creencias**. Pero si mis intenciones o mis deseos fallan, no puedo corregirlos en ese sentido. Las creencias, como las aserciones, son falsas o verda-

Mantenimiento de creencias

deras; las intenciones como los actos directivos y los comisorios, fallan o tienen éxito.

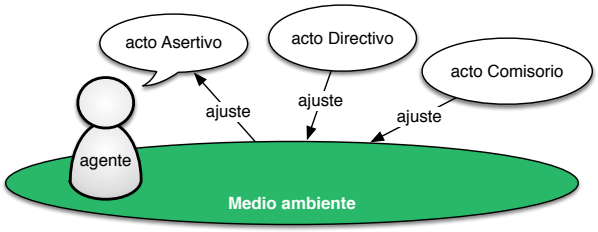

Figura 2.3: Direccionalidad en el ajuste entre los estados Intencionales BDI y el medio ambiente del agente: el agente mantiene sus creencias consistentes con el medio ambiente; pero los deseos y las intenciones buscan que el medio ambiente sea consistente con sus contenidos proposicionales.

- En general, al ejecutar un acto de habla ilocutorio con contenido proposicional, expresamos cierto estado Intencional con ese contenido proposicional, y tal estado Intencional es la **condición de sinceridad** (Ver Figura 2.4) de ese tipo de acto de habla. Por ejemplo, si afirmo que ϕ estoy expresando que creo que ϕ; si prometo que ϕ estoy expresando que intento que ϕ; si ordeno que ϕ estoy expresando mi deseo por ϕ. Esta relación es interna en el sentido de que el estado Intencional no es un acompañante de la ejecución del acto de habla, tal ejecución es necesariamente la expresión del estado Intencional. Esto se ejemplifica por la paradoja de Moore: No se puede afirmar "está nevando, pero creo que no está nevando", "le ordeno que deje de fumar, pero no quiero que deje de fumar", etc. Las declaraciones constituyen una excepción a todo esto, pués no tienen condiciones de sinceridad ni estado Intencional. Y por supuesto, siempre es posible mentir, pero un acto de habla insincero, consiste en ejecutar el acto para expresar un estado Intencional que no se tiene; por lo cual lo dicho aquí se mantiene.

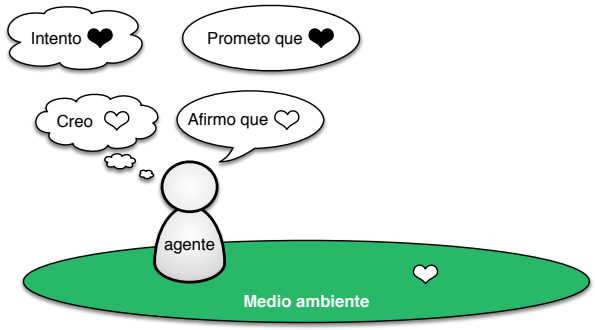

Figura 2.4: Los actos de habla ilocutorios expresan estados Intencionales que son a su vez la condición de sinceridad de lo expresado.

- El concepto de **condiciones de satisfacción** aplica en ambos casos. Por ejemplo, una afirmación se satisface si y sólo si es verdadera; una promesa se satisface si y sólo si se cumple; una orden se satisface si y sólo si se obedece; etc. Lo mismo para los estados Intencionales: mi creencia se satisface si las cosas son como creo; mis deseos se satisfacen si

son logrados; mis intenciones se satisfacen si son llevadas a cabo. La noción de satisfacción parece natural y aplica siempre que haya una dirección de ajuste presente. Lo relevante aquí es que el acto de habla se satisface si y sólo si el estado Intencional por él expresado se satisface. La excepción aquí se da cuando el estado Intencional se satisface por causa ajenas a la ejecución del acto de habla, por ejemplo: prometí hacerme de una Gibson Les Paul y ¡me la regalaron! El estado Intencional, que intentaba hacerme de una guitarra, se satisface; pero no así mi promesa.

Condiciones de satisfacción

Searle [164] extendió el trabajo de Austin, en su libro *Speech Acts*, identificando diversas condiciones necesarias para la ejecución exitosa de los actos de habla. Consideren el caso (tomado del texto Wooldridge [186]) de una solicitud de un emisor identificado como hablante, a un receptor identificado como oyente, para ejecutar una acción:

- **Condiciones normales de E/S.** Especifican que el oyente es capaz de escuchar la solicitud, que el acto se lleva a cabo en circunstancias normales, por ejemplo, no en una película, ni en una obra de teatro.
- **Condiciones preparatorias.** Establecen que debe ser cierto en el mundo para que el hablante elija correctamente el acto de habla. En este caso, el oyente debe ser capaz de ejecutar la acción, y el hablante debe creer que el oyente es capaz de ejecutar la acción. También, no deber ser obvio que el oyente hará la acción de cualquier forma.
- **Condiciones de sinceridad.** Distinguen las ejecuciones sinceras del acto de habla. Una ejecución insincera de la solicitud podría ocurrir si el hablante no quiere realmente que la acción se lleve a cabo.

Una taxonomia de los actos de habla es una taxonomia de los estados Intencionales. Aquellos estados Intencionales cuyo contenido son proposiciones completas, las llamadas **actitudes proposicionales**, se dividen convenientemente en aquellas que ajustan el estado mental al mundo, el mundo al estado mental y las que no tienen dirección de ajuste. Observen que no todas las intenciones tienen contenido proposicional completo, algunos estados son solo acerca de objetos.

Tipos de performativa

Usando **performativas**, Searle propone una clasificación sistemática de los actos de habla:

- **Representativas.** Cuando comprometen al emisor a la verdad de la proposición expresada. El caso paradigmático es informar algo o afirmarlo.
- **Directivas.** Se usan como un intento del emisor para que el receptor haga algo, por ejemplo, pedir algo.
- **Comisivas.** Crean el compromiso para el emisor de tomar un curso de acción dado, por ejemplo, prometer algo.
- **Expresivas.** Expresan algún estado psicológico, por ejemplo agradecimiento. El caso paradigmático es dar las gracias.

- **Declarativas.** Efectúan algún cambio en el estado institucional de las cosas. El paradigma de este caso es la declaración de guerra.

2.4.2 Retomando los estados Intencionales

La visión de intencionalidad presentada en la sección anterior, permite ver soluciones a muchos problemas, tradicionales y sorprendentes, sobre los estados mentales Intencionales:

- ¿Qué es un estado Intencional? no es una cuestión ontológica, o no debería serlo, ya que lo que hace que un estado mental sea Intencional no es su categoría ontológica, sino sus **propiedades lógicas**. Si estos estados se realizan en una red de neuronas, modificaciones de un ego Cartesiano, imágenes que flotan en la mente, palabras en nuestro pensamiento, es irrelevante en relación con dichas propiedades lógicas. Es irrelevante cómo se realiza un estado Intencional, mientras que tal realización sea la realización de su Intencionalidad.

  Las propiedades lógicas de los estados Intencionales se deben a que estos, al igual que las entidades lingüisticas, son **representaciones**. Lo que necesitamos saber acerca del estado Intencional es:

  1. Cuales son sus condiciones de satisfacción;
  2. Bajo que aspectos se representan estas condiciones en el contenido representativo; y
  3. El modo psicológico del estado en cuestión – creencia, deseo o intención.

  El tercer aspecto determina la dirección de ajuste del estado Intencional entre el contenido representativo y las condiciones de satisfacción; y el segundo, implica al primero.

- ¿Qué es un objeto Intencional? Desde esta perspectiva, los objetos intencionales no tienen un status ontológico especial, son simplemente objetos acerca de los cuales se da un estado Intencional. Si digo que Adán admira a Obama, Obama es el objeto de su admiración, Obama el hombre, no una idea sobre Obama, ni una sombra sobre él, ni una red neuronal artificial sobre él. Precaución: un estado Intencional tiene un contenido representativo, pero no está dirigido ni es acerca de su contenido representativo.

- Si todo esto es cierto, es un error pensar en las creencias como una relación entre dos términos: el agente y una proposición. Deberíamos decir que la proposición no es el **objeto** de la creencia, sino su **contenido**. El contenido de la creencia DeGaulle era francés, es la proposición que DeGaulle era francés. Pero esa proposición no es a lo que apunta la creencia o es acerca de. La creencia es acerca de DeGaulle y lo representa como francés. La relación es entre el estado Intencional y las cosas representadas por él.

- La única relación entre intensionalidad e Intencionalidad es que algunos enunciados Intencionales son intensionales (referencialmente opa-

Intensionalidad

cos). La **intensionalidad** se refiere a una propiedad de ciertas clases de enunciados que fallan ciertas pruebas de **extensionalidad** como lo son la substitución por identicos y la generalización existencial. Oraciones como "Adán creee que el rey Arturo arruinó a sir Lancelot" es intensional en el sentido de que tiene al menos una interpretación donde puede usarse para hacer un enunciado donde no se permite generalización existencia sobre la parte referida después del "cree que"; y no permite la substitución de expresiones con la misma referencia *salva veritate*. La respuesta expuesta hasta ahora da cuenta de ello. Searle no abunda en ello más allá de: la frase en cuestión es usada para hacer una afirmación acerca de un estado Intencional, la creencia de Adán, y puesto que un estado Intencional es una representación, esta afirmación es una representación de una representación; y por tanto, el valor de verdad del enunciado dependerán de las características de la representación siendo representada, en este caso las características de la creencia de Adán, y no en las características de los objetos o asuntos representados por la creencia de Adán. La creecia de Adán solo puede ser verdadera si existen las personas Arturo y Lancelot y el primero arruinó al segundo. Pero la afirmación de que Adán cree que el rey Arturo... tiene una intepretación que puede ser cierta aunque ninguna de las condiciones anteriores lo sean. En este caso mi afirmación sobre su creencia, no es una representación de una representación, sino una **presentación** de una representación, puesto que estoy presentando el contenido de su creencia, sin **comprometerme** al valor de verdad de su creencia.

- Bajo los supuestos presentados, el deseo de ejecutar una acción es precisamente una representación de la acción a ser ejecutada. Es debido a ello que mis deseos por ejecutar una acción puden ser **causales** de la ejecución de tal acción.

- La Intencionalidad ha sido aplicada a estados mentales, mientras que la intensionalidad a enunciados y otras entidades lingüisticas. Pero puede haber estados mentales intensionales, como mi creencia de que Adán cree que el rey Arturo arruinó a Lancelot. Esto no implica que la creencia de Adán sea también intensional, de hecho es totalmetne extensional. Y de hecho los enunciados y otras entidades lingüisticas son Intencionales! ¿Cómo puede esto ser?

Estados vs Actos

Finalmente, existe una **divergencia** evidente entre los estados Intencionales y los actos de habla: los primeros son estados y los segundos son acciones que necesitan producirse.

2.5 INTENCIONALIDAD, AGENCIA Y COMPUTACIÓN

McCarthy [125] fue uno de los primeros en argumentar a favor de la adscripción de estados mentales a máquinas, distinguiendo entre la legitimidad y el pragmatismo de tal práctica. En su opinión, adscribir creencias, deseos, intenciones, conciencia, o compromisos a una máquina o a un programa

Uso	Actitudes proposicionales
Interactivas	Percepciones, informaciones, comandos, peticiones, normas.
Representacionales	Creencias, hipótesis.
Conativas	Deseos, metas, impulsos, demandas, intenciones, compromisos.
Organizacionales	Métodos, tareas.

Cuadro 2.2: Clasificación de las actitudes proposicionales de acuerdo a su utilidad en el diseño de un agente. Adaptada de Ferber [60].

de cómputo es **legítimo** cuando tal adscripción expresa la misma información sobre la máquina, que expresaría sobre una persona. Es útil cuando la adscripción ayuda a entender la estructura de la máquina, su comportamiento pasado y futuro, o cómo repararla o mejorarla. Quizá **nunca** sea un requisito lógico, aún en el caso de los humanos, pero si queremos expresar brevemente lo que sabemos del estado de la máquina, es posible que necesitemos de cualidades mentales como las mencionadas, o isomorfas a ellas. Es posible construir teorías de las creencias, el conocimiento y los deseos de estas máquinas, en una configuración más simple que la usada con los humanos; para aplicarlas posteriormente a los humanos. Esta adscripción de cualidades mentales es más directa para máquinas cuya estructura es conocida, pero es más **útil** cuando se aplica a entidades cuya estructura se conoce muy parcialmente.

Tradicionalmente, tres actitudes proposicionales son consideradas para modelar **agentes racionales BDI** (*Belief-Desire-Intention*): Creencias, deseos e intenciones. El Cuadro 2.2 presenta una categorización más extensa de actitudes proposicionales y su uso en el modelado de los agentes racionales [60].

Otros **argumentos computacionales** para el uso de la postura intencional han sido formulados por Singh [176]:

- Las actitudes proposicionales nos son familiares a todos, diseñadores, analistas de sistemas, programadores y usuarios;
- La postura provee descripciones sucintas del comportamiento de los sistemas complejos, por lo que ayudan a entenderlos y explicarlos;
- Provee de ciertas regularidades y patrones de acción que son independientes de la implementación física de los agentes;
- Un agente puede razonar sobre si mismo y sobre otros agentes adoptando la postura intencional.

Históricamente, las representaciones Intencionales se han usado en arquitecturas y lenguajes de programación orientados a agentes. A continuación revisaremos en detalle estas dos aproximaciones.

2.5.1 Arquitecturas BDI

Los agentes racionales deben llevar a cabo un razonamiento medios-fines para producir cursos de acción alternativos, que serán ponderados en un proceso deliberativo. En ambos casos, debemos tomar en cuenta que la racionalidad de los agentes es acotada –Son incapaces de ejecutar cómputos arbitrariamente grandes en tiempo constante. Tal escenario se complica aú más si consideramos que el medio ambiente es dinámico [23]. En tales circunstancias los planes son centrales.

Los procesos y representaciones propias del razonamiento práctico pueden encapsularse en un tipo de arquitectura de agentes, conocido como **arquitecturas BDI** (Ver Figura 2.5). Fundamentalmente se trata de agentes con estado, donde estos incluyen **representaciones** directas para las creencias (B), los deseos (D) y las intenciones (I) de un agente, por ejemplo, una lógica de primer orden para las creencias. Los planes como tales, están representados en una librería que se considera incluida en las creencias del agente. Esto es, el agente cree que ciertos cursos de acción pueden lograr ciertos efectos, dadas ciertas condiciones. Los planes adoptados por el agente, son lo que conocemos como intenciones.

Arquitecturas BDI

Representaciones

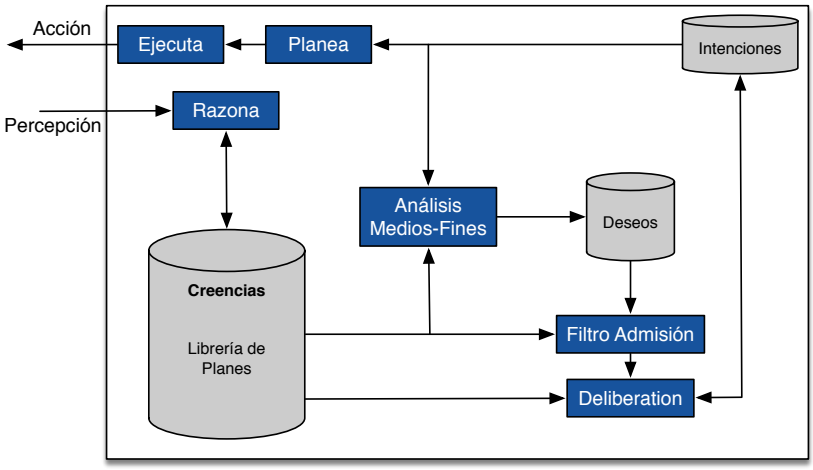

Figura 2.5: Una arquitectura BDI adaptada de Bratman, Pollak e Israel [23].

En esta arquitectura BDI, Bratman identifica cuatro **procesos** en el razonamiento práctico:

Procesos

ANÁLISIS MEDIOS-FINES. Este proceso debe proponer cursos de acción (planes) alternativos para completar los planes del agente. Es un proceso que revisa las fines actuales del sistema (intenciones actuales), para decidir que planes (intenciones futuras) podrían ser adoptados para proporcionar los medios de dichos fines. Este proceso debe ejecutarse cada vez que una intención corra el riesgo de volverse incoherente en términos medios-fines. La salida de este proceso es un conjunto de planes, todos ellos medios de los actuales fines del agente. Se trata de generar opciones de comportamiento, en un proceso **guiado por las metas** del agente.

Comportamiento guiado por metas

ANÁLISIS DE OPORTUNIDADES. No todas las opciones que el agente debe considerar provienen del análisis medios-fines. Los cambios en el am-

biente pueden cambiar las creencias del agente, y estos cambios pueden producir opciones de comportamiento que no son necesariamente medios del algún fin. Este es un componente **reactivo**, se trata de generar opciones de comportamiento guiados por el ambiente (o de manera más precisa, por los datos que reflejan cambios en el ambiente).

Comportamiento guiado por datos

FILTRADO. Las opciones generadas por los análisis anteriores deben ser filtradas para eliminar aquellas que son incompatibles con las intenciones actuales del agente; sin embargo el proceso de filtrado puede pasarse por alto de ser necesario. Esto refleja la tensión entre el carácter revocable de las intenciones y su inercia característica. Diferentes filtros pueden ser propuestos, reflejando diferentes formas de **reconsideración**; pero independientemente de su forma, deben ser procesos computacionales eficientes –Menos costosos que los procesos de análisis y deliberación.

Reconsideración

DELIBERACIÓN. Este proceso **pondera** las opciones sobrevivientes al filtrado, para incorporarlas como intenciones del agente. Este suele ser el caso cuando dos o más opciones sobrevivientes son incompatibles, se deberá elegir la "mejor" de ellas para promoverla como intención.

Ponderación

La idea central es que los planes no solo producen comportamiento, sino que tienen **roles funcionales** en estos procesos, haciendo el computo más tratable en dos aspectos: Son una entrada en el análisis medios-fines, que establece de manera clara y precisa el propósito del cómputo; y son una entrada en el proceso de filtrado, reduciendo el alcance de la deliberación al limitar el conjunto de opciones a considerar.

Roles funcionales de los planes

Asumimos que el agente está **comprometido** a hacer lo planeado. Este compromiso es complejo, por ejemplo: Si intento tener una cita a las 10, no debería preocuparme por reconsiderar esta decisión continuamente, sino en cómo llegar a la cita. Esto pasa por no intentar llevar a cabo actividades incompatibles con mi cita y asumir que estaré ocupado a las 10, y esto debe tomarse en cuenta en el resto de mis decisiones.

Compromiso

Los planes conducen el análisis medios-fines; proveen restricciones sobre qué opciones debemos considerar seriamente; influencian las creencias sobre las cuales haremos nuevos razonamientos prácticos.

Los planes deben ser **consistentes** internamente y con las creencias. Los planes que no se crean aplicables deben ser reconsiderados o abandonados, aunque deben ser razonablemente estables para poder contender con la racionalidad acotada –Deben ser relativamente resistentes al abandono y la reconsideración.

Consistencia

La racionalidad acotada implica recursos computacionales y conocimiento limitados. Por ello, los planes deben ser **parciales**. Planes muy detallados sobre futuros distantes, suelen ser de poca utilidad. Los planes pueden ser parciales estructuralmente y en el tiempo. Aunque parciales, los planes deben ser coherentes en términos medios-fines.

Parcialidad

La arquitectura delineada por Bratman, Pollak e Israel [23] se caracteriza por su proceso de filtrado de opciones, que reduce el costo del razonamiento práctico que un agente debe llevar a cabo. El proceso de filtrado se basa en el papel funcional de los planes en el razonamiento práctico. Otros procesos

de la arquitectura, como los análisis y la deliberación, son más comunes en la Inteligencia Artificial tradicional. Algunas arquitecturas BDI reportadas en la literatura incluyen el venerable PRS [74, 108] y su revisión multiagente, conocida como dMARS [50]; TileWorld [147], el más cercano a lo aquí expuesto; InteRRapt [63], otra de las primera arquitectura multiagente. Y luego una serie de arquitecturas basadas en Java como JAM [103], Jack [28], Jadex [146] y el BDI modular [136].

2.5.2 Comunicación

Comunicación y computación

La comunicación es considerada como un tema central en Ciencias de la Computación, particularmente en el contexto de **sistemas concurrentes** donde se han desarrollado diferentes formalismos para representar sus propiedades, por ejemplo, en el trabajo de Hoare [100]. Desde la perspectiva de la ingeniería de software, la comunicación en los sistemas distribuidos, nos hace pensar en **protocolos** –especificaciones modulares, potencialmente reusables, de la interacción entre dos o más componentes. La idea es que el protocolo hace referencia de manera abstracta a una serie de roles que los componentes del sistema pueden adoptar y se verifica que la interacción entre los componentes sea conforme a dichos roles. Diversos **formalismos** para la especificación de protocolos han sido propuestos en la literatura: redes de Petri [133], diagramas de secuencia UML [137], etc.

Protocolos

Formalismos

Ejemplo 2.10. *La Figura 2.6 muestra la especificación de parte de un protocolo de compra venta. El formalismo usado es una máquina de estado finito. Hay dos roles en el protocolo, un comprador (c) y un vendedor (v). Las transiciones están etiquetadas con mensajes: el vendedor envía una oferta al comprador; éste puede aceptar o rechazar la oferta; después de aceptar, el vendedor envía un mensaje de actualización a comprador.*

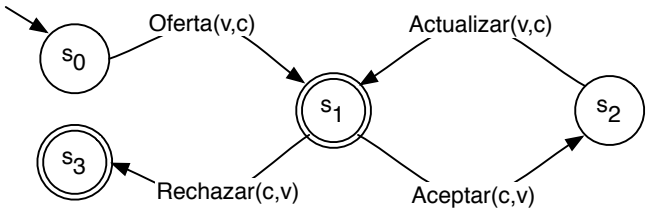

Figura 2.6: Un protocolo de actualización de oferta entre dos agentes: un vendedor (v) y un comprador (c).

Sin embargo, la autonomía y la heterogeneidad de los SMA hace que los protocolos entre agentes resulten particulares. Hay varias consideraciones a tomar en cuenta: Primero, en protocolos como el del ejemplo 2.10 ¿Cómo podemos garantizar que los agentes actuarán de **conformidad** sin atentar contra su **autonomía**? El vendedor no debería tener control sobre los posibles compradores, por ejemplo, si el vendedor puede obligar a todos los posibles compradores a hacer una oferta, pués el protocolo deja de ser realista. En el caso SMA, los protocolos no deben sobre restringir las interacciones entre los agentes. Segundo, debido a la autonomía, los agentes son la **unidad lógica**

Conformidad vs autonomía

Agente como unidad lógica

de toda posible distribución del sistema, independientemente de las posibles consideraciones de viabilidad física o de eficiencia.

Semántica

Hay un elemento faltante en la especificación de la Figura 2.6, al menos desde el punto de vista de los SMA ¿Cual es el **significado** de los mensajes en el mundo real? Usualmente, hacer una oferta para comprar un bien implica un compromiso social; mientras que actualizar la oferta significa que un nuevo compromiso se ha adoptado en lugar del original. Estas cuestiones no están presentes en los protocolos especificados como un flujo de mensajes. En algunas ocasiones parece que este significado de los mensajes no está presente en el protocolo especificado, y sin embargo, el sistema funciona. Esto suele deberse a que hay un acuerdo de diseño sobre como interpretar y procesar los mensajes, a riesgo de sobre restringir las interacciones entre los agentes.

Ingeniería de Software

Flexibilidad

Verificación

Entonces ¿Cual es el modelo de protocolo de comunicación adecuado para los agentes? Esto depende de los criterios de evaluación considerados. Desde la perspectiva de la **ingeniería del software**, los protocolos deberían estar especificados usando abstracciones de alto nivel con un significado pretendido claro. Algunas propiedades deseables incluyen: facilidad de modificación, de entendimiento y composición. Deberían promover el acoplamiento flexible de los componentes del sistema. Tomando en cuenta la **flexibilidad**, los protocolos SMA deberían permitir facilitar su adopción. Dada la naturaleza de los agentes, un protocolo no debería restringir más allá de lo estrictamente necesario para garantizar el funcionamiento correcto del sistema. Finalmente, los protocolos deberían ser **verificables** a partir de su definición precisa y la información de la que disponen los agentes participantes. Como podrán suponer, desde una perspectiva Intencional, abordaremos la comunicación desde la Teoría de los **Actos de Habla** [6, 164] revisados en la sección 2.4.

Agentes y comunicación

En el contexto de los Sistemas Multi-Agentes (SMA), algunos autores consideran que el término agente social es equivalente al de **agente comunicativo**, por ejemplo Huns y Stephens [106] afirma –Los agentes comunican para satisfacer mejor sus metas o las metas de la sociedad en la cual están inmersos. De alguna manera, éste es un punto de vista aceptado por Russell y Subramanian [162], Wooldridge y Jennings [187], e incluso por Ferber [60], quién define a los agentes comunicativos como un agente particular, aquel cuyas habilidades solo incluyen acciones de comunicación. Genesereth y Ketchel [69] van más lejos al definir agencia como la habilidad de intercambiar información usando un lenguaje de comunicación orientado a agentes (ACL).

ACL

Los **lenguajes de comunicación para agentes**, como KQML [62] y FIPA ACL [4], son lenguajes de comunicación de alto nivel, basados en las performativas de los actos de habla. Se sitúan en una capa lógica por arriba de los protocolos de transporte como TCP/IP ó HTTP. FIPA ACL incluye un modelo semántico preciso, expresado en una lógica modal de primer orden con igualdad. Esta semántica hace posible describir actos comunicativos y sus efectos Intencionales de manera extremadamente precisa. El problema es que tal semántica es tan expresiva, que el receptor no pueden en lo ge-

4 Aquí pondremos énfasis en KQML, por la mera razón de que es el ACL utilzado por Jason. Una especificación de FIPA puede encontrarse en http://www.fipa.org/repository/aclspecs.html

neral deducir la intención del emisor. Algunas estrategias de simplificación han sido propuestas, por ejemplo **políticas de conversación**. KQML tiene una semántica más simple y puede aplicarse fácilmente en aplicaciones con dominios razonablemente restringidos. La liga entre actos de habla y teorías de agente, y su arquitectura, se derivan del hecho de que los actos de habla se perciben como cualquier otra acción de los agentes. Jason utiliza KQML para comunicar.

Políticas de conversación

Castelfranchi [32] argumenta que la comunicación puede verse bien como un **instrumento** para la acción social; o como una **acción social** encaminada a modificar las creencias de otro agente. Observen que, como instrumento de la acción social, la comunicación puede darse en interacciones agresivas o cooperativas [30]. En cualquier caso, la comunicación puede verse como la típica meta social, puesto que su resultado Intencional concierne el estado mental de otro agente.

Comunicación y acción social

Cuando hablamos de comunicación Intencional, nos referimos a las dos acepciones de término intencionalidad que hemos visto en el curso: Por un lado, la **escala de Intencionalidad** que Dennett [48] plantea en *Intentional Systems in Cognitive Ethology: the Panglosian paradigm defended* aplica a los actos comunicativos. En particular, él aplica su escala para presentar una hipótesis plausible sobre la naturaleza de los actos comunicativos entre un grupo de monos. Supongan que el mono Chita lanza una alarma "leopardo". Si los monos fuesen sistemas Intencionales de cuarto orden, he aquí una hipótesis plausible sobre este acto: Chita quiere que Zira crea que Chita quiere que Zira crea que hay un leopardo. El tercer orden da lugar a: Chita quiere que Zira crea que Chita quiere que Zira corra a los árboles. Observen el cambio de una comunicación declarativa a una imperativa. La versión de segundo orden sería: Chita quiere que Zira crea que hay un leopardo. En este caso no suponemos que la acción de Chita involucre ningún reconocimiento por parte de Zira del rol de Chita en la situación de interacción. La versión de primer orden es muy simple: Chita quiere que Zira suba a los árboles. Por otro lado, la comunicación es **acción** como lo expesan los actos de habla.

Comunicación e Intencionalidad

2.5.3 Programación Orientada a Agentes

La presentación que Shoham [168] hace del paradigma de **Programación Orientada a Agentes**, es quizá la mejor explicación del interés computacional en los estados Intencionales. Como el nombre lo sugiere, este paradigma es propuesto como una especialización de la Programación Orientada a Objetos, al menos bajo su noción original, tal y como aparece en el trabajo sobre Actores [96]. En este sentido, los objetos pasarían a ser **agentes**, cuya característica principal es que su estado es Intencional, son sujetos de creencias, deseos, intenciones, compromisos, etc. Estos agentes pueden comunicarse con otros agentes para informar algo, solicitar algo, ordenar algo, etc., es decir, son capaces de ejecutar **actos de habla** [164]. Ya hemos revisado algunos argumentos a favor de tal posición, pero ¿Qué necesitamos computacionalmente para justificar el uso de esta terminología pseudo-mental? A saber:

- Una teoría precisa y de semántica clara, sobre cada categoría mental. Es deseable que la teoría tenga correspondencia con el sentido común de estas categorías (!No notación, sin denotación! [126]);
- Una demostración de que los componentes de la máquina, o programa, obedecen esta teoría; y
- Una demostración de que la teoría formal juega un papel no trivial en el análisis y diseño de la máquina (!No notación, sin explotación! [168]).

La computación distribuida provee un claro ejemplo de lo que Shoham requiere. Los investigadores de esta área se encontraron con que el razonamiento intuitivo sobre los protocolos de distribución, incluía normalmente frases como: "El procesador *A* no sabe aún que la información se está respaldando, pero el procesador *B* sabe que *A* no lo sabe! *B* no enviará el siguiente mensaje". Buscando formalizar tales explicaciones, Adoptaron las lógicas modales Del conocimiento [99] donde "saber" se formaliza como un operador de la **lógica modal** S5. Esto, como veremos más adelante, tiene algunos efectos contra intuitivos como que el saber está cerrado bajo la implicación (sabemos todas las consecuencias lógicas de lo que sabemos); y es introspectivo positiva y negativamente (si sabemos algo, sabemos que lo sabemos; y en caso contrario, sabemos que no lo sabemos). Evidentemente estas propiedades del saber son demasiado fuertes para las capacidades de cómputo de un agente artificial, aunque no lo son para el dominio del cómputo distribuido.

Lógica modal

Mientras que la teoría formal de la categoría mental debería corresponderse con el uso de sentido común de la categoría, esta correspondencia no será exacta. Por ejemplo, "saber" como un operador modal S5, es adecuado para representaciones basadas en cláusulas proposicionales, donde existen procedimientos de decisión lineales en el tiempo. Si queremos usarlo para razonar sobre protocolos de criptografía basados en teoría de números, entonces la elección no es adecuada. Pero esto sólo lo podemos saber si la teoría de "saber" está claramente definida.

Para ello, la semántica de operadores como "saber" se especifica usando el estándar de mundos posibles [104]. En el contexto de la Programación Orientada a Agentes, mundo posible sería una posible configuración global del sistema. Estas semánticas basadas en el estado de la máquina, o el programa, han permitido avances significativos en el área.

Lo ideal es que todo este trabajo de formalización y diseño, sirva para probar ciertas propiedades sobre los protocolos distribuidos de cómputo. La lógica del conocimiento no es indispensable en este caso, pero substituir los enunciados sobre el "saber" de los procesadores, por un vocabulario basado en el estado de estos, puede resultar en expresiones más complejas y por ende, difíciles de entender.

En el ejemplo anterior hicimos uso de la postura Intencional para razonar acerca de un caso de procesamiento computacional distribuido. El uso que haremos en este curso de los estados Intencionales es diferente: no los utilizaremos únicamente para analizar las entidades a las que atribuimos creencias, deseos, intenciones, etc., sino que los utilizaremos para diseñar e implementar sistemas computacionales. Los estados u operadores Intencionales formarán parte de nuestros lenguajes de programación.

Concepto	OOP	AOP
Unidad básica	Objeto	Agente
Estado	No restringido	Creencias, Deseos, Intenciones...
Cómputo	Mensajes y métodos	Paso de mensajes y métodos
Tipos de Mensajes	No restringido	Informes, Solicitudes, Promesas...
Restricciones	Ninguno	Honestidad, Coherencia...

Cuadro 2.3: Comparación entre la Programación Orientada a Agentes (AOP) y la Orientada a Objetos (OOP)[168].

La idea es entonces que en la Programación Orientada a Agentes, los agentes son como los objetos de la Programación Orientada a Objetos (en su vertiente actores), cuyos estados son Intencionales. Una computación en este paradigma está relacionada con los agentes informado, requiriendo, ofertando, aceptando, rechazando, compitiendo, y ayudándose. El Cuadro 2.3 resume esta comparación.

Un sistema completo de programación orientada a agentes, debería contar con:

- Un lenguaje **formal** restringido, de sintaxis y semántica claras, para describir los estados Intencionales;
- Un lenguaje de programación **interpretado** para definir los programas de agentes; debe ser fiel a la semántica de los estados Intencionales; y
- Un "agentificador" que convierta entidades neutras en agentes programables.

2.6 LECTURAS Y EJERCICIOS SUGERIDOS

Las referencias originales a los sistemas Intencionales aquí descritos incluyen a Dennett [47] definiendo su postura intencional, Searle [165] relacionando los estados intencionales con los actos de habla y Bratman, Pollak e Israel [23] en una aproximación más computacional al razonamiento práctico acotado. Estas nociones de intencionalidad, no son las únicas existentes en la literatura, Lyons [120] ofrece una revisión del concepto de intencionalidad, muy interesante para quien busca nuevas y válidas interpretaciones de sus constructores, incluyendo una aproximación neurofisiológica. De igual manera, Millgram [129] reporta otras concepciones del razonamiento práctico.

Numerosos lenguajes de programación orientados a agentes han sido propuestos después de `Agent0`. Bordini y col. [18] hacen una revisión de varios de ellos. Sterling y Taveter [179] discuten también estas ideas en el contexto del modelado orientado a agentes. Dastani [45] provee una revisión corta, y más reciente, del concepto de programación de SMA.

Ejercicios sugeridos

Ejercicio 2.1. *Ejemplifique las posibles descripciones del robot limpiador de los ejercicios del capítulo anterior, desde las tres posturas propuestas por Dennett. ¿Qué descripción es la más concisa? ¿Qué descripción es la más adecuada?*

Ejercicio 2.2. *Ejemplifique un caso en el escenario propuesto anteriormente, donde dos robots necesitan comunicarse ¿Qué tipos de actos de habla incluiría?*

Ejercicio 2.3. *Ejemplifique un razonamiento práctico, llevado a cabo por alguno de estos robots.*