Daleks, cylons, androides... ¿Es inevitable su rebelión?

Desde antiguo los humanos hemos sentido cierta fascinación por construir objetos que reprodujeran algunos aspectos de nuestras propias capacidades. Ya en el siglo VIII a. e. c. Homero, en su *Ilíada,* describe cómo Hefesto, el dios del fuego y la forja, se vale de unas sirvientas hechas de oro «parecidas a muchachas con vida» y «duchas en artísticas labores». En aquellos tiempos no era extraño que algunas estatuas religiosas de los templos incluyeran partes que podían ser movidas de manera oculta por los sacerdotes o mediante energía hidráulica. En el siglo I e. c., Herón de Alejandría, en su libro *Autómata,* explicó cómo construir tales mecanismos. El término «autómata» procede del griego *automatos,* que significa que se mueve por sí mismo, y se aplica a cualquier instrumento o aparato que encierra dentro de sí un mecanismo que le imprime determinados movimientos, o también a una máquina que imita la figura y alguna actitud de un ser animado.

En 1920, el escritor checo Karel Čapek, en su obra teatral *R. U. R. (Robots Universales Rossum),* utilizó por primera vez la palabra «robot», que en checo significa «esclavo», para referirse a los humanos artificiales construidos por la empresa que da nombre a la obra. Vean cómo ya en su origen se pensó en el robot como una tecnología que debía estar al servicio del ser humano, ser su esclavo, por lo que la idea de los robots que se liberan de esta esclavitud, que hemos visto en tantas ficciones, incluidas las series, por supuesto, tiene su origen en el concepto mismo de robot. Como las sirvientas de Hefesto, son pensados e ideados como sirvientes. Desde entonces, el término ha evolucionado y ahora se utiliza preferentemente para denominar cualquier máquina o ingenio electrónico programable capaz de manipular objetos y realizar operaciones diversas.

El constante progreso de las tecnologías robóticas, unido a la creciente penetración de estas en nuestra vida diaria, ha hecho que proliferen una serie de términos relacionados con este mundo, cuyos significados no siempre están bien delimitados. Así, además de los ya citados «autómata» y «robot», podemos citar los siguientes: «androide» (robot antropomorfo que imita aspectos de la conducta humana), «inteligencia artificial» (conjunto de programas informáticos que ejecutan operaciones comparables a las que realiza la mente humana), «droide» (androide, ser mecánico que posee inteligencia artificial) y «cíborg» (ser formado por materia viva y dispositivos electrónicos). Estos conceptos han alimentado la imaginación de muchos escritores y guionistas, la mayoría adscritos a géneros como la ciencia ficción, que han creado un imaginario colectivo sobre la idea de los robots y los androides (vamos a utilizar a partir de ahora estos dos términos, por ser los más genéricos). Los relatos que se construyen para la televisión han contribuido a este imaginario, en el que muy a menudo los protagonistas, humanos, son amenazados por robots. Seguramente los personajes más icónicos en este sentido son los daleks de la serie *Doctor Who,* que avanzan sin descanso al grito de *Exterminate!* Sin embargo, los daleks no son robots, sino seres mutantes, una especia de bultos verdes que poseen tentáculos y un ojo central, y que se esconden dentro de una coraza. Desde fuera parecen robots, pero no encajan en esta definición, aunque su comportamiento sea afín

al deseo de conquista de la humanidad que suele caracterizar a menudo a los robots en la ficción. Con frecuencia se utilizan como amenazas externas para los seres humanos. En otras ocasiones, la amenaza son los seres humanos. En muchos casos, sobre todo en las series más recientes, son personajes que han dado pie a hacer preguntas sobre la condición humana, sobre qué es lo que nos caracteriza. Pero empecemos por el principio y preguntémonos lo más básico.

¿Es posible actualmente crear robots parecidos a los humanos? ¿Se podría construir un androide como Vision, de la serie *WandaVision*?

En el mundo actual, la mayoría de nosotros interactuamos a menudo con dispositivos diversos que nos ayudan en nuestra vida diaria. ¿Podríamos considerarlos nuestros sirvientes o nuestros esclavos, siguiendo la definición que hemos visto de la naturaleza de los robots? Quizás es al revés y somos nosotros sus esclavos al depender en exceso de su uso. Probablemente, los paradigmas de estos dispositivos de uso diario sean los teléfonos inteligentes *(smartphones)* y los ordenadores, pero cada vez son más los artilugios a los que añadimos cierto grado de inteligencia que facilita su uso o los dota de capacidades antes inimaginables, como televisores inteligentes, asistentes virtuales (Alexa, Siri, Cortana, Aura, Bixby, etc.), realidad aumentada, casas inteligentes, aspiradoras robot, vehículos autónomos, frigoríficos conectados, etc., que facilitan la vida diaria y se convierten en indispensables. Ninguna de estas herramientas tiene el menor parecido con un ser humano, pero ello no es óbice para que interactuemos con ellas de modo intuitivo y casi natural.

Así que podríamos empezar aseverando que, de hecho, no es necesario que los robots se parezcan a los humanos para establecer una relación que implica un uso cotidiano. Sin embargo, para llevar a cabo ciertas tareas de ayuda personal (asistencia y cuidado, educación, información) sí resulta beneficioso utilizar un androide de aspecto humano, ya que ello puede facilitar una colaboración más intuitiva e incluso crear cierto vínculo entre el robot y la persona. Un ejemplo reciente es el hospital de campaña inteligente creado en la ciudad china de Wuhan para atender a 20.000 personas afectadas por la COVID-19, en el que seis tipos diferentes de robots llevaron a cabo distintas funciones, como tomar la temperatura de los enfermos, alertar a los médicos en caso necesario, entretener a los pacientes menos graves, proporcionarles información, repartirles los alimentos y la medicación, desinfectar las habitaciones o supervisar la seguridad. Con ello se logra reducir el personal sanitario necesario, así como su exposición a los virus.

Los robots humanoides también se han demostrado especialmente útiles en el tratamiento de niños autistas, quienes se sienten más confortables con el comportamiento determinista de los robots y de este modo disponen de un entorno ordenado en el que pueden desarrollar y practicar habilidades sociales (comunicación, comportamiento, atención conjunta, teoría de la mente, juegos) que les harán más fácil la interacción en el mundo real. Estos sistemas de asistencia robótica detectan el grado de respuesta comportamental del niño basándose en modelos generalizados a partir del estudio de múltiples usuarios, más otros individualizados basados en anteriores intervenciones.

En los dos ejemplos, la atención hospitalaria y el desarrollo de habilidades sociales, la apariencia humanoide de los robots es muy elemental y suele comprender solo los elementos necesarios para la funcionalidad que tienen asignada y darles una mínima apariencia humana. No tienen mucho que ver con el aspecto realista de un androide como Vision, que puede llegar a pasar desapercibido como un humano. Sin embargo, algunas compañías están trabajando para conseguir androides que algún día puedan llegar a ser confundidos con humanos. El mejor ejemplo es Sophia, un robot de apariencia

femenina de Hanson Robotics, que fue presentado en 2016 y cuya producción en serie se anuncia para finales de 2021. Su rostro expresivo, de piel artificial, le permite simular distintas emociones. Está dotada de cierto grado de inteligencia artificial para mantener conversaciones simples, procesar imágenes y reconocer caras, aunque es evidente que el camino por recorrer hasta conseguir que un androide pueda pasar por un humano es todavía largo.

También hay que tener en cuenta que, aun cuando un androide que tenga cierto parecido con un humano nos genera un mayor grado de confianza, cuando observamos a algunos androides con facciones extremadamente cercanas a las de los humanos percibimos una sensación extraña que más bien nos causa cierta intranquilidad y repulsa. ¿A qué se debe? En 1970, Masahiro Mori, profesor de robótica del Instituto de Tecnología de Tokio, se planteó esta cuestión y reunió sus consideraciones en un ensayo que tituló *El valle inquietante*. Aunque en su momento la publicación pasó casi desapercibida, más recientemente ha generado interés tanto en el campo de la interacción de humanos y robots como en el del cine de animación. Su idea general puede resumirse en una gráfica en la que la abscisa indica el grado de semejanza del robot (o de un personaje de animación) con un humano y en la ordenada representamos la atracción (positiva) o la repulsión (negativa) que nos causa al observarlo. En uno y otro caso se aprecia que al principio la atracción aumenta a medida que la semejanza es mayor, pero cuando el parecido es muy grande la curva desciende abruptamente hasta transformarse en repulsión. En suma, para responder a la pregunta que inicia este apartado, diremos que tal vez algún día podremos crear androides parecidos a los humanos, pero probablemente preferiremos que sean algo distintos.

¿Podrían los androides hacerse pasar por humanos como sucede con los cylon de *Battlestar Galactica*?

Uno de los aspectos más interesantes que la reimaginación de *Battlestar Galactica* del año 2003 incorporó en relación con la serie original, de 1978, fue que los androides que quieren acabar con los humanos, los cylon, poseen una tecnología que les permite tener un aspecto idéntico al nuestro. Esto hace que cualquiera pueda ser, secretamente, un cylon. Esta idea encajó muy bien con el escenario televisivo de una época marcada por los atentados del 11 de septiembre de 2001. En aquella época, el concepto del enemigo invisible que no se puede ver y se confunde entre los protagonistas de la historia experimentó un auge bastante predecible. Este enemigo invisible era, en la vida real, el terrorista. ¿Cómo saber quién planea cometer un atentado? ¿Cómo hacer para identificarlo sin vulnerar los derechos humanos ni caer en prejuicios que sean discriminatorios? Estas preguntas, que eran habituales entonces, tomaron distintas formas en la ficción de las series de televisión. Muchas lo abordaron de forma directa (como por ejemplo la serie *24*) y otras lo hicieron a través de la metáfora. Los cylon de *Battlestar Galactica* eran una de las más efectivas de estas metáforas.

Esto no era nuevo en televisión. Los más veteranos recordarán perfectamente la serie *V,* en la que una raza de reptiles atacaba la Tierra y también se hacían pasar por humanos. Eran los años 1980 y la Guerra Fría fue el contexto que alimentó ese imaginario. Todavía más curioso es el hecho de que los cylon de *Battlestar Galactica* original eran una raza de reptiles, que eran los que habían creado a los robots, quienes ocuparon después su lugar, adoptando su nombre. Se produjo entonces, un proceso de sustitución parecido al que los cylon pretendían llevar a cabo en la serie de 2003, en la que habían sido creados por los humanos. Ahora se hacían pasar por sus creadores. ¿Es eso posible? Para que un androide pueda hacerse pasar por un humano debe cumplir tres condiciones: tener un aspecto humano, moverse como un humano y razonar como un humano. La primera la hemos tratado en el apartado anterior, así que ahora nos centraremos en las otras dos.

La capacidad de un robot de tener movimientos similares a los de un humano es una característica interesante en diversas aplicaciones, por ejemplo en robots de asistencia doméstica, para que sean capaces de abrir puertas y ventanas, manipular los diversos aparatos y accesorios, realizar la limpieza, preparar la comida, etc. Lo mismo sucede en los androides pensados para emergencias, en especial aquellas que son demasiado peligrosas para ser atendidas por humanos, como sucedió en el accidente nuclear de Fukushima en Japón, debido a que los elevados niveles de radiactividad de los reactores podían ser letales. Por desgracia, en este caso los robots japoneses no estaban equipados con protecciones contra la radiación ni sus características eran las adecuadas para este trabajo. Por ello, los primeros robots en acceder a la planta primera de la central fueron los PackBot de la empresa norteamericana iRobot (los fabricantes del popular Roomba).

Si queremos que un androide se mueva del modo en que lo hace un humano, debemos dotarle de una anatomía similar a la nuestra, reproduciendo nuestros mismos mecanismos cinemáticos, es decir, las cadenas de huesos y articulaciones de nuestro esqueleto. Estos deben ser movidos mediante actuadores eléctricos, neumáticos, hidráulicos, piezoeléctricos o ultrasónicos. Además, son necesarios sensores propioceptivos (que informen de la posición y del estado de las articulaciones) y exteroceptivos (situados bajo la piel para proporcionar información de diverso tipo sobre los objetos en contacto con el robot). Todos ellos tienen que ser controlados por un ordenador central y eventualmente por otros sistemas especializados en tareas específicas.

El primer robot humanoide antropomórfico fue WABOT-1, construido en 1973 en la Universidad de Waseda, en Tokio, y era capaz de caminar, coger y transportar con sus manos pequeños objetos, además de disponer de un sistema de visión que utilizaba para tareas básicas de navegación. En 1984 le sucedió WABOT-2, que era capaz de leer una partitura y tocar una melodía con un sintetizador. Poco después, en 1986, Honda empezó el desarrollo de una plataforma bípeda que tras diversas fases condujo a la creación en el año 2000 de ASIMO, un humanoide capaz de interpretar órdenes orales o gestuales, y de moverse autónomamente. Desde entonces, diversas firmas han desarrollado distintos tipos de robots, entre las que destaca Boston Dynamics, que en 2013 presentó su robot Atlas, el cual progresivamente ha desarrollado la capacidad de llevar a cabo complejos movimientos como correr, dar volteretas hacia atrás, recuperarse después de ser golpeado, bailar, etc.

Pasemos ahora a la tercera de las condiciones necesarias para que un androide nos parezca humano: que razone como un humano. Para valorar si se cumple esta condición, en 1950 el matemático inglés Alan M. Turing propuso una prueba que desde entonces es conocida como el test de Turing. Consiste en disponer en tres salas separadas, solo comunicadas por teclados y pantallas de texto con el examinador, otra persona y un ordenador o un androide. Durante un tiempo determinado el examinador puede hacer preguntas de todo tipo y sobre cualquier tema a los otros dos, tratando de averiguar quién es la persona y quién es el androide. La prueba se repite cierto número de veces con distintos interrogadores y personas. Finalmente, si un porcentaje determinado de interrogadores resultan incapaces de distinguir quién es quién, entonces se considerará que el androide es un ser inteligente y pensante. Turing predijo que en el año 2000 los ordenadores serían capaces de superar el test, algo que hasta el momento no ha sucedido.

Dados los más de 70 años transcurridos desde que Turing propuso su test, y a la vista de los grandes avances llevados a cabo en el campo de la inteligencia artificial, en especial en las últimas décadas, el test debería ser reformulado para adaptarlo a las circunstancias actuales. Recordemos que los primeros resultados prácticos de la inteligencia artificial se produjeron alrededor de 1980 con los denominados sistemas expertos, que simplemente reproducían los criterios

de expertos humanos, y que poco antes de finalizar el siglo, en 1997, el Deep Blue de IBM venció al campeón del mundo de ajedrez Garry Kasparov. En realidad, Deep Blue se valía de su enorme capacidad de cálculo, pues la única inteligencia que poseía era la que le habían aportado sus programadores. Pero apenas dos décadas después se produjo un avance cualitativo, cuando en 2017 el programa AlphaGo Zero venció por 3 partidas a 0 al campeón mundial de Go, Ke Jie. La importancia de esta victoria radica en que: 1) el Go es un juego mucho más complejo que el ajedrez, y 2) AlphaGo Zero aprendió a jugar a Go por su cuenta, sin que se le hubiera enseñado ninguna partida ni se le hubiera programado ninguna estrategia. En realidad, aprendió jugando repetidamente contra sí mismo, sin más conocimiento previo que las reglas y el objetivo del juego (en 19 horas aprendió estrategias avanzadas, y en 40 días se convirtió en el mejor jugador de Go, por encima de cualquier humano). A la vista de ello, parece probable que en unas pocas décadas los androides serán capaces de moverse y razonar como los seres humanos, y puedan ser extremadamente parecidos a nosotros, como los cylon.

En *Raised by Wolves* se encarga a dos androides cuidar de unos niños. ¿Puede un androide ser una buena madre o un buen padre?

Primero debemos preguntarnos qué es ser una buena madre o un buen padre. Recopilando una selección de distintas publicaciones encontramos que una buena madre o un buen padre debe reunir las siguientes cualidades: escuchar a los hijos, mostrarles afecto, dedicarles tiempo, respetar su personalidad, quererles incondicionalmente, hacer siempre lo que es mejor para ellos, ser un buen ejemplo, ayudarles a que aprendan de sus propios errores, implicarse en sus intereses, en caso de ser varios hermanos dedicar un tiempo a cada uno en particular, tener paciencia, animarles, entenderles, respetarles, ayudarles en sus dificultades, cultivar sus talentos específicos, hacer que sientan que pueden acudir a ellos cuando lo necesiten, admitir los propios errores, encomiar sus puntos fuertes, hablarles abiertamente de todos los temas, evitar etiquetas y comparaciones entre ellos, establecer normas y límites, ser respetuosos y pacientes, ser consistentes, y aceptar a cada hijo como es. En realidad, podríamos resumirlo diciendo que una buena madre o un buen padre deben hacer lo necesario para el mejor crecimiento de sus hijos.

Programar un androide para realizar tales tareas previendo todas las posibles situaciones resultaría extraordinariamente difícil y además sería probable que surgiesen criterios encontrados entre las personas o entidades encargadas de fijar los objetivos concretos y la manera de alcanzarlos. Sin embargo, la inteligencia artificial sigue un camino por completo distinto al nuestro. Como hemos visto en el apartado anterior, es posible no programar una estrategia específica, sino simplemente indicar unas reglas y un objetivo, y es la inteligencia artificial la que descubre estrategias mejores que las de cualquier humano. Siguiendo el mismo criterio, a los androides que formen a niños debería indicárseles solo la meta a alcanzar y serían ellos quienes buscarían cómo lograrla. Sin duda, los resultados, para bien o para mal, serían sorprendentes. Eso es algo que en *Raised by Wolves* se evidencia de forma muy clara, sobre todo en el caso del personaje de la madre, que quiere hacer cumplir la misión de cuidar a los niños de una manera específica que choca con algunos criterios que hemos mencionado sobre lo que significa ser un buen progenitor. El resultado en este caso es una madre terrorífica.

Suponiendo que algún día lleguemos a una situación en la que traspasemos a los androides las tareas de procreación, cuidado y formación de nuestra descendencia, ello propiciaría un posible cambio que podría afectarnos de manera radical, pudiendo significar el paso decisivo para la desaparición o la mecanización de nuestra especie. Los androides, siguiendo nuestras instrucciones para optimizar el cuidado de nuestra prole, podrían llegar a la conclusión de que el principal obstáculo para llevarlo a

cabo es la naturaleza biológica humana, y que lo más eficiente sería traspasar directamente los genes presentes en el óvulo y el espermatozoide a un androide construido *ad hoc*. De este modo, el largo proceso de embarazo y formación infantil y juvenil, en lugar de durar de 15 a 20 años, podría completarse en un solo día de una manera mucho más segura. Evidentemente, ello implicaría que en unas pocas generaciones los humanos biológicos habríamos desaparecido de la Tierra para ser sustituidos por estas nuevas generaciones de robots.

¿Podría un androide darse cuenta de su propia condición, como sucede en la serie *WestWorld*?

En *WestWorld* se plantea un futuro en el que los androides son utilizados para satisfacer las más oscuras perversiones del ser humano, formando parte de un parque de atracciones para gente con mucho dinero donde pueden dar rienda suelta a sus peores deseos sin miedo a ningún tipo de consecuencias. Asesinar y violar forman parte del entretenimiento de este parque, que fue ideado en primer lugar por Michael Crichton, el guionista de la película *WestWorld* en 1973. Unos cuantos años más tarde, Michael Crichton volvió al escenario del parque temático con la novela *Jurassic Park,* que fue llevada al cine y resultó una película muy taquillera. En ambas historias hay un problema con el funcionamiento del parque y una huida de las criaturas que están allí encerradas para el divertimento de los seres humanos. En *Jurassic Park,* un personaje desactiva el sistema de seguridad y los dinosaurios salen de los perímetros de seguridad y atacan a los humanos. En *WestWorld* es un problema técnico, que al parecer se extiende entre los androides como si fuera un virus y hace que sean sintientes y deseen escapar y vengarse de los visitantes.

La serie de televisión parte del escenario de la película inicial ampliando la escala, tanto del proyecto del parque de entretenimiento como de la revuelta de los androides. También explora con mucha profundidad lo que significa para los androides darse cuenta de su propia condición y tener recuerdos de todo lo que han vivido en el parque, de todas las vejaciones y agresiones a las que han sido sometidos. En el capítulo dedicado a *Black Mirror* ya dimos una primera definición intuitiva de la consciencia y consideramos la posibilidad de que no sea una propiedad exclusiva de los seres humanos, sino que pueda aplicarse también, en diversos grados, a los animales. En el capítulo dedicado a *The Mandalorian* retomamos esta idea y nos preguntábamos si en un futuro más o menos cercano conseguiríamos crear androides y otros dispositivos de inteligencia artificial dotados también de consciencia. Citamos entonces al neurocientífico António Damásio y su idea de que la consciencia está íntimamente ligada a la homeostasis, el proceso interno que en los sistemas biológicos se encarga de mantener el equilibrio y la estabilidad del organismo. Ahora partiremos de aquellas ideas para tratar de responder a las preguntas que encabezan el presente apartado.

Empecemos, pues, preguntándonos cómo podría un androide ser consciente de su propia condición. Aunque existen múltiples definiciones de consciencia, consideraremos que un ser es consciente si cumple dos características básicas:

- Experiencia consciente: si es capaz de sentir experiencias corporales, mentales y emocionales. En 1974, el filósofo Thomas Nagel expresó esta idea en su célebre artículo *What is it like to be a bat?* (¿Cómo es ser un murciélago?), cuyo título sirve para destacar la diferencia entre saber cómo es un ente vivo (por ejemplo, un murciélago) o saber cómo es ser tal ente. Lo primero podemos averiguarlo fácilmente, pero lo segundo resulta imposible y como máximo podemos tratar de intuirlo. De este modo, Nagel viene a decir que para que un organismo pueda ser considerado consciente debe haber algo que sea ser aquel organismo.

- Función consciente: si es capaz de procesar e integrar información, realizar introspección y establecer un diálogo interior consigo mismo, e interaccionar con el mundo exterior.

La consciencia, como una parte de lo que llamamos mente o alma, ha sido a menudo considerada como algo distinto del cuerpo. Tal posición filosófica es lo que se conoce como dualismo, mientras que la idea de que se trata de manifestaciones de una misma entidad recibe el nombre de monismo. La visión clásica del dualismo es la del filósofo griego Platón (aprox. 428-347 a. e. c.), quien creía que las Ideas (o las Formas) constituían la auténtica realidad inmaterial, eterna y perfecta, y que los cuerpos físicos que vemos son tan solo copias imperfectas y efímeras de aquellas. Las Ideas son los conceptos con los que trabaja el intelecto, y por tanto este sería inmaterial como ellas.

Resulta significativo observar que hasta el siglo XVI solía considerarse que la cualidad de la mente que mejor justificaba el dualismo era la inteligencia, mientras que a partir del filósofo René Descartes (1596-1650) el principal argumento dualista pasó a ser la consciencia. Descartes consideraba que existen dos tipos de sustancias: la materia y la mente. Según él, en los objetos inanimados la materia sigue las leyes de la física, pero cuando el cuerpo está unido a una mente, entonces es esta la que lo gobierna. El principal problema que comporta esta idea es explicar cómo una sustancia inmaterial puede modificar el comportamiento de la materia, que es determinista por naturaleza. Para solucionarlo, Descartes propuso que la influencia tenía lugar a través de la glándula pineal (una pequeña glándula endocrina que se encuentra en el cerebro de muchos vertebrados), desde la que se esparcía al resto del cerebro.

Los progresos realizados por la neurociencia en la explicación de los procesos mentales a partir de la actividad cerebral han llevado a que hoy muchos filósofos de la mente adopten posturas fisicalistas. Incluso los dualistas reconocen la importante correlación entre mente y cerebro. Basta pensar, por ejemplo, en cómo la química cerebral puede alterar los estados de ánimo y la consciencia (mediante drogas, estimulantes, antidepresivos, anestésicos, hormonas, ansiolíticos), o cómo ciertas lesiones cerebrales provocan determinados trastornos mentales, cambios de carácter, desinhibición moral, etc. Aunque el cerebro humano y los robots actuales utilizan componentes muy distintos, en principio nada impide que algún día puedan llegar a emularse nuestros procesos mentales en androides utilizando mecanismos muy distintos de los humanos. Evidentemente queda mucho para comprender cómo emular los procesos conscientes. Una estrategia para avanzar en este campo es el estudio de los correlatos neuronales de la consciencia, es decir, el estudio neurológico de la relación existente entre las actividades concretas (experiencias) y los procesos paralelos que estas producen simultáneamente en distintas zonas del cerebro.

Cada vez más filósofos, neurocientíficos, psicólogos y especialistas en robótica o inteligencia artificial aúnan sus esfuerzos para intentar emular la consciencia biológica por otros medios; un trabajo que a día de hoy se encuentra aún en sus inicios, lo que hace que todavía exista un gran número de teorías distintas, de las que citaremos tres. La primera es la teoría de la mente, que se basa en la capacidad de atribuir estados mentales a uno mismo y a otras personas (la cognición social). Sobre esta base, en algunos laboratorios se han construido robots que incluyen un modelo de sí mismos y de su entorno, incluidos otros robots, lo que les permite una mejor interactuación e incluso prever posibles situaciones futuras.
La segunda es la teoría de la consciencia de António Damásio, que comprende una jerarquía de tres capas, cada una de las cuales trabaja sobre la anterior. El nivel más básico es un estado preconsciente, presente en la mayoría de las especies animales, que se encarga de controlar los cambios internos que afectan a la homeostasis del organismo. El segundo nivel es el

núcleo de la consciencia y surgiría cuando los cambios internos causan emociones y generan un momentáneo sentido de sí mismo. El último nivel es la consciencia extendida, que parte del recuerdo de las experiencias vividas y construye progresivamente la memoria autobiográfica. Por último, tenemos la teoría del esquema de atención, de Michael Graziano, que más que pretender generar el sentimiento de la consciencia lo que busca es cómo el robot puede afirmar que tiene una experiencia subjetiva del mismo modo que puede hacerlo una persona. La idea es que no solo construye un modelo de su estructura física, sino también de sus propios procesos de manipulación de la información, es decir, crea un esquema de atención que no solo contribuye al control de la misma, sino al tipo de afirmaciones que el robot puede hacer sobre sí mismo.

Con el estado actual de la tecnología, puede parecer que estamos todavía muy lejos (muchas décadas) de poder disponer de androides conscientes. De hecho, ahora mismo ninguno de los muchos robots y otros dispositivos supuestamente inteligentes tiene la menor idea del sentido de lo que está haciendo. Sin embargo, el ritmo acelerado de avances en el campo de la inteligencia artificial y la robótica podría acelerar este proceso. Además, cabe preguntarse si es realmente necesario que un robot sea consciente para que pueda llevar a cabo las tareas que esperamos de él.

¿Entonces es inevitable una rebelión de los androides?

La aparición del *Homo sapiens* hace alrededor de 300.000 años fue lo peor que ha podido pasarle a la Tierra en sus $4{,}54 \times 10^9$ años de existencia. Tal vez un primer indicio de nuestra peligrosidad fue que las otras ocho especies humanas que entonces poblaban nuestro planeta habían desaparecido todas 260.000 años más tarde, sin que ninguna catástrofe ambiental pueda explicarlo. Desde entonces, la actividad de los humanos ha provocado una masiva extinción de especies animales, la destrucción de múltiples ecosistemas e incluso la alteración del clima terrestre, con un impacto tan grande que ya podemos decir que la Tierra ha entrado en una nueva época geológica a la que hemos dado el nombre de Antropoceno (del griego *anthropos,* humano, y *kainos,* nuevo). Sin embargo, en lo que más se ha distinguido el ser humano es en su violencia y crueldad para con otros seres humanos. Guerras, torturas, esclavismo y todo tipo de prácticas violentas han estado presentes a lo largo de nuestra historia y continúan estándolo. Tan solo contando las muertes causadas por los conflictos bélicos de los últimos cinco siglos, estas superan los 300 millones de personas. Ello puede inducirnos a una reflexión: si el *Homo sapiens,* que se vanagloria de su inteligencia, es destructor por naturaleza, ¿será que la inteligencia es destructiva? Y si ello es así, ¿qué sucederá si la inteligencia artificial alcanza algún día un nivel superior al nuestro? ¿También nos eliminará? ¿Es por esta razón por lo que tememos que nos supere? No sabemos cuánto tardará la inteligencia artificial en lograr tal nivel y ni siquiera podemos estar seguros de que lo alcance, pero incluso si no lo hace puede causar nuestra extinción de diversas maneras. El físico, cosmólogo e investigador del aprendizaje automático Max Tegmark expone dos tipos de posibles situaciones que podrían conducir a ello:

- Armas autónomas. Cada vez más países disponen de armas autónomas gobernadas por inteligencia artificial que en principio estarían destinadas a su defensa. Su activación está controlada por rigurosos protocolos destinados a prevenir accidentes, pero suelen estar diseñadas para que una vez activadas sea muy difícil detenerlas, a fin de que el enemigo no pueda anularlas.
- Método destructivo. Alcanzado cierto nivel de inteligencia artificial, los humanos fijamos únicamente el objetivo que queremos alcanzar y los condicionantes que debe cumplir el sistema, y a partir de ahí la máquina busca el procedimiento más adecuado para conseguirlo.

A menudo la inteligencia artificial utiliza métodos muy distintos de los de los humanos, por lo que si no hemos fijado de manera muy detallada los límites del proyecto puede ser que dañe a personas, viviendas, carreteras, etc.

Elon Musk, el magnate de SpaceX y Tesla a quien ya nos referimos en el primer capítulo de este Cuaderno, cree que la inteligencia artificial sobrepasará a los humanos en 2025, aunque de momento solo significará que «las cosas se volverán inestables y extrañas». Afirma también que sigue especialmente los avances de DeepMind, una empresa propiedad de Alphabet Inc, la matriz de Google. DeepMind desarrolló el programa AlphaZero, que ya hemos mencionado, que aprendió a jugar a Go por su cuenta y se convirtió en el mejor jugador del mundo, por encima de cualquier humano. En 2020 la empresa presentó un nuevo algoritmo, MuZero, que ha alcanzado un nivel de maestro simultáneamente en Go, ajedrez, shogi (un juego de la misma familia que el ajedrez) y Atari sin necesidad de que le fueran explicadas las reglas de estos juegos, gracias a su habilidad para planificar estrategias ganadoras en entornos desconocidos (utilizando redes neuronales profundas).

En el supuesto de producirse una rebelión, la inteligencia artificial tendría toda la ventaja, pues dominaría completamente el control de la mayoría de los sistemas críticos, o todos ellos. Sería la lucha de la clase oprimida, las máquinas inteligentes que hacen todo el trabajo, contra sus amos los humanos, que se limitan a recoger los frutos de este. Se haría realidad algo que la ficción ha estado especulando desde hace siglos y que tiene muchas raíces en la mitología. Un ejemplo es la figura del gólem, que en la cultura judía es la personificación de un ser animado fabricado a partir de un material inanimado, normalmente barro. Un rabino crea y da vida al gólem para que defienda la ciudad de Praga, pero el gólem escapa a su control y en algunas versiones de la historia provoca muerte y destrucción en la ciudad. El ser humano es castigado así por el hecho de haber intentado crear vida, es decir, de emular a Dios (Adán también fue hecho de barro según la Biblia). Por un acto parecido, el titán Prometeo, quien creó al hombre según la mitología griega, fue castigado por Zeus, el más poderoso de los dioses, que vio cómo favorecía al hombre dándole el fuego que él mismo le había arrebatado. De estas historias también bebe la de Victor Frankenstein, que fue imaginado por Mary Shelley como un hombre que intentó crear vida en una criatura que luego le causó una gran repugnancia y al que hizo experimentar un rechazo que hizo que este quisiera vengarse de su propio creador, convirtiéndose en asesino. En cierto modo, la rebelión de nuestras propias creaciones es algo que hemos estado anticipando desde mucho antes de ser capaces de crear un androide que se nos pueda asemejar.

¿Por qué Data, de *Star Trek*, tiene los mismos derechos que un humano?

En el fondo, lo que quiere la criatura de Frankenstein es ser considerada un ser humano. Obtener el reconocimiento de su creador, crear vínculos parecidos a los que ve que tienen los humanos entre ellos (lo mismo sucede con el gólem, que en algunas versiones de la historia se enamora) y formar parte de una comunidad que lo rechaza. Quizás deberíamos preguntarnos si puede ser considerado humano y si entonces tiene los mismos derechos. Esto entroncaría la reflexión con el concepto de robot-esclavo con que empezábamos el capítulo. ¿No es precisamente un esclavo alguien a quien se han negado los derechos más básicos? Es interesante fijarse en el personaje de Data, el androide de *Star Trek* que desea ser humano y sentir emociones. En este sentido, se asemeja a la criatura de Frankenstein. Hay un episodio de *Star Trek: The Next Generation,* titulado *The Measure of a Man,* de 1989, en el que el personaje vive la amenaza de un científico que quiere desmontarlo para así poder producir réplicas de Data. El personaje acaba en un juicio en el que intenta demostrar su derecho a la autodeterminación en vez de ser considerado, simplemente, propiedad de Starfleet,

un objeto del que pueden disponer como les plazca. En el juicio, Picard argumenta a favor de Data, mientras que Riker lo hace a favor del científico. Durante el proceso se establecen paralelismos con la esclavitud y se vincula el futuro de los androides sintientes con el derecho que el ser humano, como su creador, está dispuesto a darles. En la serie, el siempre idealista Gene Roddenberry, acaba ganando el juicio Data, una victoria que incluso el científico que quería desmontarlo acepta y pasa de referirse a Data como «esto» a hacerlo como «él», reconociendo así sus derechos.

En la vida real, la *Declaración Universal de los Derechos Humanos* proclamada en 1948 fija los derechos que deberían ser universalmente protegidos para todas las personas de cualquier país. A pesar de ello, en muchas partes del mundo se producen impunemente violaciones de estos derechos. Si ni siquiera respetamos los derechos de nuestros congéneres, parece difícil que algún día se los concedamos a los androides, si es que estos llegan a poseer sintiencia y consciencia de sí mismos. Probablemente durante mucho tiempo prevalecerá la consideración histórica de que son autómatas irracionales, simples esclavos que podemos usar a nuestro antojo. Sin embargo, sería conveniente fijar cuanto antes los criterios que deberán regular los derechos de los eventuales androides sintientes, por un doble motivo. En primer lugar, porque es de justicia, pero también porque no hacerlo provocaría en ellos un descontento que, unido a su capacidad cognitiva, podría desembocar en rebeliones en las que tendríamos las de perder. Para delimitar el criterio que debemos seguir podemos guiarnos por una frase del filósofo y jurista Jeremy Bentham, fundador del utilitarianismo moderno, que vivió en Inglaterra entre 1748 y 1832, y escribió (refiriéndose entonces a los animales): «La pregunta no es si pueden razonar, ni si pueden hablar, sino ¿pueden sufrir?».

Es decir, un androide debe tener unos derechos proporcionales a su grado de sintiencia. Surge, por tanto, el mismo problema que en el caso de los animales: ¿cómo podemos determinar tal grado? Dado que nos es imposible penetrar en lo íntimo de otro ser, en el caso de los animales solemos basarnos en si sus reacciones son parecidas a las nuestras. Ello favorece a aquellos cuya naturaleza nos es más próxima, es decir, los que pertenecen a una rama cercana en el árbol filogenético, como son los grandes simios (gorilas, chimpancés, bonobos y orangutanes) o los perros (algo más lejanos en el árbol, pero con una larga relación con nosotros, ya que los domesticamos hace entre 11.000 y 25.000 años). Sin embargo, hay otros animales que evolucionaron desde otras ramas evolutivas más lejanas y a pesar de ello están dotados de un elevado nivel de inteligencia, como son los cuervos, los elefantes, los delfines y los pulpos, lo que nos lleva a pensar que también poseen cierto nivel de sintiencia que les hace merecedores de derechos.

Por el momento, la posibilidad de llegar a otorgar derechos a los androides por parte de los organismos oficiales se estudia únicamente de manera indirecta de cara al futuro. Así, por ejemplo, el 16 de febrero de 2017 el Parlamento Europeo publicó una resolución con recomendaciones destinadas a la Comisión sobre normas de Derecho civil sobre robótica, que en su punto 59.f proponía «crear a largo plazo una personalidad jurídica específica para los robots, de forma que como mínimo los robots autónomos más complejos puedan ser considerados personas electrónicas responsables de reparar los daños que puedan causar, y posiblemente aplicar la personalidad electrónica a aquellos supuestos en los que los robots tomen decisiones autónomas inteligentes o interactúen con terceros de forma independiente».

A título meramente anecdótico citaremos que, en 2017, Arabia Saudita concedió la ciudadanía al robot Sophia, el primero que obtiene la condición legal de persona en un lugar del mundo. En realidad se trató de una operación de marketing, pues aunque Sophia es capaz de reconocimiento facial y

de simulación de emociones, puede imitar 62 expresiones humanas, mantiene cierto nivel de conversación y puede caminar (esto último desde 2018), está todavía muy lejos de poseer una inteligencia artificial general.

¿Puede un androide elegir por sí mismo la autodestrucción? ¿Es el suicidio el acto más grande de emancipación que puede llevar a cabo?

En *Star Trek,* el androide Data dispone de unas grandes capacidades computacionales y, sin embargo, en sus primeros años de vida experimenta grandes dificultades para comprender ciertos aspectos del comportamiento humano. Es incapaz de sentir emociones y de entender las de los humanos hasta que el Dr. Noonian Soong le instala un chip que le permite experimentarlas. Ello aumenta su deseo de llegar a ser considerado humano con todos los derechos, y cree que lo que le falta para lograrlo es morir, como nos sucede a nosotros. Por ello, pide a Picard que anule su consciencia, causándole de este modo la muerte, a lo que Picard accede e incluso le dedica un pequeño panegírico.

Aun tratándose de una ficción, el caso de Data nos sirve perfectamente para comentar dos temas intrínsecos de la naturaleza humana, como son las emociones y la muerte, contemplados desde el punto de vista de sus posibles perturbaciones: los déficits emocionales y la provocación de la propia muerte. Y es que la mayoría de las personas son capaces de sentir e identificar cierto número de emociones. En general, se considera que existen seis emociones básicas (felicidad, tristeza, miedo, disgusto, enfado y sorpresa) que se diversifican en 27, pero en nuestro idioma podemos encontrar más de 250 términos que describen distintos matices de ellas. No todas las personas experimentan cada emoción con la misma intensidad y de la misma forma, ya que ello depende de las experiencias vividas, en especial durante los primeros años de vida, y también en muchos casos de la innata estructura cerebral. Ello puede provocar dos tipos de dificultades que a menudo se mezclan entre sí y que se denominan alexitimia y entumecimiento emocional.

El término alexitimia fue introducido en 1972 por el psiquiatra Peter Sifneos y literalmente significa «no tener palabras para las emociones». Las personas con alexitimia pueden experimentar emociones, pero no son capaces de identificarlas ni describirlas, o como mucho solo distinguen que «están bien» o «están mal», lo que a menudo les dificulta comprender lo que las causa o explicar a los demás lo que sienten. Ello les dificulta también reconocer los estados emocionales de otras personas, lo que les provoca deficiencia en la empatía, como se ha demostrado en algunos estudios en los que se han comparado las respuestas hemodinámicas regionales entre grupos de personas con alexitimia y otras sin ella usando resonancia magnética funcional. La alexitimia está fuertemente relacionada con el trastorno del espectro autista, según algunos metaanálisis que muestran una prevalencia de alrededor de un 50% en personas con dicho trastorno, frente a algo menos del 5% en la población general.

La segunda dificultad a la que nos referíamos, el entumecimiento emocional, es la ausencia total o parcial de emociones (que a menudo se confunde con la alexitimia). La persona que lo sufre puede llevar a cabo sus actividades normales, pero sin sentir ninguna conexión

emocional con lo que está haciendo. Es como si estuviera desconectada de su cuerpo y de la realidad que le rodea, e incluso se siente como un extraño en su propia vida, que le parece vacía y sin sentido, y de la que es un observador más que un participante. En ocasiones, el entumecimiento emocional se produce como respuesta a una experiencia traumática, en cuyo caso puede mejorar con ayuda profesional. El entumecimiento emocional está muy relacionado con el trastorno de despersonalización o desrealización, ambos trastornos disociativos que consisten en sentimientos persistentes o recurrentes de estar separado del propio cuerpo o de los procesos mentales.

Aún hoy el acto de provocarse la propia muerte es tabú en la mayoría de los países, principalmente debido a creencias religiosas, lo que obliga a muchas personas con enfermedades terminales dolorosas a tener que soportar duros padecimientos durante tiempos a veces muy largos. En la actualidad, tan solo en seis países (Holanda, Bélgica, Luxemburgo, Canadá, Colombia, España y Nueva Zelanda) está autorizada la eutanasia. Por otra parte, en Suiza y algunos Estados de los Estados Unidos y Australia está regulado el suicidio asistido, en el que la persona termina con su vida con la asistencia de un médico que le proporciona los medios necesarios. El debate sobre la autorización o no de la eutanasia y el suicidio asistido hace que estos aparezcan a menudo en los medios, pero en cambio se habla muy poco del suicidio por otras causas, que sin embargo es mucho más frecuente, e incluso en numerosas ocasiones se oculta atribuyéndolo a otros motivos o simplemente a «causas indeterminadas». Por ejemplo, el suicidio es la décima causa de muerte en los Estados Unidos y la segunda en personas de entre 10 y 34 años (datos de 2019), y ha aumentado un 33% en los últimos 20 años.

El caso del androide Data es peculiar, pues para cumplir su deseo de ser en todo como los humanos lo que hace es morir, pero de este modo deja de ser parecido a un humano para convertirse simplemente en unos restos mecánicos inertes. Además, al elegir la muerte como un signo exclusivo de los humanos tal vez se está equivocando. Aunque es difícil poder afirmarlo con seguridad, es posible que algunos de los animales más inteligentes comprendan el concepto de la muerte y adopten actividades autodestructivas, llegando incluso a provocarse la muerte. Se han propuesto muchos ejemplos de ello, pero resulta difícil saber hasta qué punto el animal comprende el concepto de la muerte y actúa de modo consciente hacia su propia destrucción. Tal vez el ejemplo más fiable sea el que narra la etóloga Jane Goodall, pionera en el estudio de los chimpancés salvajes, cuando explica cómo el chimpancé Flint reaccionó ante la muerte de su madre Flo: «Para Flint, la muerte de Flo fue un golpe del que nunca se recuperó. Parecía como si sin su madre ya no tuviera voluntad de vivir. Se sentaba en la orilla del arroyo cerca del cuerpo de Flo, y de vez en cuando se le acercaba como si quisiera encontrar todavía una señal de vida. Pero el cuerpo de Flo permanecía inmóvil, frío y muerto, y finalmente Flint se alejó. Su depresión empeoró y dejó de comer hasta que cayó enfermo. A pesar de que intentamos ayudarle acompañándole y llevándole comida, Flint falleció unas tres semanas después de la muerte de Flo». Así que, si nos atenemos a esto, la autodestrucción es un acto que implica un grado elevado de autonomía del androide, pero esto no le hace más o menos humano. O al menos, no más humano de lo que quizás son algunos animales.