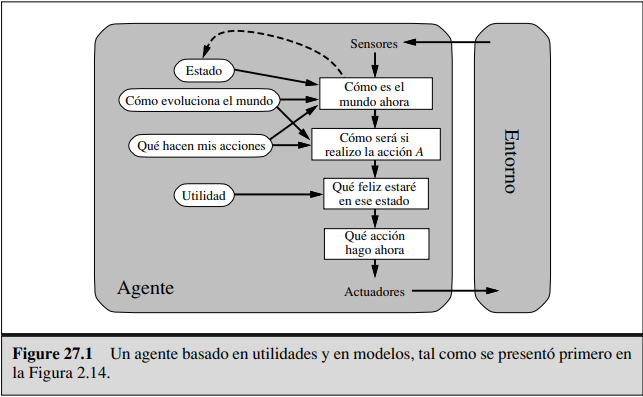
**IA: presente y futuro**

En en el que nos interesamos por dónde nos encontramos y hacia dónde vamos, lo cual es bueno antes de seguir hacia adelante.

En la Parte I, propusimos una visión unificada de la IA como diseño racional de agentes. Demostramos que el problema del diseño depende de las percepciones y acciones de los que disponga ese agente, las metas que debería satisfacer el comportamiento del agente, y la naturaleza del entorno. Pueden existir diferentes diseños de agentes, desde agentes reactivos hasta agentes basados en conocimiento y completamente deliberativos. Además, los componentes de estos diseños pueden tener diferentes instanciaciones, por ejemplo, lógicas, probabilísticas o «neuronales». En los capítulos intermedios se presentaron los principios por los que funcionan estos componentes. Ha habido un tremendo avance tanto en el entendimiento científico como en nuestras habilidades tecnológicas en lo referente a los diseños y componentes de agentes. En este capítulo, no entraremos en detalles y nos preguntaremos, ¿Nos va a llevar este progreso a un agente inteligente de propósito general que pueda actuar correctamente en diferentes entornos? El Apartado 27.1 estudia los componentes de un agente inteligente para evaluar lo que se conoce y lo que no. El Apartado 27.2 hace lo mismo, pero en la arquitectura completa de los agentes. El Apartado 27.3 pregunta si «el diseño de agentes racionales» es el objetivo adecuado en primer lugar. (La respuesta es, «Realmente no, pero por ahora está bien».) Finalmente, el Apartado 27.4 examina las consecuencias del éxito en nuestros esfuerzos.

**27.1 Componentes de los agentes**

En el Capítulo 2 presentamos varios diseños de agentes y sus componentes. Para centrar nuestro estudio, examinaremos el agente basado en la utilidad, como se muestra en la Figura 27.1. Este es el diseño de agentes más general de todos. Estudiaremos también su extensión con capacidades de aprendizaje, como se representa en la Figura 2.15.



Interacción con el entorno a través de sensores y actuadores: durante mucho tiempo en la historia de la IA, esto ha sido un notorio punto débil. Con unas pocas excepciones honorables, los sistemas IA se construyeron de tal forma que los humanos tenían que proporcionar las entradas e interpretar las salidas, mientras que los sistemas robóticos se centraban en tareas de bajo nivel en donde el razonamiento y la planificación de alto nivel estaban totalmente ausentes. Esto fue debido en parte al gran esfuerzo que requería la ingeniería y el alto coste para hacer funcionar robots listos. En los últimos años la situación ha cambiado rápidamente con la disponibilidad de robots programables listos para utilizarse o prefabricados, tales como los robots de cuatro patas que se muestran en la Figura 25.4(b). A su vez, estos se han beneficiado de las pequeñas cámaras CCD baratas y de alta resolución, y de las unidades motoras compactas y fiables. La tecnología MEMS (sistemas micro-electromecánicos) ha facilitado acelerómetros miniaturizados y está elaborando actuadores (efectores) que, por ejemplo, podrán activar a un insecto volador artificial. (También es posible combinar millones de actuadores MEMS para producir actuadores macroscópicos muy potentes.) Para los entornos físicos, entonces, los sistemas IA ya no tienen realmente excusa. Además, ya se dispone de un entorno enteramente nuevo como es Internet. Seguir la pista del estado del mundo: ésta es una de las capacidades centrales que se requieren para un agente inteligente. Requiere tanto percepción como actualización de las representaciones internas. En el Capítulo 7 describimos los métodos para seguir la pista de los mundos descritos por la lógica proposicional; el Capítulo 10 lo amplió hasta llegar a la lógica de primer orden; y el Capítulo 15 describió los algoritmos de filtrado para rastrear entornos inciertos. Las herramientas de filtrado se necesitan cuando está involucrada la percepción real (y por tanto imperfecta). Los algoritmos actuales de filtrado y de percepción pueden combinarse para hacer un trabajo razonable de informar predicados de bajo nivel, tales como «la taza está sobre la mesa», sin embargo, tenemos mucho que recorrer antes de que puedan informar de que «el Doctor Russell está tomando una taza de té con el Doctor Norvig». Otro problema es que aunque los algoritmos de filtrado aproximados puedan manejar una cantidad bastante grande de entornos, todavía son esencialmente proposicionales; como en la lógica proposicional, no pueden representar objetos ni relaciones de forma explícita. El Capítulo 14, explicó cómo se pueden combinar la probabilidad y la lógica de primer orden para resolver este problema; esperemos que la aplicación de estas ideas para rastrear entornos complejos produzca grandes beneficios. Casualmente, tan pronto como empecemos a hablar de objetos y de un entorno incierto, nos encontraremos con la incertidumbre de identidad, no conocemos qué objeto es cuál. Este problema ha sido el gran ignorado por la IA basada en la lógica, en la cual se ha supuesto generalmente que las percepciones incorporan símbolos constantes que identifican a los objetos. Proyección, evaluación y selección de cursos futuros de acción: los requisitos básicos de representación del conocimiento son los mismos aquí que para seguir la pista del mundo; la dificultad básica es hacer frente a los cursos de acción, tales como tener una conversación o tomar una taza de té, que finalmente constan de miles y millones de pasos primitivos para un agente real. Es sólo al imponer una estructura jerárquica que los humanos podemos abordarlos. Algunos de los algoritmos de planificación del Capítulo 12 utilizan representaciones jerárquicas y representaciones de primer orden para manejar problemas de esta escala; por otro lado, los algoritmos ofrecidos en el Capítulo 17 para la toma de decisiones bajo incertidumbre están utilizando esencialmente las mismas ideas que los algoritmos de búsqueda basados en el estado del Capítulo 3. Obviamente, aquí todavía queda mucho por hacer, quizás en la línea de los desarrollos actuales en el aprendizaje de refuerzo jerárquico. La utilidad como expresión de preferencias: en principio, basar las decisiones en la maximización de la utilidad esperada es completamente general y evita muchos de los problemas de los enfoques basados puramente en objetivos, tales como objetivos conflictivos y consecución incierta. Sin embargo, hasta ahora, no se ha trabajado mucho en la construcción de funciones de utilidades realistas; imaginemos, por ejemplo, una red compleja de preferencias interactuantes que debe ser entendida por un agente que funciona como un asistente de oficina para un ser humano. Resulta muy difícil descomponer las preferencias en estados complejos de la misma forma que las redes de Bayes descomponen las creencias sobre los estados complejos. Una razón puede ser que las preferencias sobre los estados se compilan desde las preferencias sobre las historias de estados descritas por las funciones de recompensa (véase el Capítulo 17). Aunque la función de recompensa sea simple, la función de utilidad correspondiente puede ser muy compleja. Esto sugiere que hay que tomar muy en serio la tarea de ingeniería del conocimiento para las funciones de recompensa como una forma de transmitir a nuestros agentes qué es lo que queremos que hagan. Aprendizaje: los Capítulos 18 y 20 describen la forma en que se puede formular el aprendizaje en un agente como aprendizaje inductivo (supervisado, sin supervisar o basado en el refuerzo) de las funciones que constituyen los diferentes componentes del agente. Se han desarrollado técnicas estadísticas y lógicas muy poderosas que pueden hacer frente a problemas bastante grandes, alcanzando o excediendo las capacidades humanas en la identificación de patrones predictivos definidos en un vocabulario dado. Por otro lado, el aprendizaje de la máquina no ha avanzado mucho en cuanto al importante problema de construir representaciones nuevas a niveles de abstracción mayores que el vocabulario de entrada. Por ejemplo, ¿cómo puede generar un robot autónomo predicados útiles tales como Oficina y Café si no son proporcionados por los humanos? Consideraciones similares se aplican al comportamiento del aprendizaje, Tomar Una Taza De Té es una acción de alto nivel importante, pero, ¿cómo introducirla en una biblioteca de acciones que inicialmente contiene acciones mucho más simples tales como Levante el Brazo y Trague? A menos que entendamos estos temas, nos enfrentamos a la tarea desoladora de construir a mano grandes bases de conocimiento de sentido común.

**27.2 Arquitecturas de agentes**

Es natural que preguntemos, «¿Cuál de las arquitecturas de agentes del Capítulo 2 deberían utilizar los agentes?» La respuesta es «¡Todas ellas!». Hemos visto que para situaciones en las que el tiempo no es esencial, se necesitan respuestas reflejas, mientras que la deliberación basada en el conocimiento permite que el agente planifique con antelación. Un agente completo debe ser capaz de hacer las dos cosas, utilizando una arquitectura híbrida. Una propiedad importante de las arquitecturas híbridas es que los límites entre los diferentes componentes de decisión no son fijos. Por ejemplo, la compilación, a nivel deliberativo, convierte continuamente la información declarativa en representaciones más eficientes, alcanzando finalmente el nivel reflejo, véase la Figura 27.2 (Este es el propósito del aprendizaje basado en explicaciones que se examinó en el Capítulo 19). Las arquitecturas de agentes tales como SOAR (Laird et al., 1987) y THEO (Mitchell, 1990) tienen exactamente esta estructura. Cada vez que resuelven un problema mediante deliberación explícita se almacena una versión generalizada de la solución para usarla en un componente reflejo. Un problema menos abordado es la inversión de este proceso: cuando el entorno cambia, los reflejos aprendidos ya no son apropiados y el agente debe retornar al nivel deliberativo para producir comportamientos nuevos.

Diagrama

Descripción generada automáticamente

Los agentes también necesitan formas de controlar sus propias deliberaciones. Deben poder parar de deliberar cuando se exige la acción, y deben poder utilizar el tiempo del que disponen para la deliberación y ejecutar los cómputos más rentables. Por ejemplo, un agente que conduce un taxi y ve un accidente delante, debe decidir en menos de un segundo si frenar o actuar de forma evasiva. También debería pensar en menos de un segundo las cuestiones más importantes, tales como si los carriles de la izquierda y derecha están libres y si tiene un camión detrás muy cerca, en vez de preocuparse de la tensión y el desgaste en las ruedas (neumáticos, gomas) o de dónde recoger al siguiente pasajero. Estos temas normalmente se estudian bajo el encabezamiento de IA en tiempo real. A medida que los sistemas IA entran en dominios más complejos, todos los problemas serán de tiempo real, porque el agente nunca tendrá tiempo suficiente como para resolver el problema de la decisión de forma exacta. Obviamente, existe una necesidad urgente de métodos que funcionen en situaciones de toma de decisiones más generales. En los últimos años han surgido dos técnicas prometedoras. La primera conlleva la utilización de algoritmos de cualquier momento (anytime) (Dean y Boddy, 1988; Horvitz, 1987). Un algoritmo de esta clase es un algoritmo cuya calidad de salida mejora gradualmente con el tiempo, de manera que tiene preparada una decisión razonable siempre que tenga una interrupción. Dichos algoritmos se controlan mediante un procedimiento de decisiones de metanivel que evalúa si merece la pena realizar más cómputos. Una búsqueda iterativa cada vez más en profundidad en los juegos es un ejemplo simple de un algoritmo de cualquier momento. También se pueden construir sistemas más complejos, que se compongan de muchos algoritmos de esta clase trabajando juntos (Zilberstein y Russell, 1996). La segunda técnica es el metarazonamiento teórico para las decisiones (Horvitz, 1989; Russell y Wefald, 1991; Horvitz y Breese, 1996). Este método aplica la teoría del valor de la información (véase el Capítulo 16) para la selección de cómputos. El valor de un cómputo depende tanto de sus costes (en lo que se refiere a demorar la acción) como de sus beneficios (en cuanto a la calidad de decisión mejorada). Las técnicas del meta-razonamiento se pueden utilizar para diseñar mejores algoritmos de búsqueda y para garantizar que los algoritmos tienen la propiedad «cualquier momento» (anytime). Desde luego que el meta-razonamiento es caro y que los métodos de compilación se pueden aplicar para que los sobrecostes sean pequeños en comparación con los costes de los cómputos que se están controlando. El meta-razonamiento no es sino un aspecto de una arquitectura reflexiva general, es decir, una arquitectura que permite la deliberación sobre las entidades y las acciones computacionales que ocurren dentro de la misma arquitectura. Un fundamento teórico de las arquitecturas reflexivas se puede construir mediante la definición de un espacio de estados conjunto compuesto por el estado del entorno y del estado computacional del mismo agente. Se pueden diseñar algoritmos de aprendizaje y de toma de decisiones que funcionen en este espacio de estados conjunto y que por tanto sirvan para implementar y mejorar las actividades computacionales del agente. Finalmente, se espera que los algoritmos específicos para unas tareas tales como la búsqueda alfa-beta y el encadenamiento hacia atrás (backward) desaparezcan de los sistemas IA, para ser sustituidos por métodos generales que dirijan los cómputos del agente hacia la generación eficiente de decisiones de alta calidad.

**27.3 ¿Estamos llevando la dirección adecuada?**

En la sección anterior mencionamos muchos avances y oportunidades para progresar. Pero, ¿a qué nos lleva todo eso? Dreyfus (1992) ofrece la analogía de intentar llegar a la luna trepando un árbol, de manera que se puede informar de un progreso continuo hasta llegar a conseguir la copa del árbol. En esta sección, estudiaremos si el camino actual de la IA es más parecido a trepar un árbol o a un viaje en cohete. En el Capítulo 1 dijimos que nuestro objetivo era construir agentes que actuaran racionalmente. Sin embargo, también dijimos que …alcanzar la racionalidad perfecta, haciendo siempre lo correcto, no es viable en entornos complicados. Las exigencias computacionales son simplemente muy elevadas. Sin embargo, en la mayor parte del libro, adoptaremos la hipótesis de trabajo en que la racionalidad perfecta es un buen punto de comienzo (arranque) para el análisis. Ahora, es momento de considerar una vez más cuál es exactamente el objetivo de la IA. Queremos construir agentes, pero, ¿qué especificación debemos tener en mente? Existen cuatro posibilidades: Racionalidad perfecta. Un agente perfectamente racional actúa en cualquier instante de tal manera que maximiza la utilidad esperada, dada la información que haya adquirido del entorno. Hemos observado que los cálculos necesarios para lograr la racionalidad perfecta en la mayoría de los entornos llevan demasiado tiempo, de forma que la racionalidad perfecta no es un objetivo realista. Racionalidad calculadora. Esta es la noción de la racionalidad que hemos utilizado implícitamente al diseñar agentes lógicos y teóricos para las decisiones. Un agente calculadoramente racional finalmente devuelve lo que habría sido la opción racional al comienzo de su deliberación. Esta es una propiedad interesante para que un sistema la presente; sin embargo en la mayoría de los entornos, la respuesta adecuada no tiene valor en el momento equivocado. En la práctica, los diseñadores de sistemas IA se ven obligados a llegar a un compromiso sobre la calidad de las decisiones para obtener un rendimiento global razonable; desgraciadamente, la base teórica de la racionalidad calculadora no proporciona un modo bien fundado para llegar a dichos compromisos. Racionalidad limitada. Herbert Simon (1957) rechazó la noción de racionalidad perfecta (o incluso aproximadamente perfecta) y la reemplazó con la racionalidad limitada, una teoría descriptiva de la toma de decisiones por agentes reales. Y en sus propias palabras, él escribió: La capacidad de la mente humana para formular y resolver problemas complejos es muy pequeña en comparación con el tamaño de los problemas cuya solución se requiere para un comportamiento objetivamente racional en el mundo real, o incluso para una aproximación razonable a dicha racionalidad objetiva. Simon sugirió que la racionalidad limitada funciona en primer lugar satisfaciendo, es decir, deliberando sólo el tiempo necesario para elaborar una respuesta que sea lo «suficientemente buena». Simon ganó el premio Nobel en economía por este trabajo y ha escrito mucho en profundidad acerca de este tema (Simon, 1982). En muchos casos, parece que es un modelo útil para los comportamientos humanos. Sin embargo, no es una 1104 INTELIGENCIA ARTIFICIAL. UN ENFOQUE MODERNO RACIONALIDAD PERFECTA RACIONALIDAD CALCULADORA RACIONALIDAD LIMITADA especificación formal para los agentes inteligentes, porque la definición de «suficientemente bueno» no la ofrece esta teoría. Además, satisfacer parece ser sólo uno de la gran cantidad de métodos que se utilizan para hacer frente a recursos limitados. Optimalidad limitada (Bounded optimality, BO). Un agente óptimo limitado se comporta todo lo bien que puede, dado sus recursos computacionales. Es decir, la utilidad esperada del programa agente para un agente óptimo limitado es por lo menos tan elevada como la utilidad esperada de cualquier otro programa agente que se ejecute en la misma máquina. De estas cuatro posibilidades, la optimalidad limitada parece ofrecer la mejor esperanza de conseguir un fundamento teórico fuerte para la IA. Tiene la ventaja de que es posible conseguirla; existe siempre por lo menos un mejor programa, algo de lo que carece la racionalidad perfecta. Los agentes óptimos limitados son realmente útiles en el mundo real, mientras que los agentes calculadoramente racionales normalmente no lo son, y los agentes satisfactores podrían serlo o no dependiendo de sus propios caprichos. El enfoque tradicional de la IA ha sido empezar con la racionalidad calculadora y a continuación hacer concesiones para satisfacer las restricciones de los recursos. Si los problemas impuestos por las restricciones son de menor importancia, se podría esperar que el diseño final sea similar al diseño de agentes BO. Pero a medida que las restricciones de los recursos se hacen más críticas, por ejemplo, a medida que el entorno es más complejo se podría esperar que los diseños vayan a divergir. En la teoría de la optimalidad limitada, estas restricciones se pueden abordar de una manera escrupulosa. Hasta ahora, no se conoce mucho sobre la optimalidad limitada. Es posible construir programas de optimalidad limitada para máquinas muy simples y para clases de entornos algo restringidos (Etzioni, 1989; Russell et al., 1993), pero todavía no tenemos idea de cómo serán los programas BO para grandes computadores de uso general en entornos complejos. Si va a existir una teoría constructiva de la optimalidad limitada, tenemos que esperar que el diseño de los programas óptimos limitados no dependa demasiado de los detalles del computador que se está utilizando. Ello haría que la investigación científica fuera muy difícil si añadir unos pocos kilobytes de memoria a una máquina de gigabytes cambiara significativamente el diseño del programa BO. Una forma de asegurarnos de que esto no ocurra es estar ligeramente más tranquilos con los criterios de optimalidad limitada. Por analogía con la noción de complejidad asintótica (ABO)(Apéndice A), la optimalidad limitada asintótica (ABO, Asymptotic Bounded Optimality) se puede definir de la siguiente forma (Russell y Subramanian, 1995). Supongamos que el programa P es un programa óptimo limitado para la máquina M en la clase de entornos E, en donde la complejidad de los entornos en E no está limitada. Entonces el programa P’ es ABO para M en E si puede superar a P ejecutándose en una máquina kM que es k veces más rápida (o mayor) que M. A menos que k fuera enorme, estaríamos encantados de que un programa fuera ABO para un entorno no trivial en una arquitectura no trivial. No tendría sentido esforzarnos por encontrar programas BO en vez de ABO, porque de todas formas el tamaño y la velocidad de las máquinas disponibles tiende a multiplicarse por un factor constante en una cantidad fija de tiempo. Podríamos arriesgar el pronóstico de que los programas BO o ABO para computadores potentes en entornos complejos no tendrán necesariamente una estructura simple y elegante. Y hemos visto ya que la inteligencia de propósito general requiere capacidad de reflejos y alguna capacidad deliberativa, una variedad de formas de conocimiento y toma de decisiones, mecanismos de aprendizaje y de compilación para todas esas formas, métodos de controlar el razonamiento, y un gran almacén de conocimientos específicos del dominio. Un agente óptimo limitado se debe adaptar al entorno en el que se encuentra de manera que su organización interna pueda reflejar las optimizaciones que sean específicas para ese entorno en particular. Esto es esperable, y es similar a la manera en que los coches de carreras restringidos por la capacidad del motor han evolucionado hasta tener diseños extremadamente complejos. Sospechamos que una ciencia de inteligencia artificial basada en la optimalidad limitada conllevará realizar un estudio amplio de los procesos que permiten a un programa de agentes converger hacia la optimalidad limitada y quizá con una concentración inferior en los detalles de los programas complicados que se obtengan como resultado. Resumiendo, se propone un concepto de optimalidad limitada como una tarea formal para la investigación de la IA que esté bien definida y que sea viable. La optimalidad limitada especifica programas óptimos en vez de acciones óptimas. Las acciones, después de todo, son generadas por programas, y es sobre los programas en donde tienen control los diseñadores.

**27.4 ¿Qué ocurriría si la IA tuviera éxito?**

En la novela de David Lodge, Small World (1984) sobre el mundo académico de la crítica literaria, el protagonista causa consternación cuando pregunta a un panel de teóricos literarios eminentes, pero contradictorios la siguiente cuestión: «¿Qué ocurriría si ustedes tuvieran razón?» Ninguno de los teóricos parece haber considerado antes esta cuestión, quizá porque debatir teorías infalsificables es un fin en sí mismo. Una confusión similar algunas veces se puede evocar preguntando a los investigadores en IA, «¿Qué ocurriría si tuvieran éxito?». La IA es fascinante, los computadores inteligentes son obviamente más útiles que los computadores no inteligentes, por tanto ¿por qué hay que preocuparse? Como se dijo en el Apartado 26.3, existen temas éticos a tener en cuenta. Los computadores inteligentes son más potentes, pero ¿se va a utilizar ese poder para fines buenos o malos? Aquellos que se esfuerzan por desarrollar la IA tienen la responsabilidad de ver que el impacto de su trabajo es positivo. El alcance de este impacto dependerá del grado de éxito de la IA. Y ya se han obtenido éxitos modestos que han cambiado las formas de enseñar la informática (Stein, 2002) y las formas en que se practica el desarrollo del software. La IA ha hecho posibles aplicaciones nuevas tales como los sistemas de reconocimiento de voz, sistemas de control de inventarios, sistemas de vigilancia, robots y motores de búsqueda. Se puede esperar que los éxitos de medio nivel en IA afectarán a toda clase de gente en sus vidas cotidianas. Hasta ahora, las redes de comunicaciones computerizadas, tales como los teléfonos móviles (celulares) e Internet, han tenido este tipo de efecto penetrante o dominante en la sociedad, pero la IA no. Podemos imaginar que los asistentes personales verdaderamente útiles para la oficina o para casa tendrían un impacto 1106 INTELIGENCIA ARTIFICIAL. UN ENFOQUE MODERNO positivo sobre las vidas de las personas, aunque a corto plazo podrían causar dislocaciones económicas. Una capacidad tecnológica a este nivel también se puede aplicar al desarrollo de armas autónomas, que muchos ven como un avance indeseable. Finalmente, parece probable que un éxito en IA a gran escala, la creación de inteligencia en el nivel humano y más allá, cambiaría las vidas a una mayoría de la humanidad. La verdadera naturaleza de nuestro trabajo y de nuestro papel cambiaría así como nuestro punto de vista de la inteligencia, la consciencia y el destino futuro de la raza humana. A este nivel, los sistemas IA podrían suponer una amenaza directa a la autonomía humana, la libertad, e incluso la supervivencia. Por estas razones, la investigación en IA no se puede divorciar de sus consecuencias éticas. ¿Qué camino deparará el futuro? Los autores de ciencia ficción parecen estar a favor de futuros anti-utópicos en vez de los utópicos, probablemente porque se producen argumentos más interesantes. Sin embargo hasta la fecha, la IA parece encajar con otras tecnologías revolucionarias (imprenta, fontanería, viajes en avión, telefonía) cuyas repercusiones negativas son compensadas por sus aspectos positivos. En conclusión, se puede observar que la IA ha progresado a lo largo de su corta existencia, pero todavía hoy tiene validez la frase del ensayo de Alan Turing sobre Computing Machinery and Intelligence: «Podemos observar sólo a poca distancia hacia delante, pero también se puede ver que todavía queda mucho por hacer.»