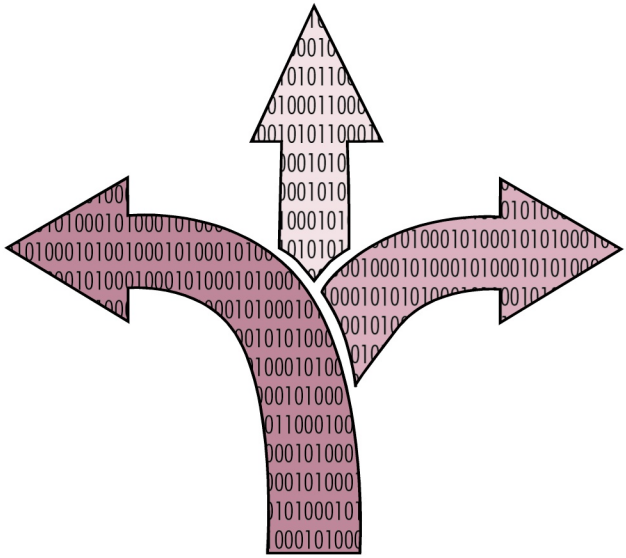
ÉTICA DE LA INTELIGENCIA ARTIFICIAL

Mark Coeckelbergh



CÁTEDRA

¿Es una IA «una simple máquina», o merece alguna forma de consideración moral? ¿Deberíamos tratarla de forma diferente a, por ejemplo, una tostadora o una lavadora?

este sentido, que es similar a, por ejemplo, la mayoría de los coches actuales: estos últimos también pueden generar problemas de índole moral. Pero dado que la IA se está volviendo cada vez más inteligente y autónoma, ¿puede tener una forma más fuerte de agencia moral? ¿Se le debería otorgar, o desarrollará la IA, alguna capacidad para el razonamiento moral, el juicio y la toma de decisiones? Por ejemplo: ¿pueden y deben los coches autónomos que usan IA ser considerados agentes morales? Estas cuestiones atañen a la ética de la IA, pues plantean qué tipo de capacidades morales tiene o debería tener una IA. Pero las cuestiones sobre el «estatus moral» pueden también referirse a cómo debemos tratar a una IA. ¿Es una IA «una simple máquina», o merece alguna forma de consideración moral? ¿Deberíamos tratarla de forma diferente a, por ejemplo, una tostadora o una lavadora? ¿Deberíamos otorgar derechos a una entidad artificial altamente inteligente, si tal entidad se desarrollase algún día, incluso si no fuese humana? Esto es lo que los filósofos denominan el problema de la *paciencia moral,* que no gira en torno a la ética *de* o *en* la IA sino a *nuestra* ética para *con* la IA. Aquí, la IA es el objeto de la preocupación ética, no un agente ético potencial en sí mismo.

Agencia moral

Comencemos con la cuestión de la agencia moral. Si existiese una IA mucho más inteligente que las que existen en la actualidad, podemos suponer que podría desarrollar un razonamiento moral y que podría aprender cómo toman decisiones los seres humanos sobre problemas éticos. ¿Pero, sería esto suficiente para otorgarle una agencia moral completa, esto es, una agencia moral similar a la humana? La cuestión no incumbe solo a la ciencia ficción. Si ya delegamos algunas de nuestras decisiones a algoritmos, por ejemplo en los coches o en los juzgados, sería positivo que estas decisiones fuesen moralmente correctas. Pero no está claro que las máquinas tengan las mismas capacidades morales que los seres humanos. Se les otorga agencia en el sentido de que hacen cosas en el mundo, y estas acciones tienen consecuencias morales. Por ejemplo, un coche autónomo puede causar un accidente, o una IA puede recomendar enviar a una persona concreta a la cárcel. Estos comportamientos y elecciones no son neutrales en términos de moral: claramente tienen consecuencias morales para la gente que está involucrada en

dichos acontecimientos. ¿Pero debería, para lidiar con este problema, otorgarse agencia moral a las IAs? ¿Pueden tener agencia moral completa?

Hay distintas posturas filosóficas en torno a estas cuestiones. Algunos dicen que las máquinas jamás podrán ser agentes morales. Las máquinas, argumentan, no tienen las capacidades necesarias para la agencia moral, tales como estados mentales, emociones o libre albedrío. De ahí que sea peligroso suponer que puedan tomar decisiones morales correctas y delegar totalmente en ellas dichas decisiones. Por ejemplo, Deborah Johnson (2006) ha argumentado que los sistemas de ordenadores no tienen agencia moral por sí mismos: están producidos y son utilizados por humanos, y solo estos humanos tienen libertad y son capaces de actuar y decidir moralmente. De forma similar, se podría decir que las IAs están creadas por humanos y que, por tanto, la toma de decisiones morales en prácticas tecnológicas debería ser ejercida por humanos. En el otro extremo del espectro están aquellos que piensan que las máquinas pueden ser agentes morales completos de la misma forma que lo son los seres humanos. Investigadores tales como Michael y Susan Anderson, por ejemplo, afirman que, en principio, es posible y deseable otorgar a las máquinas una moralidad de tipo humano (Anderson y Anderson 2011). Podemos dar principios morales a las IAs, y las máquinas pueden incluso ser mejores que los seres humanos en el razonamiento moral, ya que son más racionales y no se dejan llevar por sus emociones. Contra esta postura, algunos han argumentado que las reglas morales entran a menudo en conflicto (considérense, por ejemplo, las historias de robots de Asimov, en las que las leyes morales para los robots siempre meten en problemas a los robots y a los humanos) y que el proyecto completo de construir «máquinas morales» dándoles reglas se basa en suposiciones erróneas con respecto a la naturaleza de la moralidad. La moralidad no puede reducirse a seguir reglas y no es enteramente una cuestión de emociones humanas: pero estas pueden bien ser indispensables para el juicio moral. Si fuese posible la IA general, resultaría indeseable una «IA psicópata» que fuera perfectamente racional pero insensible a las preocupaciones humanas porque carece de emociones (Coeckelbergh 2010).

Según estas razones, podría o bien rechazarse en su conjunto la idea misma de la agencia moral completa, o bien adoptarse una posición intermedia: tenemos que dar a las IAs algún tipo de moralidad, pero no moralidad completa. Wendell Wallach y Colin Allen usan el término

«moralidad funcional» (2009, pág. 39). Los sistemas de IA necesitan alguna capacidad para evaluar las consecuencias éticas de sus acciones. La razón de esta decisión está clara en el caso de los coches autónomos: el coche probablemente estará involucrado en situaciones donde haya que tomar una decisión moral, pero no hay tiempo suficiente para una toma de decisión o intervención humana. A veces estas elecciones se presentan como dilemas. Los filósofos hablan del *dilema del tranvía,* llamado así a partir del experimento mental en el que un tranvía avanza sin control por unos raíles y alguien tiene que escoger entre no hacer nada, con lo que morirán cinco personas que están atadas a la vía, o tirar de una palanca que hace que el tranvía siga por otro raíl en el que solo hay atada una persona, a la que no se conoce. ¿Cuál es la solución moralmente correcta? De forma similar, los que proponen este enfoque argumentan que un coche autónomo puede tener que hacer la elección moral entre, por ejemplo, matar a unos peatones que cruzan la carretera y estrellarse contra un muro, matando, así, al conductor. ¿Qué debería escoger el coche? Parece que tendremos que tomar estas decisiones morales (de antemano) y asegurarnos de que los diseñadores las implementan en los coches. O, quizás, necesitemos construir coches con IAs que aprendan de las elecciones humanas. Sin embargo, se podría cuestionar si proporcionar reglas a las IAs es una buena forma de representar la moralidad humana, si es que la moralidad puede «representarse» y reproducirse, y si los dilemas del tranvía capturan de alguna forma algo que es central para la vida y la experiencia moral. O, desde una perspectiva totalmente diferente, se puede cuestionar si los humanos son, de hecho, buenos a la hora de tomar decisiones morales. ¿Por qué imitar en cualquier caso la moralidad humana? Los transhumanistas, por ejemplo, pueden argumentar que las IAs tendrán una moralidad superior porque serán más inteligentes que nosotros.

Este cuestionamiento del énfasis en lo humano nos lleva a otra posición que no necesita de una agencia moral completa, y que intenta alejarse de la posición ética antropocéntrica. Luciano Floridi y J. W. Sanders (2004) han apostado por una moralidad *a-mental,* que no se base en las características de los seres humanos. Podríamos hacer que la agencia moral dependiera de tener un nivel suficiente de interactividad, autonomía y adaptabilidad, así como de ser capaces de acción moralmente calificable. De acuerdo con estos criterios, un perro de rescate es un agente moral, pero también lo es un *bot* web con IA que filtra los emails no deseados. De forma similar, se podrían aplicar criterios no an-

tropocéntricos para la agencia moral de los robots, como propuso John Sullins (2006): si una IA es autónoma respecto de los programadores y podemos explicar su comportamiento atribuyéndole intenciones morales (como la intención de hacer bien o mal), y si se comporta de una forma que muestra un entendimiento de su responsabilidad frente a otros agentes morales, entonces esta IA es un agente moral. Desde esta perspectiva, pues, las IAs no necesitan de una agencia moral completa en tanto que esto implica agencia moral humana, sino que definen la agencia moral de una forma que es, en principio, independiente de la agencia moral humana completa y de las capacidades humanas necesarias para ello. Sin embargo, ¿sería suficiente tal agencia moral artificial si la juzgáramos desde los estándares morales humanos? El problema en la practica es que, por ejemplo, los coches autónomos pueden no ser lo bastante morales. La preocupación por lo que respecta a los principios es que aquí nos apartamos demasiado de la moralidad humana. Mucha gente piensa que la agencia moral es y debe estar vinculada a lo que entendemos por ser humano y ser persona. No están dispuestos a apoyar nociones posthumanistas o transhumanistas.

Paciencia moral

Otra controversia se refiere a la paciencia moral de la IA. Imaginemos que tenemos una IA superinteligente. ¿Es moralmente aceptable apagarla, «matarla»? Y algo más cercano a la IA actual: ¿está bien golpear a un robot con IA?[1]. Si las IAs van a ser parte de la vida cotidiana, como muchos investigadores predicen, entonces estos casos aparecerán inevitablemente y harán surgir la cuestión de cómo nos deberíamos comportar los humanos con estas entidades artificiales. Una vez más, sin embargo, no es necesario que busquemos respuestas en un futuro lejano o en la ciencia ficción. La investigación ha demostrado que la gente empatiza con los robots y que duda a la hora de «matarlos» o «torturarlos» (Suzuki et al. 2015; Darling, Nandy y Breazeal 2015), incluso si estos robots no tienen IA. Los humanos parecen necesitar muy poco

[1] Un caso del mundo real fue el perro robot Spot, al que patearon sus desarrolladores patearon para probarlo, algo que, sorprendentemente, encontró muchas reacciones empáticas: <https://www.youtube.com/watch?v=aR5Z6AoMh6U>.

de los agentes artificiales para proyectar sobre ellos personalidad o humanidad, o para empatizar con ellos. Si estos agentes se convirtieran ahora en IA, lo cual los haría potencialmente más parecidos a los humanos (o a los animales), es posible que esta cuestión de la paciencia moral se hiciera más acuciante. Por ejemplo, ¿cómo deberíamos reaccionar ante las personas que empatizan con la IA? ¿Están equivocados?

Decir que las IAs son simples máquinas y que la gente que empatiza con ellas está equivocada en su juicio, emociones y experiencia moral es, quizás, la posición más intuitiva. A primera vista, parece que no tenemos obligación alguna con respecto a las máquinas. Son cosas, no personas. Muchos investigadores de la IA siguen esta línea de pensamiento. Por ejemplo, Joanna Bryson ha argumentado que los robots son herramientas, tienen propietarios y no tenemos obligaciones hacia ellos (Bryson 2010). Aquellos que mantienen esta posición puede que también estén de acuerdo en que en el caso de que las IAs llegaran a ser conscientes, a tener estados mentales, y demás, deberíamos otorgarles un estatus moral, aunque dirán que esta condición no se cumple actualmente. Como hemos visto en los capítulos anteriores, algunos argumentarán que dicha situación no se dará nunca; otros que podría darse en principio, pero que no ocurrirá en un lapso de tiempo cercano. En cualquier caso, su respuesta a la pregunta sobre el estatus moral es que hoy en día y en el futuro próximo las IAs tienen que ser tratadas como cosas, a menos que se demuestre que son lo contrario.

Sin embargo, un problema de esta posición es que no explica ni justifica nuestras intuiciones y las experiencias morales que nos dicen que hay *algo* que está mal cuando se «maltrata» a una IA, incluso si esta no tiene propiedades similares a las humanas o animales como la consciencia o la sensibilidad. Para encontrar tales justificaciones se podría acudir a Kant, que argumentó que está mal dispararle a un perro, no porque hacerlo incumpla ninguna obligación para con el perro, sino porque tal persona «daña la amabilidad y las cualidades humanas que posee en sí misma, que debería ejercitar en virtud de sus obligaciones para con la humanidad» (Kant 1997). Hoy en día tendemos a pensar en los perros de otro modo (aunque no todo el mundo ni en todos los sitios), pero parece que el argumento podría aplicarse a las IAs: podríamos decir que no tenemos obligaciones con respecto a las IAs, pero aun así no deberíamos patearlas o «torturarlas» porque ello nos hace crueles para con los humanos. Se podría usar también el argumento de la ética de la virtud, que es también indirecto, ya que versa sobre los humanos,

Hay quien considera que «maltratar» a una IA está mal no porque se le haga ningún daño a la IA, sino porque en caso de hacerlo se daña nuestra integridad moral.

no sobre la IA: «maltratar» a una IA está mal no porque se le haga ningún daño a la IA, sino porque en caso de hacerlo se daña nuestra integridad moral. No nos hace mejores personas. Contra este enfoque podemos argüir que en el futuro algunas IAs puedan tener un valor intrínseco y merezcan nuestra atención moral, suponiendo que tengan cualidades tales como la sensibilidad. Una obligación indirecta o un enfoque basado en la virtud no parece tomarse en serio esta «otra» cara de la relación moral. Solo se preocupa por los humanos. ¿Qué pasa con las IAs? ¿Pueden las IAs o los robots ser *otros,* como ha preguntado David Gunkel (2018)? De nuevo, el sentido común apunta a que no: las IAs no reúnen las cualidades necesarias.

Desde un enfoque totalmente distinto se argumenta que la manera en que nos cuestionamos el estatus moral es problemática. El razonamiento moral usual sobre el estatus moral está basado en cuáles son las propiedades moralmente relevantes de las entidades en cuestión: por ejemplo, consciencia o sensibilidad. ¿Pero cómo sabemos si la IA realmente reúne o no las cualidades moralmente relevantes? ¿Estamos seguros en el caso de *los humanos?* Los escépticos dirían que no estamos seguros. Incluso sin dicha certeza epistemológica atribuimos un estatus moral a los humanos sobre la base de la apariencia. Puede que esto sucediera si en el futuro las IAs tuvieran una apariencia y comportamiento similares a los de los humanos. Parece que, independientemente de lo que los filósofos consideren moralmente *correcto,* los humanos continuarán atribuyendo estatus moral a máquinas y, en consecuencia, les concederán derechos. Además, si analizamos más de cerca cómo los humanos otorgamos estatus moral *de facto,* resulta que, por ejemplo, las relaciones sociales existentes y el lenguaje desempeñan un notable papel. Por ejemplo, si tratamos a nuestro gato con cariño, no es porque llevemos a cabo un juicio moral sobre él, sino porque hemos adquirido un tipo de relación social con él. Ya es una mascota y un compañero antes de que realicemos la operación filosófica de otorgarle un estatus moral (si es que llegamos alguna vez a sentir dicha necesidad). Y, si le ponemos a nuestro perro un nombre, entonces (en contraste con los animales sin nombre que nos comemos), ya le hemos conferido un estatus moral particular independientemente de sus cualidades objetivas. Bajo un enfoque de este tipo, relacional, crítico y no dogmático (Coeckelbergh 2012), podríamos concluir, del mismo modo, que los seres humanos otorgaremos el estatus a las IAs y que este dependerá de cuán integradas estén ellas en nuestra vida social, en el lenguaje y en la cultura humana.

Además, dado que tales condiciones son susceptibles de cambiar a lo largo de la historia (pensemos de nuevo en cómo tratábamos a los animales y pensábamos en ellos), quizás fuera necesaria alguna precaución antes de «fijar» el estatus moral de la IA en general o de cualquier IA concreta. ¿Y por qué incluso hablar de la IA en general o en abstracto? Parece que hay algo que no encaja en el proceso por el cual otorgamos estatus moral: para juzgarlo, tomamos en consideración la entidad fuera de su contexto relacional, y antes de que tengamos el resultado de nuestro procedimiento moral ya la tratamos, de forma jerárquica, condescendiente y hegemónica, como una entidad sobre la que tomamos decisiones los seres humanos como jueces superiores. Parece que antes de que llevemos a cabo el juicio sobre su estatus moral, ya nos hemos posicionado ante esa entidad y, quizás, incluso la hayamos violentado al tratarla como el objeto de nuestra toma de decisión, estableciéndonos a nosotros mismos como dioses de la Tierra, poderosos y omniscientes, que se reservan el derecho a conferir un estatus moral a otras entidades. También hacemos invisibles todos los contextos situacionales y sociales. Como en el caso del dilema del tranvía, hemos reducido la ética a una caricatura. Con tales razonamientos, los filósofos morales parecen hacer lo que los filósofos que siguen a Dreyfus acusan de hacer a los investigadores de la IA simbólica: formalizar y abstraer una gran fuente de experiencia moral y de conocimiento a costa de dejar fuera lo que nos hace humanos y, además, con el riesgo de pedir la cuestión misma del estatus moral de los no humanos. Independientemente de cuál «sea» el estatus moral actual de las IAs, si es que este puede acotarse por completo e independientemente de la subjetividad humana, merece la pena examinar con perspectiva crítica nuestra propia actitud moral y el proyecto del razonamiento moral abstracto mismo.

Hacia los problemas prácticos

Como demuestran las cuestiones examinadas en este capítulo y el previo, pensar sobre la IA no solo nos permite aprender cosas sobre ética en sí. También nos enseña cosas de nosotros mismos: de cómo pensamos y cómo nos relacionamos y deberíamos relacionarnos con los no humanos. Si estudiamos los fundamentos filosóficos de la ética de la IA, encontramos profundos desacuerdos sobre la naturaleza y el futuro de la humanidad, la ciencia y la modernidad. Cuestionar la IA abre un

abismo de problemas críticos sobre el conocimiento, la sociedad y la naturaleza de la moralidad humanas.

Los debates filosóficos en torno a esto son menos rebuscados y menos «académicos» de lo que se podría pensar. Continuarán resurgiendo cuando consideremos más adelante los problemas éticos, legales y políticos concretos suscitados por la IA. Si intentamos abordar temas como la responsabilidad y los coches autónomos, la transparencia del aprendizaje automático, la IA sesgada, o la ética de los robots sexuales, pronto nos encontraremos de nuevo confrontados con ellos. Si la ética de la IA quiere ser más que una lista de problemas a tener en cuenta, debería también aportar algo sobre tales cuestiones.

Dicho esto, es hora de volver a las cuestiones prácticas. Estas no conciernen ni a los problemas filosóficos planteados por una hipotética inteligencia artificial general ni a los riesgos que lleva aparejados una superinteligencia en un futuro lejano, ni cualesquiera otros monstruos espectaculares de la ciencia ficción. Por el contrario, afectan a las realidades menos visibles y probablemente menos atractivas, pero aun así importantes, de las IAs ya implantadas. La IA tal como ya funciona actualmente no desempeña el papel del monstruo de Frankenstein o del espectacular robot que amenaza la civilización, y es más que un experimento mental filosófico. La IA concierne a las tecnologías menos visibles, ocultas pero dominantes, poderosas, y cada vez más inteligentes que ya conforman nuestras vidas a día de hoy. Las éticas de la IA se ocupan de los desafíos planteados por la IA actual y del futuro cercano y su impacto en nuestras sociedades y en las democracias vulnerables. La ética de la IA se ocupa de la vida de la gente y de la política. Se ocupa de nuestra necesidad, como personas y sociedades, de lidiar con los problemas éticos de *ahora.*

Capítulo 5

La tecnología

Antes de discutir problemas éticos de la IA de forma más detallada y concreta, tenemos otra tarea pendiente para despejar el camino: una vez superada la exageración, nos hace falta una mejor comprensión de la tecnología y sus aplicaciones. Dejando de lado la ciencia ficción transhumanista y la especulación filosófica sobre la IA general, echemos un vistazo a lo que es y hace la IA actualmente. Ya que las definiciones de la IA y otros términos son controvertidas en sí mismas, no ahondaré demasiado en los debates filosóficos o en la contextualización histórica. Mi principal propósito es ofrecer al lector una idea de la tecnología en cuestión y de cómo se usa. Déjenme comenzar diciendo algo sobre la IA en general; el siguiente capítulo se centra en el aprendizaje automático, la ciencia de datos y sus aplicaciones.

¿Qué es la Inteligencia Artificial?

La IA se puede definir como una inteligencia desplegada o simulada por un código (algoritmos) o por máquinas. Esta definición de la IA plantea el problema de cómo definir la inteligencia. Hablando filosóficamente, es un concepto vago. Una comparación obvia es la inteligencia de tipo humano. Por ejemplo, Philip Jansen *et al.* definen la IA como «la ciencia y la ingeniería de máquinas con capacidades que se

consideran inteligentes según los estándares de la inteligencia humana» (2018, pág. 5). Desde esta perspectiva, la IA tiene como objeto el crear máquinas inteligentes que puedan pensar o (re)accionar como lo hacen los humanos. Sin embargo, muchos investigadores de la IA piensan que la inteligencia no necesita ser de corte humano y prefieren una definición más neutral que se formula en términos independientes de la inteligencia humana y de las metas relacionadas con la IA general o fuerte. Enumeran todo tipo de funciones cognitivas y tareas tales como aprendizaje, percepción, planificación, procesamiento del lenguaje natural, razonamiento, toma de decisiones y solución de problemas (la última a menudo se hace equivaler con la inteligencia *per se).* Por ejemplo, Margaret Boden declara que la IA «busca hacer que los ordenadores hagan el tipo de cosas que puede hacer la mente». A primera vista, esto parece sonar como si los humanos fueran el único modelo. Sin embargo, Boden enumera a continuación todo tipo de destrezas psicológicas como la percepción, la predicción y la planificación, que forman parte de los «entornos ricamente estructurados de distintas capacidades de procesamiento de la información» (2016, pág. 1). Y este procesamiento de la información no tiene por qué ser una cuestión exclusivamente humana. La inteligencia general, de acuerdo con Boden, no es humana por necesidad. Algunos animales también pueden ser considerados inteligentes. Y los transhumanistas sueñan con mentes futuras que ya no estén biológicamente integradas. Dicho esto, la meta de alcanzar capacidades similares a las humanas y, posiblemente, una inteligencia general de tipo humano ha sido parte de la IA desde el inicio.

La historia de la IA está estrechamente conectada con la informática y con disciplinas relacionadas como las matemáticas y la filosofía y, por tanto, se remonta al menos hasta el comienzo de la Edad Moderna (Gottfried Wilhelm Leibniz y René Descartes, por ejemplo) si no a la antigüedad, con sus historias sobre artesanos que creaban seres artificiales e ingeniosos artefactos mecánicos capaces de engañar a la gente (piensen en las figuras animadas de la antigua Grecia o en las figuras mecánicas con forma humana de la antigua China). Pero como disciplina propiamente dicha se considera que la IA comenzó en los años 50 del siglo pasado, tras la invención de los ordenadores programables en los años 40 y con el nacimiento de la disciplina de la cibernética, definida por Norbert Wiener en 1948 como el estudio científico del «control y comunicación en el animal y en la máquina» (Wiener 1948). Un momento importante en la historia de la IA fue la publicación en *Mind*

en1950 del artículo de Alan Turing «Computing Machinery and Intelligence» [«Maquinaria computacional e inteligencia»], en el que introdujo el famoso «test de Turing», aunque trataba más ampliamente la cuestión de si las máquinas pueden pensar y ya especulaba acerca de si las máquinas podrían aprender y llevar a cabo tareas abstractas. Pero el seminario Dartmouth que tuvo lugar en verano de 1956 en Hannover, New Hampshire, se considera generalmente el lugar de nacimiento de la IA contemporánea. Su organizador, John McCarthy, acuñó el término IA, y entre los participantes estaban incluidos los nombres de Marvin Minsky, Claude Shannon, Allen Newell y Herbert Simon. Dado que la cibernética parecía estar demasiado dedicada a las máquinas analógicas, la IA de Dartmouth se centró las máquinas digitales. La idea era *simular* la inteligencia humana (que no recrearla: el proceso no es el mismo que en los humanos). Muchos participantes pensaban que una máquina tan inteligente como un ser humano estaría a la vuelta de la esquina, y de hecho que no llevaría más que una generación.

Esta es la meta de la *IA fuerte.* La IA *fuerte* o *general* es capaz de llevar a cabo cualquier tarea cognitiva que puedan realizar los humanos, mientras que la IA *débil* o *estrecha* solo puede operar en ámbitos específicos como el ajedrez, la clasificación de imágenes, etc. Hoy por hoy, no hemos alcanzado la IA general y, como hemos visto en los capítulos anteriores, se duda de que jamás lleguemos a alcanzarla. A pesar de que algunos investigadores y empresas están intentando desarrollarla, especialmente aquellas que creen en la teoría computacional de la mente, la IA general no se vislumbra en el horizonte. De ahí que las cuestiones éticas y políticas del capítulo siguiente se centren en la IA débil o estrecha, que ya tenemos hoy en día y que es probable que se vuelva más poderosa y dominante en un futuro próximo.

La IA se puede definir o bien como una ciencia o bien como una *tecnología.* Su objetivo puede interpretarse como el de alcanzar una mejor explicación científica de la inteligencia y de las mencionadas funciones cognitivas. Puede ayudarnos a comprender mejor a los seres humanos y a otros seres dotados de inteligencia natural. En este sentido, es una ciencia y una disciplina que estudia sistemáticamente el fenómeno de la inteligencia (Jansen et al. 2018) y, a veces, la mente o el cerebro. Como tal, la IA está relacionada con otras ciencias como la ciencia cognitiva, la psicología, la ciencia de datos (véase más adelante) y, a veces, también la neurociencia, que llega a sus propias conclusiones acerca de la comprensión de la inteligencia natural. Pero la IA también puede bus-

car el desarrollo de tecnologías para diversos objetivos prácticos, o para «lograr cosas útiles», como dice Boden: puede darse en forma de herramientas, diseñadas por humanos, que generen una apariencia de inteligencia y de comportamiento inteligente con fines prácticos. Las IAs pueden hacer esto analizando el entorno (datos) y actuando con un grado importante de autonomía. A veces, los intereses científico-teóricos y los fines tecnológicos coinciden; esto sucede, por ejemplo, en la neurociencia computacional, que utiliza herramientas de la informática para comprender el sistema nervioso, o en proyectos particulares como el «Proyecto cerebro humano»[1] europeo, que implica a la neurociencia pero también a la robótica y a la IA. Algunos de estos proyectos combinan neurociencia con aprendizaje automático y con la llamada neurociencia del *big data* (por ejemplo, Vu et al. 2018).

De forma más general, la IA se basa en y está relacionada con muchas disciplinas, incluyendo las matemáticas (por ejemplo, la estadística), la ingeniería, la lingüística, la ciencia cognitiva, la informática, la psicología e, incluso, la filosofía. Como hemos visto, tanto los filósofos como los investigadores de la IA están interesados en comprender la mente y fenómenos como la inteligencia, la consciencia, la percepción, la acción y la creatividad. La IA está influenciada por la filosofía y viceversa. Keith Frankish y William Ramsey reconocen esta conexión con la filosofía, subrayan la multidisciplinariedad de la IA y combinan los aspectos científicos y tecnológicos cuando definen la IA como «un enfoque multidisciplinar para la comprensión, modelado y reproducción de la inteligencia y los procesos cognitivos mediante el uso de varios principios y dispositivos computacionales, matemáticos, lógicos, mecánicos e, incluso, biológicos» (2014, 1). La IA es, pues, tanto teórica como práctica, tanto ciencia como tecnología. Este libro se centra en la IA como una tecnología, en el sentido más práctico: no solo porque dentro de la IA se ha desplazado el foco en esta dirección, sino especialmente porque es sobre todo en esta forma en la que la IA conlleva consecuencias éticas y sociales (a pesar de que la investigación científica tampoco es totalmente neutral en cuanto a la ética).

Como tecnología, la IA puede adoptar varias formas y normalmente es parte de sistemas tecnológicos más amplios: algoritmos, máquinas, robots, etc. Por ello, aunque la IA puede tener que ver con «máquinas»,

[1] Véase <https://www.humanbrainproject.eu/en/>.

este término no solo se refiere a los robots, y menos todavía a robots humanoides. La IA puede estar integrada en muchos otros tipos de sistemas y dispositivos tecnológicos. Los sistemas de IA pueden adoptar la forma de un *software* que funciona en la web (por ejemplo, *bots* de *chat*, motores de búsqueda, análisis de imagen), pero también puede estar integrada en aparatos físicos como robots, coches, o en aplicaciones del «internet de las cosas»[2]. Para el internet de las cosas se usa a veces el término «sistemas ciberfísicos»: dispositivos que funcionan en, e interactúan con, el mundo físico. Los robots son un tipo de sistemas ciberfísicos que influyen directamente en el mundo (Lin, Abney y Bekey 2011).

Si la IA está integrada en un robot, a veces se dice que es una IA *corporeizada.* Al ejercer una influencia directa en el mundo físico, la robótica resulta altamente dependiente de sus componentes materiales. Pero cada IA, incluyendo el software activo en la web, «hace» algo y también tiene aspectos materiales, tales como el ordenador en el que se ejecuta, los aspectos materiales de la red y de la infraestructura de la que depende, y un largo etcétera. Esta cuestión hace que sea problemática la distinción entre, por una parte, las aplicaciones «virtuales» basadas en la web y el «software» y, por otra, las aplicaciones físicas o «hardware». El software de IA necesita infraestructura física y *hardware* para ejecutarse, y los sistemas ciberfísicos *son* «IA» solamente si están conectados al software pertinente. Además, fenomenológicamente hablando, el hardware y el software a veces se fusionan en nuestra experiencia y uso de dispositivos: no experimentamos un robot interactivo humanoide impulsado por IA, o un aparato de IA conversacional como Alexa, bien como software o bien como hardware, sino como un único dispositivo tecnológico (y, a veces, casi como una persona, como sucede, por poner un caso, con Hello Barbie).

Es probable que la IA ejerza una influencia importante en la robótica, por ejemplo, a través del progreso en el procesamiento de lenguajes naturales y en las comunicaciones que simulan a la humana. A menudo se les llama a estos robots «robots sociales» porque están diseñados para participar en la vida social ordinaria de los seres humanos, por ejemplo, interactuando, ya sea como acompañantes o como asistentes, con los

---

[2] Véase, por ejemplo, la definición de la IA del Grupo de Expertos de Alto Nivel de IA de la Comisión Europea (2018).

humanos de forma natural. Así, la IA puede fomentar un mayor desarrollo de la robótica social.

Sin embargo, independientemente de la apariencia y del comportamiento del sistema como un todo y su influencia en su entorno, que es muy importante fenomenológica y éticamente hablando, la base de la «inteligencia» de una IA es el software: un *algoritmo* o una combinación de algoritmos. Un algoritmo es un conjunto y secuencia de instrucciones, como una receta, que le dice qué hacer al ordenador, *smartphone,* máquina, robot, o cualquier cosa en la que esté integrado. Conduce a un resultado *(output)* particular basándose en la información disponible *(input).* Se utiliza para resolver un problema. Para entender la ética en la IA, necesitamos comprender cómo funcionan los algoritmos de IA y qué hacen. Diré más sobre este tema en este y en el siguiente capítulo.

Diferentes enfoques y subcampos

Existen diferentes tipos de IA. También se podría decir que hay diferentes *enfoques* o *paradigmas de investigación.* Como vimos a partir de la crítica de Dreyfus, por lo general, la IA ha sido fundamentalmente *simbólica.* Este fue el paradigma dominante hasta finales de los años 80. La IA simbólica se basa en las representaciones simbólicas de tareas cognitivas superiores tales como el razonamiento abstracto y la toma de decisiones. Por ejemplo, puede decidir basándose en un *árbol de decisiones* (un modelo de decisiones y sus posibles consecuencias, a menudo representado visualmente como un gráfico de flujo). Un algoritmo que hace esto contiene afirmaciones condicionales: reglas de decisión de forma *si* ... (condiciones) ... *entonces* ... (resultado). El proceso es determinista. Mediante el empleo de una base de datos que representa el conocimiento humano experto, una IA puede razonar utilizando una gran cantidad de información y actuar como un *sistema experto.* Puede tomar decisiones complejas o recomendaciones basándose en un amplio corpus de conocimiento, en muchas ocasiones difícil o imposible de revisar para los seres humanos. Los sistemas expertos se usan, por ejemplo, en el sector médico para la diagnosis y la planificación de tratamientos. Durante mucho tiempo este fue el tipo de software de IA con más éxito.

Hoy en día la IA simbólica sigue siendo útil, pero han aparecido nuevos tipos de IA, que pueden (o no) combinarse con ella y que a di-

ferencia de los sistemas expertos son capaces de aprender autónomamente a partir de datos. Esto se logra mediante un enfoque completamente distinto. El paradigma de investigación del *conexionismo* —que se desarrolló en los años 80 como una alternativa a lo que acabó llamándose la Inteligencia Artificial Anticuada (GOFAI, del inglés Good Old-Fashioned Artificial Intelligence)—, y la tecnología de las *redes neuronales* se basan en la idea de que, en lugar de representar funciones cognitivas superiores, necesitamos construir redes interconectadas basadas en unidades simples. Sus defensores afirman que es similar al modo en que funciona el cerebro humano: la cognición surge de interacciones entre unidades de procesado simples, llamadas «neuronas» (que, sin embargo, no son como las neuronas biológicas) y se utilizan muchas interconexiones neuronales. Este enfoque y esta tecnología se usan a menudo para el *aprendizaje automático* (véase el capítulo siguiente), también denominado *aprendizaje profundo (deep learning)* cuando las redes neuronales tienen varias capas de neuronas. Algunos sistemas son híbridos; AlphaGo de DeepMind, por ejemplo, lo es. El aprendizaje profundo ha permitido progresos en campos como el de la visión artificial o el del procesamiento de lenguajes naturales. El aprendizaje automático que emplea una red neuronal puede convertirse en una «caja negra» en el sentido de que, aunque los programadores conocen la arquitectura de la red, no está claro para otras personas qué es lo que ocurre en sus capas intermedias (entre el *input* y el *output)* y, por tanto, tampoco cómo se alcanza una decisión. Esta situación contrasta con la de los árboles de decisión, que son transparentes e interpretables y, por tanto, pueden ser revisados y evaluados por seres humanos.

Otro paradigma importante en la IA es el que usa enfoques corporeizados y situacionales, centrándose en las tareas motoras y en las interacciones más que en las llamadas tareas cognitivas superiores. Los robots construidos por los investigadores de IA, tales como el Rodney Brooks del MIT, no resuelven problemas usando representaciones simbólicas, sino interactuando con su entorno circundante. Por ejemplo, Cog, el robot humanoide construido por Brooks y desarrollado en los años 90, fue diseñado para que aprendiese a interactuar con el mundo tal como lo hacen los niños. Además, no son pocos los que piensan que la mente solo puede surgir de la vida. Así, para crear una IA, necesitamos intentar crear vida artificial. Algunos ingenieros adoptan un enfoque menos metafísico y más práctico: toman la biología como modelo a partir del cual desarrollar las aplicaciones tecnológicas prácticas. Tam-

bién hay IAs evolutivas que tienen la capacidad de evolucionar. Algunos programas, empleando los llamados algoritmos genéticos, pueden incluso cambiarse a sí mismos.

Esta diversidad de enfoques y funciones de la IA implica también que hoy en la actualidad existan varios *subcampos:* aprendizaje automático, visión artificial, procesamiento de lenguajes naturales, sistemas expertos, computación evolutiva, etc. Actualmente el énfasis a menudo se pone en el aprendizaje automático, pero esta es solo una área de la IA, incluso si estas otras áreas están conectadas a menudo con el aprendizaje automático. Se ha alcanzado un gran progreso recientemente en la visión artificial, el procesamiento de lenguajes naturales y en el análisis de grandes cantidades de datos mediante el aprendizaje automático. Este último puede utilizarse, por ejemplo, para procesar lenguajes naturales basándose en el análisis de fuentes habladas y escritas, como textos extraídos de internet. Este tipo de trabajo es el que dio lugar a los agentes conversacionales actuales. Otro ejemplo es el reconocimiento facial basado en la visión artificial y en el aprendizaje profundo, que puede usarse, por ejemplo, para vigilancia.

Aplicaciones e impacto

La tecnología de IA se puede aplicar en distintos campos, desde la fabricación industrial, pasando por la agricultura y el transporte, hasta el sistema sanitario, las finanzas, el *marketing,* el sexo y el entretenimiento, la educación y las redes sociales. En las ventas al por menor y el *marketing,* se usan recomendadores o sistemas de recomendación para influir en las decisiones de compra y para ofrecer publicidad dirigida a un objetivo concreto. En las redes sociales, la IA es capaz de impulsar *bots:* cuentas de usuarios que parecen ser de gente real pero que son, en realidad, software. Estos *bots* pueden publicar mensajes de contenido político o conversar con usuarios humanos. En atención médica, se emplea la IA para analizar datos de millones de pacientes. Los sistemas expertos también se siguen utilizando en dicha área. En finanzas, la IA se utiliza para analizar grandes conjuntos de datos para el análisis de mercado y los sistemas automáticos de *trading.* Los robots (de compañía) incluyen a menudo IA. Los pilotos automáticos y los coches autónomos usan IA. Los empresarios pueden usar IA para monitorizar a sus empleados. Los videojuegos tienen personajes que fun-

cionan con IA. Las IAs pueden componer música o escribir artículos de periódico. También pueden imitar voces de personas e, incluso, falsificar discursos.

Dadas sus numerosas aplicaciones, es probable que la IA tenga un impacto generalizado, ahora y en un futuro próximo. Considérese la policía predictiva y el reconocimiento de habla, que crean nuevas posibilidades de seguridad y vigilancia, el transporte *peer-to-peer* y los coches autónomos que pueden transformar ciudades enteras, el *trading* algorítmico de alta frecuencia que ya afecta a los mercados financieros, o las aplicaciones diagnósticas en el sector médico que influyen en la toma de decisiones de los especialistas. No deberíamos olvidar que el de la ciencia es uno de los sectores en los que más impacto tiene la IA: mediante el análisis de grandes conjuntos de datos, la IA puede ayudar a los científicos a descubrir conexiones que, de otra manera, habrían pasado por alto. Esto es aplicable en ciencias naturales como la física, pero también en ciencias sociales y en humanidades. La IA ha afectado con seguridad al campo emergente de las humanidades digitales, por ejemplo, pues nos permite aprender cosas nuevas de los seres humanos y sus sociedades.

La IA tiene también un impacto en las relaciones sociales y una influencia amplia en la sociedad, la economía y el medio ambiente (Jansen *et al*. 2018). Es probable que condicione las interacciones humanas y tenga un impacto en la privacidad. Se dice que potencialmente incrementará el prejuicio y la discriminación. Se ha predicho que conducirá a la pérdida de empleo y, quizás, a la transformación de la economía al completo. Podría incrementar la brecha entre ricos y pobres y entre poderosos e indefensos, acelerando la injusticia y la desigualdad. Las aplicaciones militares pueden cambiar la forma en que se libran las guerras gracias al desarrollo de armas letales automáticas. También tenemos que tener en cuenta el impacto medioambiental, que incluye el incremento del consumo de energía y de la contaminación. Más adelante examinaré con más detalle algunas de sus implicaciones éticas y sociales, centrándome en los problemas y en los riesgos de la IA. Pero es probable que la IA también nos traiga cosas positivas; por ejemplo, puede generar nuevas comunidades a través de las redes sociales, reducir tareas repetitivas y peligrosas al permitir que los robots las lleven a cabo, mejorar la cadena de suministros, reducir el consumo de agua, etc.

Pero no solo debemos preguntarnos por la naturaleza y el alcance de su impacto (positivo o negativo); es también importante preguntar-

¿Quién tendrá acceso a la tecnología y será capaz de obtener sus beneficios? ¿Quiénes serán capaces de fortalecerse gracias a la IA? ¿A quién se excluirá de estos logros?

se *a quién* afecta y de qué manera. Un impacto concreto puede resultar más positivo para unos que para otros. Hay muchas partes interesadas, desde trabajadores, pacientes y consumidores a gobiernos, inversores y empresas, y cada una de ellas puede verse afectada de forma distinta. Y estas diferencias en ganancias y vulnerabilidad ante los impactos de la IA surgen no solo dentro de los Estados, sino también entre países y regiones del mundo. ¿Beneficiará principalmente la IA a los países altamente avanzados y desarrollados? ¿Podría también beneficiar a gente menos educada y con menos ingresos, por ejemplo? ¿Quién tendrá acceso a la tecnología y será capaz de obtener sus beneficios? ¿Quiénes serán capaces de fortalecerse gracias a la IA? ¿A quién se excluirá de estos logros?

La IA no es la única tecnología digital que plantea estos problemas. Otras tecnologías de la información y de la comunicación digital también tienen un enorme impacto en nuestras vidas y sociedades. Como veremos, algunos problemas éticos de la IA no son exclusivos de esta. Por ejemplo, hay paralelismos con otras tecnologías de automatización. Considérense los robots industriales que, sin tener estatus de IA, generan desempleo. Y algunos de los problemas de la IA están relacionados con tecnologías asociadas a ella, como las redes sociales e internet, que, cuando se combinan con la IA, plantean nuevos desafíos. Por ejemplo, cuando las plataformas de redes sociales como Facebook la utilizan para saber más sobre sus usuarios, surge la preocupación por la privacidad.

Esta relación con otras tecnologías también significa que, a veces, la IA no es visible. Esto ocurre, en primer lugar, porque se ha convertido en una parte que ya está engranada en nuestra vida cotidiana. La IA se usa a menudo en aplicaciones nuevas y llamativas como AlphaGo. Pero no debemos olvidar que la IA ya impulsa plataformas de redes sociales, motores de búsqueda y otros medios y tecnologías que se han convertido en parte de nuestra experiencia de la vida cotidiana. La IA está en todas partes. La línea entre IA propiamente dicha y otras formas de tecnología puede ser borrosa, convirtiendo a la IA en invisible: si los sistemas de IA están integrados en las tecnologías, no solemos percatarnos su presencia. Y si sabemos que la IA está implicada, entonces es difícil decir si es la IA la que crea el problema o impacto, o si es la otra tecnología la que lo hace. En cierto sentido, no existe una «IA» en sí misma: la IA siempre se basa en otras tecnologías y está integrada en prácticas y procedimientos científicos y tecnológicos más amplios. Puesto que la IA también hace que surjan sus propios problemas éticos

No debemos olvidar que la IA ya impulsa plataformas de redes sociales, motores de búsqueda y otros medios y tecnologías que se han convertido en parte de nuestra experiencia de la vida cotidiana. La IA está en todas partes.

específicos, cualquier «ética de la IA» necesitará estar conectada con una ética más general de las tecnologías de la información y de la comunicación digitales, con una ética de la informática, etc.

Otro sentido en el que podemos afirmar que la IA no es algo que exista por sí solo es que la tecnología es siempre social y humana: la IA no tiene que ver solamente con la tecnología, sino también de lo que los humanos hacen con ella, cómo la usan, cómo la perciben y la experimentan, y cómo la integran en entornos sociotécnicos más amplios. Esto es importante para la ética (que trata también las decisiones humanas) e implica que en ella debe incluirse una perspectiva histórica y sociocultural. El actual revuelo que la IA genera en los medios no es el primero que provoca una tecnología puntera. Antes de la IA, las palabras clave eran «robots» o «máquinas». Y otras tecnologías avanzadas como la nuclear, la nanotecnología, internet y la biotecnología también han producido muchos debates. Merece la pena tener esto en la cabeza cuando se discute sobre la ética de la IA, ya que quizás podamos aprender algo de estas controversias. El uso y el desarrollo de la tecnología se producen en un contexto social. Como sabe la gente que trabaja en evaluación de la tecnología, cuando esta es nueva tiende a ser muy controvertida, pero una vez que la tecnología se integra en la vida cotidiana, la exageración y la controversia se desinflan significativamente. Es probable que esto también suceda con la IA. Si bien es cierto que dicha predicción no es una buena razón para abandonar la tarea de evaluar los aspectos éticos y sociales de las consecuencias de la IA, nos ayuda a ver la IA en contexto y, en consecuencia, a comprenderla mejor.

Capítulo 6

Que no se nos olvide la ciencia de datos

Aprendizaje automático

Dado que muchas cuestiones éticas sobre la IA conciernen a tecnologías que están basadas entera o parcialmente en el aprendizaje automático y están relacionadas con la ciencia de datos, merece la pena acercarse a esta tecnología y a esta ciencia.

El *aprendizaje automático* se refiere al software que puede «aprender». El término es controvertido: algunos dicen que no es aprendizaje real porque no tiene una cognición real: solo los humanos pueden aprender. En cualquier caso, el aprendizaje automático moderno guarda «poca o ninguna similaridad con lo que plausiblemente podría estar ocurriendo en las cabezas humanas» (Boden 2016, pág. 46). El aprendizaje automático está basado en la estadística: es un proceso estadístico. Puede usarse para varias tareas, pero la tarea subyacente es a menudo el reconocimiento de patrones. Los algoritmos pueden identificar patrones o reglas observando conjuntos de datos y utilizar estos patrones o reglas para explicarlos y hacer predicciones.

Esto se consigue autónomamente en el sentido de que ocurre sin instrucciones ni reglas directas dadas por los programadores. En contraste con los sistemas expertos, que dependen de especialistas humanos en el campo en cuestión, quienes explican las reglas a los programadores que, a su vez, codifican estas reglas, el algoritmo de aprendizaje

automático encuentra reglas o patrones que el programador no ha especificado. Solo se le da un objetivo o tarea. El software puede adaptar su comportamiento para cumplir mejor con los requisitos de la tarea. Por ejemplo, el aprendizaje automático puede ayudar a distinguir correo no deseado de e-mails importantes buscando entre un gran número de mensajes y aprendiendo cuáles cuentan como correo no deseado. Otro ejemplo: para desarrollar un algoritmo que reconoce imágenes de gatos, el programador no da un conjunto de reglas al ordenador para definir lo que son los gatos, sino que hace que el algoritmo cree su propio modelo de imágenes de gatos. Se optimizará para alcanzar la mayor tasa de predicción en un conjunto de imágenes de gatos y no-gatos. Así, el algoritmo trata de aprender a reconocer las imágenes de gatos. Los humanos le dan retroalimentación, pero no le aportan instrucciones o reglas específicas.

Los científicos solían crear teorías para explicar datos y hacer predicciones; en el aprendizaje automático, los ordenadores crean sus propios modelos que encajan con los datos. El punto de partida son los datos, no las teorías. En este sentido, los datos dejan de ser «pasivos» y pasan a ser «activos»: son los «datos mismos los que definen lo que hacer a continuación» (Alpaydin 2016, 11). Los investigadores entrenan al algoritmo usando conjuntos de datos existentes (por ejemplo, viejos e-mails) y a continuación el algoritmo puede predecir resultados a partir de nuevos datos (por ejemplo, nuevos mensajes recibidos) (CDT 2018). Identificar patrones en grandes cantidades de información *(big data)* es lo que se llama a menudo «minería de datos» *(data mining),* haciendo una analogía con la extracción de minerales valiosos de la tierra. Sin embargo, el término es engañoso porque la meta es la extracción de patrones a partir de los datos, es decir, de su análisis; no la extracción de los datos mismos.

El aprendizaje automático puede estar *supervisado,* lo que significa que el algoritmo se centra en una variable particular que se designa como blanco para la predicción. Por ejemplo, si el objetivo es dividir a la gente en categorías (por ejemplo, riesgo de seguridad alto o bajo), las variables que predicen estas categorías ya son conocidas, y el algoritmo aprende entonces a predecir la categoría de pertenencia (riesgo de seguridad alto/riesgo de seguridad bajo). El programador entrena al sistema dándole ejemplos y contraejemplos, tales como imágenes de gente que suponen un alto riesgo de seguridad y de personas que no lo suponen. La meta es que el sistema aprenda a predecir quién pertenece a cada

categoría, quién supone un riesgo de seguridad alto y quién no, basándose en nuevos datos. Si se le otorgan al sistema los ejemplos suficientes, será capaz de generalizar a partir de estos ejemplos y sabrá categorizar nuevos datos, tales como una imagen nueva de un pasajero que está pasando un control de seguridad en un aeropuerto. Si está *no supervisado* quiere decir que este tipo de entrenamiento no se hace y que las categorías no son conocidas: los algoritmos crean sus propios grupos *(clusters).* Por ejemplo, la IA crea sus propias categorías de seguridad basándose en variables que ella misma selecciona; el programador no las facilita. La IA puede encontrar patrones que los especialistas en el campo (en este caso, personal de seguridad) aún no ha identificado. Es posible que sus categorías parezcan completamente arbitrarias para los humanos. Puede que no tengan sentido, pero cabe identificarlas estadísticamente. En otras ocasiones tienen sentido, y entonces este método nos permite adquirir nuevos conocimientos sobre las categorías del mundo real. El *aprendizaje por refuerzo,* finalmente, requiere que se indique si el *output* es bueno o malo. Su lógica es análoga a la de la recompensa y el castigo. No se le dicen al programa las acciones que debe llevar a cabo, sino que este «aprende» a través de un proceso iterativo en el que las acciones producen recompensas. Retomando el ejemplo de la seguridad: el sistema recibe retroalimentación de (datos suministrados por) el personal de seguridad para que «sepa» si ha hecho un buen trabajo en el caso de una predicción particular. Si una persona de la que se ha predicho que plantea un riesgo de seguridad bajo no causa ningún problema de seguridad, el sistema obtiene la retroalimentación de que este *output* era correcto y «aprende» de él. Nótese que siempre hay un porcentaje de error: el sistema nunca es cien por cien preciso. Nótese también que los términos técnicos «supervisado» y «no supervisado» tienen poco que ver con cómo de involucrados estén los humanos en el uso de la tecnología: aunque al algoritmo se le dé autonomía, en los tres casos los humanos estarán involucrados de alguna manera.

Esto también se aplica a todo lo que respecta al uso de datos en IA, incluido el llamado *big data.* El aprendizaje automático basado en el *big data* ha despertado mucho interés debido a la disponibilidad de grandes cantidades de datos y a un incremento en la potencia de los ordenadores. Algunos investigadores hablan de un «terremoto de datos» (Alpaydin 2016, pág. x). Todos producimos datos derivados de nuestras actividades digitales, por ejemplo, cuando usamos las redes sociales o

cuando compramos productos *online.* Estos datos son de interés para los actores comerciales, pero también para los gobiernos y los científicos. Nunca la recogida, el almacenaje y el procesamiento de datos habían sido tan sencillos para las organizaciones (Kelleher y Tierney 2018). Este hecho no se debe solamente al aprendizaje automático: el entorno digital más amplio y otras tecnologías digitales han desempeñado un importante papel. Las aplicaciones *online* y las redes sociales hacen más sencillo recolectar datos de los usuarios. Además, es más barato almacenar datos y los ordenadores se han vuelto más potentes. Todos estos factores han sido importantes para el desarrollo de la IA en general, pero también para la ciencia de datos.

Ciencia de datos

El aprendizaje automático está así vinculado a la *ciencia de datos.* Esta busca extraer patrones útiles y significativos de los conjuntos de datos, que actualmente son enormes. El aprendizaje automático es capaz de analizar estos grandes conjuntos. Tanto este como la ciencia de datos están basados en la estadística, que parte de observaciones particulares para alcanzar descripciones generales. Los especialistas están interesados en encontrar correlaciones en los datos. El modelado estadístico busca las relaciones matemáticas entre el *input* y el *output,* tarea para la cual el aprendizaje automático es de gran utilidad.

Pero la ciencia de datos supone más que su simple análisis mediante el aprendizaje automático. Los datos tienen que ser recogidos y preparados antes de analizarse, y después tienen que interpretarse los resultados de los análisis. La ciencia de datos incluye desafíos tales como el de obtener y limpiar los datos (por ejemplo, de las redes sociales y de la web), conseguir los datos suficientes, agrupar conjuntos de datos, reestructurarlos, seleccionar los relevantes y cuál es el tipo que vamos a utilizar. De este modo, los humanos desempeñan aún un papel importante en todas las fases y en relación con todos estos aspectos, incluyendo el encuadre del problema, la obtención de datos, la preparación de estos (el conjunto de datos con el que se entrena al algoritmo y el conjunto al que se lo destinará), la creación o selección del algoritmo de aprendizaje, la interpretación de los resultados y la elección de una acción (Kelleher y Tierney 2018).

A lo largo de las páginas de este libro, se tratan, de manera concisa y accesible, los principales problemas éticos que el desarrollo de la inteligencia artificial y su aplicación a un gran número de ámbitos de nuestra vida cotidiana han planteado en los últimos años.

Huyendo del sensacionalismo y la exageración, Coeckelbergh despliega un mapa de los desafíos que supone esta tecnología, deteniéndose en aquellos que afectan transversalmente al desarrollo de las sociedades contemporáneas, y en los que, por estar directamente relacionados con la naturaleza de los seres humanos, presentan mayores dificultades.

¿A qué ética responde la decisión de una máquina? ¿En qué consiste exactamente tomar decisiones? ¿Podemos considerar a la máquinas responsables de sus actos y de las consecuencias que conllevan? ¿Cómo aprende y actúa una inteligencia artificial?

Estas preguntas y otras muchas exigen respuestas urgentes.

De la publicidad a los mercados financieros, de la industria armamentística a la de la automoción, de las redes sociales al internet de las cosas, la IA ocupa un lugar cada vez más relevante en nuestra vida, aunque solo en algunas ocasiones seamos conscientes de ello.

0112127

ISBN 978-84-376-4212-3

9 788437 642123

colección TEOREMA

serie mayor