Helga Nowotny

La fe en la inteligencia artificial

Los algoritmos predictivos y el futuro de la humanidad

Galaxia Gutenberg

HELGA NOWOTNY

La fe en la inteligencia artificial

Los algoritmos predictivos y el futuro de la humanidad

Traducción de Alfred Bosch

Galaxia Gutenberg

Título de la edición original: *In AI We Trust. Power, Illusion and Control of Predictive Algorithms*
Traducción del inglés: Alfred Bosch

Publicado por
Galaxia Gutenberg, S.L.
Av. Diagonal, 361, 2.º 1.ª
08037-Barcelona
info@galaxiagutenberg.com
www.galaxiagutenberg.com

Primera edición: octubre de 2022

Preimpresión: Fotocomposición Gama, SL
Impresión y encuadernación: Sagrafic
Depósito legal: B 12262-2022
ISBN: 978-84-18526-33-6

Introducción: un viaje personal a Digilandia

Orígenes: tiempo e incertidumbre. Ciencia, tecnología y sociedad

Este libro es el resultado de un largo viaje personal y profesional. Reúne dos vertientes de mi trabajo anterior, al abordar las principales transformaciones sociales que la humanidad está experimentando en este momento: los procesos en curso de digitalización y nuestra llegada a la época del Antropoceno. La digitalización nos empuja a un proceso coevolutivo entre humanos y máquinas, un proceso que va acompañado de hazañas tecnológicas sin precedentes, y de la confianza que depositamos en la inteligencia artificial (IA). Preocupan las continuas pérdidas de privacidad, cómo será el futuro del trabajo y los riesgos que la IA puede representar para las democracias libres. Todo esto genera sentimientos de ambivalencia generalizados: confiamos en la IA como una apuesta de futuro, pero a su vez nos damos cuenta de que hay motivos para la desconfianza. Estamos aprendiendo a vivir junto a los dispositivos digitales, con los que interactuamos alegremente, como si fueran nuestros nuevos parientes, nuestros *alter ego* digitales, manteniendo una profunda ambivalencia hacia ellos y hacia el entramado tecnoempresarial que los produce.

El proceso de digitalización e informatización coincide con la creciente conciencia de una crisis medioambiental. El impacto del cambio climático y la lamentable condición actual del ecosistema, del que dependemos para sobrevivir, exigen una acción urgente. Pero estamos igual de esclavizados o angustiados por las tecnologías digitales que se están generalizando en nuestras sociedades. Sin embargo, rara vez se reflexiona conjuntamente sobre estas

dos grandes transformaciones: la digitalización y la sostenibilidad. Nunca antes habíamos tenido los instrumentos tecnológicos y el conocimiento científico para ver tan bien el pasado y el futuro, ni habíamos contado con tantos recursos tecnocientíficos para actuar. Y, aun así, nos sentimos angustiados por nuestra existencia en este misterioso presente, que marca una transición hacia un futuro desconocido, que será diferente a lo que se nos prometió tiempo atrás. Tal sentimiento de ansiedad generalizado sólo se ha visto exacerbado por la pandemia de la COVID-19, en sí misma un hecho histórico traumático de consecuencias a largo plazo y a escala mundial.

El viaje que me llevó a escribir este libro fue largo y lleno de sorpresas. Para mi trabajo anterior sobre el tiempo, especialmente la estructura y la experiencia del tiempo social, tuve que indagar mucho. En concreto, sobre nuestra exposición e interacción diaria con la IA y los dispositivos digitales, que se han convertido en nuestros compañeros íntimos, y sobre cómo estos alteran todavía más nuestra experiencia del tiempo. Al contrastarlo con escalas de tiempo geológicas, procesos atmosféricos a largo plazo, o la vida media de la disolución de residuos microplásticos y tóxicos, ¿cómo afecta la IA a la temporalidad de nuestra vida diaria? ¿Cómo incide en la dimensión temporal de las relaciones entre las personas? ¿Estamos presenciando el surgimiento de algo que podemos llamar *tiempo digital,* que se entromete en las jerarquías temporales de siempre, de los tiempos físicos, biológicos y sociales? Si es así, ¿cómo nos enfrentamos y asumimos estos diferentes tipos de tiempo a medida que avanza nuestra vida?

Al abordar la otra vertiente de mi trabajo anterior, sobre la incertidumbre, dirigí mi investigación hacia formas de asumir y gestionar las viejas y nuevas incertidumbres, con la ayuda de poderosas herramientas informáticas que nos acercan al futuro. Estas herramientas permiten vislumbrar la dinámica de sistemas complejos y, en principio, nos facilitan identificar los puntos de inflexión en los que los sistemas hacen su transición para cambiar el estado en el que se encuentran. Los puntos de inflexión marcan una mayor transformación, incluida la posibilidad del derrumbe. A medida que la ciencia comienza a comprender sistemas complejos, ¿cómo

se puede aprovechar este conocimiento para contrarrestar los riesgos que aparecen y fortalecer la resiliencia de las redes sociales?

Como era de esperar, encontré obstáculos en mi camino, pero también me di cuenta de que mi antiguo interés en el estudio del tiempo y de la incertidumbre astuta (que implica aceptación), me permitían conectar ciertas facetas de mi propia biografía, mis vivencias y experiencias, con estudios empíricos y hallazgos científicos. Sin embargo, esos vínculos personales ya no servían para asumir las probables consecuencias del cambio climático, la pérdida de la biodiversidad y la acidificación del océano, o temas como el impacto de la digitalización sobre las profesiones liberales. Pude ver que había mucho en juego, como ven todos los que se enfrentan a alarmas mediáticas sobre catastróficos incendios forestales, inundaciones y el hielo ártico que se derrite rápidamente. Seguí leyendo informes científicos que hacen estimaciones cuantitativas y calculan el año en que alcanzaremos una mayor degradación del medio ambiente, y por tanto asistiremos al hundimiento del ecosistema. De nuevo, como tantos otros, me sentí expuesta a las preocupaciones y esperanzas, las oportunidades y los posibles contratiempos relacionados con el actual proceso de digitalización.

Entre todas esas observaciones y reflexiones, y al margen de la escala global en la que se desarrollan semejantes amenazas, preservé un espacio para refugiarme en mi vida personal segura de que, por fortuna, continuaría ahí sin mayores alteraciones. Incluso los aspectos locales se estaban produciendo en lugares remotos o seguían siendo locales, en el sentido de que pronto serían superados por otros eventos locales. La mayoría de nosotros somos conscientes de que estas importantes transformaciones sociales tendrán enormes impactos y numerosas consecuencias no deseadas, y, no obstante, las mantenemos en un nivel de abstracción tan abrumador que son difíciles de entender intelectualmente en toda su magnitud. La brecha entre conocer y actuar, entre la percepción personal y la acción colectiva, entre el pensamiento a nivel del individuo y las instituciones pensantes a nivel mundial, parece protegernos del impacto inmediato que estos cambios de gran alcance puedan tener.

Por último, me llamó la atención que hay una manera de conectar la investigación científica rigurosa e impulsada por la cu-

riosidad con la experiencia personal y la intuición sobre lo que está en juego: se trata del papel cada vez más importante que desempeña la predicción; en particular, los algoritmos predictivos y analíticos. La predicción, obviamente, se refiere al futuro, pero versa sobre cómo vemos el futuro desde el presente. Cuando se aplica a sistemas complejos, la predicción debe lidiar con la falta de linealidad de los procesos. En un sistema no lineal, los cambios en los *inputs* ya no son proporcionales a los cambios en los *outputs*. Esa es la razón por la que tales sistemas parecen imprevisibles o caóticos. Y es el punto donde nos encontramos ahora: queremos ampliar el rango de lo que se puede predecir de manera fiable, pero también nos damos cuenta de que los sistemas complejos desafían la linealidad que, tal vez como herencia de la modernidad, aún sustenta gran parte de nuestro pensamiento.

El comportamiento de los sistemas complejos nos resulta difícil de comprender y, a menudo, nos parece contrario a la intuición lógica. Está ejemplificado por el famoso efecto mariposa, cuando una respuesta sensible, dependiendo de las condiciones iniciales, puede acabar con grandes diferencias en una etapa posterior. Como cuando el aleteo de una mariposa en el Amazonas llega a provocar un tornado que arrasa Texas. Pero tales metáforas no siempre ayudan, y empecé a preguntarme si en realidad somos capaces de pensar de manera no lineal. Las predicciones sobre el comportamiento de los sistemas dinámicos complejos a menudo se presentan en forma de ecuaciones matemáticas aplicadas a las tecnologías digitales. Los modelos de simulación no nos hablan claro y directo; sus resultados y las opciones que producen deben interpretarse y explicarse. Dado que se perciben como científicamente objetivos, a menudo no se cuestionan. Pero, entonces, las predicciones adquieren el poder activo que les atribuimos. Si se sigue ciegamente, el poder predictivo de los algoritmos se convierte en una profecía de autocumplimiento: una predicción se cumple porque la gente cree en ella y actúa en consecuencia.

Así, me propuse salvar la brecha entre el nivel personal, en este caso las predicciones que recibimos como individuos, y lo colectivo, representado por sistemas más complejos. Nos sentimos cómodos con mensajes conocidos y comunicaciones que nos llegan

a nivel personal, mientras que, a menos que adoptemos una postura profesional y científica, vivimos todo lo relacionado con sistemas complejos como una fuerza externa e impersonal. ¿No podría ser, me preguntaba, que se nos convenza tan fácilmente de confiar en un algoritmo predictivo porque nos llega a nivel personal? Y al mismo tiempo, tal vez desconfiamos del sistema digital, sea lo que sea que entendamos como tal, porque lo percibimos como impersonal.

En ciencia, hablamos de diferentes niveles, organizados de manera jerárquica, y cada nivel sigue sus propias reglas o leyes. En las ciencias sociales, incluida la economía, la brecha persiste en forma de división a nivel micro y macro. Pero ninguna de las consideraciones epistemológicas que siguen parecían proporcionar lo que yo buscaba: un modo de ver a través de estas divisiones, ya fuera cambiando de perspectiva o, mucho más ambicioso, tratando de encontrar varias perspectivas que me permitieran acceder a ambos niveles. Por lo tanto, he tratado de encontrar una manera de combinar lo personal y lo impersonal, el impacto de los algoritmos predictivos en nosotros como individuos, pero también los efectos que la digitalización causa en nosotros como sociedad.

Aunque la mayor parte de este libro la escribí antes de que un nuevo virus causara estragos en todo el mundo, agravado por unas políticas descoordinadas y a menudo irresponsables, sin duda el resultado está marcado por el impacto de la pandemia de la COVID-19. Inesperadamente, la crisis del coronavirus reveló las limitaciones de las predicciones. Una pandemia es una de esas incógnitas previsibles que se espera que ocurran. Se sabe que es probable que aparezcan, pero se desconoce cuándo y dónde. En el caso del virus SARS-CoV-2, la brecha entre las predicciones y la falta de preparación pronto se hizo evidente. Estamos preparados para creernos ciegamente las predicciones que los algoritmos arrojan sobre lo que debemos consumir, sobre cuál tiene que ser nuestro comportamiento e incluso nuestro estado mental emocional en el futuro. Creemos lo que nos dicen sobre los riesgos para la salud y los avisos sobre la necesidad de cambiar nuestro estilo de vida. Tales datos se utilizan para la elaboración de perfiles policiales, sentencias judiciales y mucho más. Y, sin em-

bargo, no estábamos preparados en lo más mínimo para una pandemia que se había pronosticado mucho tiempo atrás. ¿Cómo ha podido fallar todo?

Así pues, la crisis de la COVID-19, que lo más probable es que pase de ser una emergencia a ser una situación endémica, fortaleció mi convicción de que la clave para comprender los cambios que estamos viviendo está vinculada a lo que llamo la paradoja de la predicción. Cuando el comportamiento humano, por flexible y adaptativo que sea, comienza a ajustarse a lo que anuncian las predicciones, corremos el riesgo de volver a un mundo determinista, en el que el futuro ya está fijado. La paradoja se encuentra en la relación dinámica pero volátil entre el presente y el futuro: las predicciones, como es evidente, son sobre el futuro, pero actúan directamente sobre cómo nos comportamos en el presente.

El poder predictivo de los algoritmos nos permite ver más allá y prever los efectos de las pautas emergentes, dentro de sistemas complejos obtenidos a través de modelos de simulación. Respaldados por una enorme potencia informática, y entrenados en una ingente cantidad de datos extraídos del mundo natural y social, podemos trazar algoritmos predictivos y analizar su impacto. Pero la manera en que hacemos esto es paradójica en sí misma: anhelamos conocer el futuro, pero nos desentendemos de cómo las predicciones nos afectan en el presente. ¿Qué creemos, pues, y qué descartamos? La paradoja surge de la incompatibilidad entre una función algorítmica, que al fin y al cabo es una ecuación matemática abstracta, y esas creencias humanas lo bastante poderosas para impulsarnos (o no) a actuar.

Los algoritmos predictivos han adquirido un poder poco común que se expresa en varias dimensiones. Hemos llegado a confiar en ellos bajo formas que incluyen predicciones científicas con una amplia gama de aplicaciones, como la mejora de las previsiones meteorológicas o los numerosos instrumentos tecnológicos diseñados para abrir nuevos mercados. Se basan en técnicas de análisis predictivo que han dado como resultado una amplia gama de productos y servicios, desde el análisis de muestras de ADN para predecir el riesgo de determinadas enfermedades, hasta aplicaciones en política (se ha llegado a apuntar a grupos específicos de vo-

tantes, cuyo perfil se ha establecido a través de bases de datos, algo que se ha convertido en una característica habitual de las campañas). Las predicciones se han vuelto omnipresentes en nuestra vida diaria. Regalamos nuestros datos personales a cambio de conveniencia, eficiencia y ahorro en los productos que nos ofrecen las grandes empresas. Alimentamos su insaciable apetito por más datos y les confiamos información sobre nuestros sentimientos y comportamientos más íntimos. Parece que nos hemos adentrado en un camino irreversible de confianza en tales compañías. El análisis predictivo prevalece en los mercados financieros, donde se instalaron hace mucho tiempo las evaluaciones de riesgo automatizadas de comercio y tecnología financiera. También es la columna vertebral del desarrollo militar de armas robotizadas, cuyo despliegue real constituiría una auténtica pesadilla.

Sin embargo, la pandemia de la COVID-19 ha revelado que el control es mucho menor de lo que pensábamos. Esto no se debe a algoritmos defectuosos ni a falta de datos, aunque la pandemia ha evidenciado hasta qué punto se subestima la importancia del acceso a datos de calidad y su interoperabilidad. No hubo algoritmos predictivos cuando se advirtió de posibles epidemias; los modelos epidemiológicos y la estadística bayesiana fueron suficientes. Pero las advertencias no fueron escuchadas. La brecha entre saber y actuar seguirá existiendo si la gente no quiere saber o encuentra muchas excusas para justificar su inacción. Por tanto, las predicciones deben verse siempre en su contexto. Pueden caer en el vacío o llevarnos a seguirlas a ciegas. La analítica predictiva, aun cuando se expresa como una derivada de nuestra ignorancia, viene como un paquete digital que recibimos con gusto, pero que rara vez nos vemos en la necesidad de desempaquetar. Tiene la apariencia de productos algorítmicos refinados, producidos por un sistema que parece impenetrable para la mayoría de nosotros y, a menudo, guardado celosamente por las grandes empresas que lo poseen.

Así pues, las observaciones realizadas durante mi viaje intelectual empezaron a centrarse en el poder de la predicción y, en especial, en el poder ejercido por los algoritmos predictivos. Esto me permitió preguntarme: ¿cómo cambia la inteligencia artificial nuestra concepción del futuro y nuestra experiencia del tiempo?

Podría volver a mi antiguo compromiso con el estudio del tiempo social y, en particular, el concepto de *Eigenzeit*, que fue el tema de un libro que escribí a finales de los años ochenta. Hace unos años hice una revisión del *Eigenzeit*, y analicé los cambios debidos al uso de los medios y dispositivos digitales, que para entonces ya se habían convertido en nuestros compañeros cotidianos (Nowotny 2017). Tenemos nuevas relaciones temporales con aquellos que están físicamente distantes pero digitalmente cercanos, de modo que la ausencia y la presencia, así como la ubicación física y digital, han conducido a una experiencia del tiempo alterada.

No podíamos ni imaginar el significado que términos como distanciamiento físico y social han adquirido hoy. En medio de la pandemia de la COVID-19, vi confirmado mi diagnóstico anterior sobre un presente prolongado. Mi argumento era que la línea que separaba el presente del futuro se estaba diluyendo a medida que la dinámica de la innovación, encabezada por la ciencia y la tecnología, abría el presente a las muchas opciones nuevas que se hacían disponibles. El presente se ampliaba a medida que había que adaptarse a la selección y la apropiación social de las nuevas tecnologías. Gran parte de lo que parecía posible sólo en un futuro lejano de pronto invadía el presente. Esto alteró la experiencia del tiempo. El presente se estaba comprimiendo y densificando mientras se proyectaba hacia el futuro inmediato (Nowotny 1989).

Lo que veo ahora es que ya ha llegado el futuro. Vivimos no sólo en una era digital, sino en una máquina del tiempo digital. Una máquina alimentada por algoritmos predictivos que producen la energía para empujarnos más allá del futuro que ya ha llegado, hacia un futuro desconocido que queremos dilucidar desesperadamente. Por tanto, nos apresuramos a compilar pronósticos y a participar en múltiples ejercicios de previsión, tratando de obtener una medida de control sobre lo que de otro modo parece incontrolable debido a su complejidad. Los algoritmos y análisis predictivos nos brindan tranquilidad al trazar las trayectorias para el comportamiento futuro. Les atribuimos poderes y nos sentimos apoyados por los mensajes que transmiten sobre las incógnitas que más nos preocupan. Nuestro anhelo de certeza es tal

que incluso en los casos en que el pronóstico es negativo nos sentimos aliviados de saber lo que sucederá. Al ofrecer tal seguridad, las predicciones algorítmicas pueden ayudarnos a hacer frente a la incertidumbre y, al menos en parte, devolvernos algo de control sobre el futuro.

Mi experiencia en estudios de ciencia y tecnología (STS, por sus siglas en inglés) me permitió vencer la desconexión entre ciencia y sociedad y alcanzar una mejor comprensión de las fricciones y los malentendidos mutuos que acosan a esta relación endeble y tensa. Los STS abren la posibilidad de observar cómo se lleva a cabo la investigación en la práctica y hacen que podamos ver las estructuras y los procesos sociales que sustentan el funcionamiento de la ciencia. El virus simplemente ha agregado un nuevo giro, aunque algo desafortunado. Si bien al comienzo de la pandemia la ciencia ocupó el centro del escenario, combinada con la expectativa de que pronto se desarrollaría una vacuna y varias curas terapéuticas, pronto se vio envuelta en el oportunismo político. Surgió un desagradable *nacionalismo de la vacuna*, mientras que la ciencia fue eludida por los negacionistas de la COVID-19, así como las teorías conspiratorias –que comenzaron a florecer junto con los movimientos políticos antivacunas y de extrema derecha–. Después de un breve y brillante interludio, la relación entre ciencia, política y público volvió a ser problemática.

La pandemia ofreció un campo de pruebas avanzado; especialmente para la biomedicina, para la que la inteligencia artificial y las tecnologías digitales más recientes demostraron ser de gran ayuda. Se pudo secuenciar el genoma del virus y sus mutaciones posteriores en un tiempo récord, con investigadores que compartían muestras en todo el mundo y reciclaban equipos en sus laboratorios para aportar más recursos en los análisis. Se creó el High Performance Consortium COVID-19, una iniciativa público-privada con los grandes maestros de la IA y la NASA, agregando la capacidad informática de los ordenadores más rápidos y avanzados del mundo. Con la ayuda de los métodos *deep learning*, aprendizaje profundo, fue posible pasar de mil millones de moléculas analizables a unos pocos miles, sin afectar el valor médico de las pruebas.

La respuesta a la pandemia también asignó un papel mucho más importante a los datos. La presión fue enorme para avanzar lo más rápido posible con los datos disponibles, a fin de incorporarlos a los modelos de simulación que los informáticos, epidemiólogos y matemáticos usaban para hacer pronósticos. El objetivo era predecir las diversas trayectorias que podría tomar la pandemia, trazando el aumento, caída o aplanamiento de las curvas y analizando las implicaciones para los diferentes grupos de población, la infraestructura sanitaria, las cadenas de suministro y el previsible daño colateral socioeconómico. Sin embargo, a pesar del papel importante y visible que se les dio a los datos a lo largo de la pandemia, no surgió una rápida solución cuantitativa de datos que proporcionara una base sólida para las medidas que se deberían tomar. Si la calidad de los datos es mala o no existe el tipo correcto de datos, un buen instrumento puede convertirse rápidamente en basura que contamina los modelos de simulación y que reduce su utilidad social de forma drástica.

Hasta cierto punto, la crisis de la COVID-19 ha eclipsado la discusión en curso sobre la innovación y sobre los hallazgos científicos que hay que transferir a la sociedad. Por tanto, es apropiado recordar el trabajo de los profesionales de STS que han analizado extensamente la configuración social de las tecnologías. Sus hallazgos demuestran que las tecnologías se aplican de forma selectiva. Tienen género. Se traducen en productos que abren nuevos mercados y que dan un nuevo impulso al capitalismo global. Los beneficios de la innovación tecnológica nunca se distribuyen por igual, y las desigualdades sociales ya existentes se hacen más profundas con el cambio tecnológico acelerado. Pero nunca es la tecnología sola la que actúa como una fuerza externa que provoca el cambio social. Más bien, las tecnologías y el cambio tecnológico son consecuencia de condiciones previas sociales, culturales y económicas, y resultado de muchos procesos coproductivos.

Visto desde una perspectiva STS, lo que aparenta ser completamente nuevo y único debe contextualizarse en términos históricos y comparativos. La transformación actual se puede comparar con cambios de paradigma tecnoeconómico anteriores, que también tuvieron profundos impactos en la sociedad. En la era mo-

derna, el progreso se concibió como lineal y unidireccional. La creencia general, liderada y respaldada por las tecnociencias, era que el crecimiento económico aseguraría un futuro mejor y más brillante. Creencia aparejada a la noción de control, expresada a partir de un exceso de confianza en la planificación. Esta creencia en el progreso, sin embargo, ha ido decayendo durante algún tiempo y, más recientemente, muchos acontecimientos y novedades han inyectado otras dudas. La destrucción del medio ambiente a escala global nos confronta a todos con una verdad inconveniente, confirmada por el movimiento Fridays for Future que ha galvanizado a generaciones más jóvenes. Además, la pandemia ha demostrado la impotencia de muchos gobiernos y el cinismo de sus reacciones; hacer frente a las consecuencias exigirá un cambio de dirección.

La notable velocidad de los avances recientes en IA y su convergencia con la crisis de la sostenibilidad invita a preguntarnos: ¿Qué es diferente esta vez? Ya somos conscientes de las limitaciones de nuestro hábitat espacial, y nos enfrentamos a múltiples desafíos a la hora de utilizar los recursos disponibles de manera sostenible. Tenemos que asumir desde la gestión de una transición a energías limpias, pasando por el mantenimiento de la biodiversidad y la mejora de la habitabilidad de las ciudades, hasta echar un freno drástico a la contaminación por plásticos y gestionar la creciente cantidad de residuos. No es de extrañar que crezca la preocupación sobre un menor control. Se espera que las máquinas que hemos creado se hagan cargo de muchos trabajos que actualmente realizan los humanos, pero nuestra capacidad de control se reducirá aún más porque estas máquinas monitorearán y limitarán nuestras acciones y posibilidades. Por estas razones, se necesitará mucha sabiduría para comprender mejor cómo la IA afecta y limita la acción humana.

Pronto me di cuenta de que había tocado sólo la superficie de procesos de transformación más profundos en los que tendremos que pensar juntos. El futuro estará dominado por las tecnologías digitales mientras nos enfrentamos a una crisis de sostenibilidad, y ambas transiciones están vinculadas con cambios en las estructuras temporales que dan forma a nuestras vidas y a la sociedad. Las tecnologías digitales arrastran el futuro hacia el presente, mientras

que la crisis de la sostenibilidad nos enfrenta al pasado y nos desafía a desarrollar nuevas capacidades para el futuro. Cualquier solución que encontremos debe integrar la dimensión humana y nuestra relación alterada con un medio ambiente tecnológicamente transformado. Estas fueron algunas de las preguntas subyacentes que me mantuvieron activa, tarareando en voz baja pero persistentemente mientras proseguía con mi búsqueda. Mi viaje me llevó a una serie de reuniones, talleres y conferencias internacionales donde se debatieron algunos de estos temas. Por ejemplo, hubo reuniones sobre cómo proteger los derechos a la privacidad, que recibieron un estatus legal especial en Europa a través del Reglamento General de Protección de Datos (RGPD). Se percibe que Europa desempeña sólo un papel secundario en la competencia geopolítica entre las dos superpotencias de inteligencia artificial, Estados Unidos y China. Una competencia a la que a veces se hace referencia como la carrera por la supremacía en armamentos digitales del siglo xxi, y que se ha reavivado recientemente de un modo alarmante. Muchos europeos se consuelan con el hecho de que al menos tienen un sistema regulatorio para protegerlos, incluso si reconocen que ni el RGPD ni otras formas de vigilancia contra la intrusión de las grandes corporaciones transnacionales son suficientes en la práctica.

Otros puntos en la agenda de los foros de discusión sobre digitalización se referían a los riesgos derivados de los procesos de automatización. El principal fue el tema candente del futuro del empleo y los riesgos potenciales que la digitalización conlleva para las democracias libres. Me parecía que el temor a que se perdieran más puestos de trabajo de los que se pudieran crear se estaba sintiendo con mucha más fuerza en Estados Unidos. En Europa no tanto, en parte debido a las disposiciones europeas en materia de bienestar social todavía existentes, y en parte porque la digitalización aún no se había hecho visible. El fenómeno afectaba más a los profesionales y a la clase media, y la amenaza a la democracia liberal se hizo más evidente cuando olas populistas, nacionalistas y xenófobas se extendieron por muchos países. Fueron alimentadas por pautas siniestras como las *fake news* y los troyanos, con piratas informáticos desconocidos y presuntos ser-

vicios secretos extranjeros –involucrados en grupos concretos de microincidencia con mensajes inventados–. En términos más generales, parecían decididos a socavar las instituciones democráticas existentes al mismo tiempo que apoyaban a los líderes políticos con tendencias autoritarias. Las tecnologías digitales y las redes sociales se estaban explotando para socavar los principios democráticos y el estado de derecho, mientras que internet, al parecer, se había convertido en un espacio desenfrenado y no regulado para la difusión del odio y el desprecio.

Mis visitas regulares a Singapur me proporcionaron un ángulo diferente para ver cómo las sociedades podrían adoptar la digitalización; fue una oportunidad única de observar en acción un país digital y económicamente avanzado. Reuní información sobre el tan cacareado sistema educativo de Singapur y observé la dependencia de la burocracia en las tecnologías digitales, pero también sus altos estándares de eficiencia y el nivel de confianza en el Gobierno. Sin embargo, lo que más me impresionó fue el delicado y siempre precario equilibrio del país entre un sentido ampliamente compartido de su vulnerabilidad –pequeño, sin recursos naturales y rodeado de vecinos grandes y poderosos– y la determinación igualmente compartida de estar bien preparados para el futuro. Estaba ante un país que se percibía a sí mismo como una nación todavía joven, que obtenía gran parte de su energía de la notable riqueza económica y el bienestar social logrado. Esta energía tenía que ser canalizada hacia un futuro que estaban decididos a moldear. En ningún otro lugar encontré tantos debates, talleres, informes y medidas políticas centradas en un futuro que, a pesar de ser incierto, debía deliberarse y planificarse cuidadosamente, teniendo en cuenta los muchos imprevistos que surgirían. Por supuesto, el futuro sería digital. Debían cultivarse las habilidades digitales necesarias y ponerse en práctica todas las herramientas digitales disponibles.

Pude obtener conocimientos y recoger observaciones al asistir a reuniones internacionales sobre el futuro de la inteligencia artificial. En virtud de mi puesto anterior como presidenta del Consejo Europeo de Investigación (ERC, por sus siglas en inglés), participé en varias reuniones del Foro Económico Mundial (WEF, por

sus siglas en inglés). El WEF quiere mostrarse muy comprometido con la construcción del futuro digital. En las reuniones a las que asistí, reconocidas figuras del mundo de la tecnología y los negocios se mezclaron con académicos e investigadores corporativos que trabajan a la vanguardia de la IA. Era obvio que el entusiasmo por las oportunidades que ofrecían las tecnologías digitales debía sopesarse con sus posibles riesgos si los gobiernos y el mundo empresarial querían evitar una reacción violenta de los ciudadanos, siempre preocupados por el ritmo del cambio tecnológico. Se exploraron muchas incertidumbres con respecto a cómo se desarrollaría todo, pero las soluciones ofrecidas fueron pocas.

Otras reuniones en las que participé tenían el objetivo explícito de involucrar al público en general en un debate sobre el futuro de la IA, como el Nobel Week Dialogue 2015 en Gotemburgo, o el Falling Walls Circle en Berlín en 2018. También hubo visitas a laboratorios TIC y talleres de robótica encargados de establecer varios tipos de estrategias digitales. Aproveché mucho las discusiones en curso con colegas del Vienna Complexity Science Hub y miembros de su red internacional, lo que me permitió vislumbrar la ciencia de la complejidad. Por casualidad me topé con una conferencia reveladora sobre humanismo digital, una tendencia que se está expandiendo gradualmente para convertirse en un auténtico movimiento.

Aunque la mayoría de estas charlas fueron dispersas y poco concluyentes, proyectaron la imagen de un campo dinámico que avanzaba con rapidez. Los principales protagonistas estaban ansiosos por retratar su trabajo como la responsabilidad de avanzar hacia una IA beneficiosa o iniciativas similares. Hubo una impaciencia notable por demostrar que los investigadores y promotores de la IA eran conscientes de los riesgos involucrados, pero la línea entre la preocupación sincera y los intentos poco sinceros de las grandes corporaciones de reclamar la propiedad ética a menudo también se difuminaba. De hecho, puede que la inteligencia humana sea burlada algún día por la IA, pero los comentaristas rara vez se detienen a estudiar la diferencia entre ambas. En cambio, ofrecen garantías de que los riesgos podrán gestionarse. De vez en cuando se toca el tema de la estupidez humana y el papel

que desempeña la ignorancia. Y a veces brilla una fascinación por la tecnología amable, similar a la que describió J. Robert Oppenheimer cuando habló de su obsesión por la bomba atómica.

En una de las muchas conferencias a las que asistí sobre el futuro de la IA, los organizadores habían decidido utilizar un algoritmo para maximizar la diversidad dentro de cada grupo. A la IA también se le encomendó la tarea de crear cuatro haikus diferentes, uno para cada grupo (por cierto, la primera vez que la IA logró llevar a cabo una tarea tan creativa fue en la década de 1960). La conferencia fue un éxito y las discusiones dentro de cada grupo de haikus fueron gratificantes, pero de alguna manera me sentí insatisfecha con el haiku que la IA había producido para mi grupo. Entonces, en el avión de regreso, decidí escribir uno yo misma, el primero de mi vida. Con la suerte del principiante, la última línea de mi haiku afirmaba que «el futuro requiere sabiduría».

Se dice que un haiku captura un momento fugaz, una impresión pasajera o una sensación efímera. Obviamente, mis impresiones estaban relacionadas con el tema de la conferencia, el futuro de la IA. «El futuro requiere sabiduría»; la frase se me quedó grabada. ¿Qué futuro me preocupaba tanto? ¿Estaría dominado por algoritmos predictivos? Y si es así, ¿cómo cambiaría esto el comportamiento humano y nuestras instituciones? ¿Qué podría hacer yo para proyectar algo de sabiduría al futuro? En mi viaje a Digilandia he aprendido a escuchar con atención las disonancias y matices y sondear los medios tonos, para detectar las ambigüedades y ambivalencias en nuestros enfoques de los problemas que tratamos, y para perfeccionar la capacidad de deslizarnos entre los recuerdos selectivos del pasado, un presente que nos abruma y un futuro incierto y abierto.

El bucle y el laberinto

Ninguno de estos encuentros y discusiones me preparó para la sorpresa que tuve al empezar a revisar sistemáticamente las fuentes disponibles. Ya se ha publicado mucho sobre el tema, y hay un flujo constante de textos actualizados que siguen apareciendo.

Llegué a la conclusión de que gran parte de ellos deben haberse escrito de forma apresurada, como si se tratara de ponerse al día de las novedades en tiempo real. A veces me sentía como si sufriera un atracón involuntario, un empacho de información superflua que me dejaba intelectualmente desnutrida. Lo más sorprendente es el hecho de que la mayoría de los libros en esta área abrazan o bien una visión optimista y tecnoentusiasta o bien una visión fatalista. A menudo se basan en especulaciones o simplemente describen a un público no especializado lo que están haciendo los friquis de la IA, y cómo las tecnologías digitales cambiarán la vida de las personas. Acabé con una profunda insatisfacción acerca de cómo se estaban tratando los problemas y temas que consideraba importantes: el enfoque era en gran medida a corto plazo y ahistórico, superficial y básicamente especulativo. A menudo se adoptaba una perspectiva disciplinaria estrecha, incapaz de conectar los desarrollos tecnológicos con los procesos sociales de una manera significativa, y, en ocasiones arrogante, al descartar lo social o malinterpretarlo como un mero apéndice de lo tecnológico.

Muchos libros sobre inteligencia artificial y digitalización inundan el mercado. La mayor parte están escritos en un estilo entusiasta y amigable con la tecnología, pero también hay otro enfoque que abraza el lado oscuro de las tecnologías digitales. Los primeros proporcionan una descripción general amplia de los últimos desarrollos en IA y sus beneficios económicos, o muestran algunas características agregadas recientemente destinadas a aliviar los temores de que las máquinas pronto se harán cargo de todo. Se reconoce el impacto social de la IA, así como la conveniencia del diálogo interdisciplinario. Un guiño a las consideraciones éticas se ha vuelto obligatorio, pero se esquivan los problemas y se espera que se resuelvan en otra parte. Rara vez, por ejemplo, escuchamos consideraciones sobre la justicia social digital. Encontrar mi camino a través de la abundante literatura sobre IA a veces era como estar metida en un bucle, en un embrollo deliberadamente confuso diseñado para atraparte.

En este bucle hay muchos tramos bien iluminados; sus paredes están revestidas con los últimos artilugios y exhiben con orgullo

características diseñadas para llevar al usuario a un país de las maravillas virtual. Las arboledas más oscuras están llenas de imágenes y terribles vaticinios de cosas peores, y proyectan a veces un final digital apocalíptico. La ciencia ficción ocupa varios recodos, a menudo expresados en una sobrecarga de imaginación tecnológica y una carencia de aspectos sociales. En medio, una gran cantidad de senderos mundanos, algunos de los cuales resultan ser callejones sin salida. También se encuentran consejos útiles sobre cómo lidiar con las molestias esenciales provocadas a diario por las tecnologías digitales, o sobre cómo trabajar con el sistema. Una gran cantidad de *marketing* impregna el bucle, transmitiendo una sensación de entusiasmo a corto plazo y ganas tremendas de volver a animarse para ofrecer una nueva y mayor dosis de progreso digital.

A veces sentía que ya no estaba atrapada en un bucle, sino en lo que se había convertido en un auténtico laberinto. Era el caso de los libros que viraban hacia la singularidad y el transhumanismo, conceptos que pueden adquirir fácilmente un estatus de culto, al estar impregnados de teorías, fantasías y especulaciones de cómo la especie humana pronto superará sus actuales limitaciones cognitivas y físicas. En contraste con un bucle de características enredadas y retorcidas, callejones sin salida y caminos serpenteantes, un laberinto está cuidadosamente diseñado para tener un centro al que se puede llegar siguiendo un único camino. Está organizado de manera ingeniosa, y a menudo lúdica, con figuras geométricas como un círculo o una espiral. No es de extrañar que los laberintos hayan inspirado a muchos escritores y artistas a jugar con estas formas y con un viaje cargado de significado. Si el punto de salida y llegada es el mismo, se supone que el viaje nos habrá transformado. Por lo general, la transformación ocurrirá dentro de nosotros mismos. De ahí la estrecha asociación del laberinto con un estado superior de conciencia o iluminación espiritual.

El laberinto es un mito clásico, que implica transformación, aunque sepamos poco sobre los rituales asociados a él. En la era digital, el centro imaginado del laberinto digital o informático es la zona en que la IA desborda a la inteligencia humana, también

llamada singularidad. En ese punto, la mente humana se fusionaría con una mente superior creada artificialmente, y el cuerpo humano frágil y envejecido podría quedar atrás. El cuerpo y el mundo material se descartan a medida que el ser recién nacido es absorbido por el mundo digital o un orden superior. Aquí nos encontramos con una antigua fantasía, el sueño recurrente de la inmortalidad nacido del deseo de volverse como los dioses, esta vez reinventados como los maestros del universo digital. Me llamó la atención lo cerca que se podía combinar la discusión de temas trascendentales, como la inmortalidad o la búsqueda del alma, con temas muy tecnológicos y con la informática a ras de suelo. Me pareció que el bucle podía transformarse de repente en un laberinto y viceversa.

En la práctica, sin embargo, surgen fuertes contradicciones. Aquellos que se preocupan por los riesgos potenciales de las tecnologías digitales en las democracias liberales descubren que los expertos que trabajan sobre los riesgos se interesan poco por la democracia o la política. Quienes escriben sobre el futuro del trabajo rara vez hablan con quienes están diseñando los sistemas automatizados que dejarán a las personas sin empleo –o que crearán nuevos puestos de trabajo–. Muchos informáticos y expertos en TIC son claramente conscientes de las distorsiones y errores en sus productos, lamentando a la vez las limitaciones asociadas a los amplios sistemas tecnológicos. Pero en el fondo están convencidos de que las soluciones a muchos de los problemas que aquejan a la sociedad surgirán de la tecnología. Mientras tanto, los humanistas se retiran a su santuario histórico y actúan en defensa de los valores humanos. Al parecer, el sueño interdisciplinario, tantas veces reivindicado, no ha avanzado mucho en la práctica.

Al salir del bucle, tuve la sensación de que se trataba de un mercado sobrevalorado, donde los productos existentes son rápidamente desplazados por otros, seleccionados sólo por ser novedosos. Dependiendo del estado de ánimo de los compradores potenciales, se impondría o bien la visión utópica o bien la distópica, ambas sujetas a la volatilidad del mercado. El laberinto, por supuesto, constituiría un lugar más intrigante y encantador, donde las profundas preguntas filosóficas se cruzan con las especula-

ciones más descabelladas. Allí, por momentos, me sentía como Ariadna, siguiendo los hilos que me sacarían del centro del laberinto. Uno de esos hilos se basa en la idea de un humanismo digital, una visión de que los valores y perspectivas humanas deben ser el punto de partida para el diseño de algoritmos y sistemas de IA que pretendan servir a la humanidad. Se basa en la convicción de que semejante alternativa es posible.

Otro hilo está entretejido con la orientación que se inspira en un notable descubrimiento humano: la idea del futuro como un horizonte abierto, lleno de posibilidades aún inimaginables e intrínsecamente incierto. El horizonte abierto se extiende hacia el vasto espacio de lo desconocido, palpitando con la dinámica de lo posible. La creatividad humana está lista para explorarlo, con la ciencia y el arte a la vanguardia. Esta concepción del futuro es la que está en juego cuando los algoritmos predictivos amenazan con llenar el presente con su aparente certeza, y cuando el comportamiento humano comienza a ajustarse a estas predicciones.

El marco más amplio de este libro lo establece una trayectoria coevolutiva en la que se ha embarcado la humanidad junto con las máquinas digitales que ha inventado y desplegado. La coevolución significa que se está gestando una interdependencia mutua, con adaptaciones flexibles en ambos lados. Seres o entidades digitales como los robots, creados por nosotros, están mutando para convertirse en nuestros destacados *otros*. No tenemos ni idea de adónde conducirá semejante viaje, ni de cómo terminará. Sin embargo, en el curso prolongado de la evolución humana, es posible que nos hayamos convertido en algo parecido a una especie que se autodomestica y que aprende a valorar la cooperación o, al menos hasta cierto punto, a disminuir su potencial agresivo. Esta capacidad de cooperación podría extenderse ya a las máquinas digitales. Hemos llegado al punto de empezar a creer que el algoritmo nos conoce mejor de lo que nos conocemos a nosotros mismos, lo cual lo convierte en una nueva autoridad para guiar al yo, uno que sabe lo que es bueno para nosotros y lo que nos depara el futuro.

Lo que nos depara el camino: avanzar mirando atrás

Las predicciones científicas se consideran el sello distintivo de la ciencia moderna. En particular, la física avanza al inventar nuevos conceptos teóricos e instrumentos para las predicciones derivadas de ellos. La revolución informática que comenzó a mediados del siglo pasado se ha visto impulsada por el enorme aumento de la potencia de los ordenadores y los métodos de aprendizaje profundo que despegaron en el siglo xxi. Junto con el acceso a una cantidad de datos sin parangón, todavía en aumento, estos desarrollos han extendido el poder de las predicciones y su aplicabilidad a una enorme variedad de fenómenos naturales y sociales. Las predicciones científicas ya no se limitan a la ciencia.

Desde entonces, la analítica predictiva se ha vuelto altamente rentable para la economía y ha invadido todo el tejido social. La operación de algoritmos es la base del funcionamiento de productos tecnológicos que han alterado los modelos comerciales y creado nuevos mercados. Aprovechado por la industria del *marketing* y la publicidad, instrumentalizado por políticos que buscan maximizar los votos, y rápidamente adoptado por el oscuro mundo de los servicios secretos, *hackers* y estafadores que explotan el anonimato de internet, el uso del análisis predictivo ha convencido a consumidores, votantes y ciudadanos que creen en la bondad de estos poderosos instrumentos digitales, que se supone que están ahí para satisfacer nuestras necesidades y deseos latentes.

Gran parte de su difusión exitosa y adopción entusiasta se debe al hecho de que el poder de los algoritmos predictivos es eficaz. Un algoritmo tiene la capacidad de hacer que suceda lo que predice cuando el comportamiento humano se somete a la predicción. La llamada *performatividad* significa que lo que se promulga, pronuncia o realiza puede afectar a la acción, como se muestra en el trabajo pionero sobre los actos del habla y la comunicación no verbal de J. L. Austin, Judith Butler y otros. Otro fenómeno social bien conocido se capta en el teorema de Thomas («Si los hombres definen situaciones como reales, son reales en sus consecuencias») que se remonta a 1928 y que posteriormente fue reformulado por Robert K. Merton en términos de profecía de auto-

cumplimiento. Ha llegado el momento de reconocer lo que los sociólogos saben desde hace mucho tiempo y aplicarlo también a los algoritmos predictivos.

La propensión de las personas a orientarse en relación con lo que hacen los demás, en especial en circunstancias inesperadas o amenazantes, aumenta el poder de los algoritmos predictivos. Magnifica la ilusión de tener el control. Pero si el instrumento gana en comprensión perdemos la capacidad de pensamiento crítico. Terminamos confiando en el piloto automático mientras volamos a ciegas en la niebla. Sin embargo, hay situaciones en las que es crucial desactivar el piloto automático y ejercer nuestro propio juicio sobre lo que debemos hacer.

Al visualizar el camino por delante, veo una situación en la que hemos creado un instrumento altamente eficiente que nos permite seguir y prever la dinámica en evolución de una amplia gama de fenómenos y actividades, pero en la que en gran medida no entendemos las causas. Dependemos cada vez más de lo que nos dicen los algoritmos predictivos, sobre todo cuando las instituciones comienzan a alinearse con sus predicciones, a menudo sin darse cuenta de las consecuencias no deseadas que seguirán. Confiamos no sólo en el poder performativo de la analítica predictiva, sino también en que sabe qué opciones presentarnos, de nuevo sin considerar quién ha diseñado estas opciones y cómo, o que podría haber otras opciones igualmente dignas de considerar.

Al mismo tiempo, surge la desconfianza en la IA y aumentan las preocupaciones. Algunas de estas, como los temores sobre la vigilancia o el futuro del trabajo, son bien conocidas y ampliamente discutidas. Otras no son tan obvias. Cuando las profecías autocumplidas comienzan a proliferar, corremos el riesgo de volver a una cosmovisión determinista en la que el futuro aparece como prescrito y, por tanto, cerrado. El espacio vital para imaginar lo que podría ser de otra manera comienza a encogerse. La motivación y la capacidad de ampliar los límites de la imaginación se reducen. Depender sólo de la eficacia de la predicción oculta la necesidad de comprender por qué y cómo. El riesgo es que todo lo que atesoramos sobre nuestra cultura y nuestros valores se pueda atrofiar.

Además, en un mundo gobernado por la analítica predictiva, no existe ni lugar ni obligación de rendir cuentas. Cuando el poder político deja de rendir cuentas a aquellos sobre quienes se ejerce, corremos el riesgo de destruir la democracia. La rendición de cuentas se basa en una comprensión básica de causa y efecto. En una democracia, esto se enmarca en términos legales y es una parte integral de las instituciones democráticamente legitimadas. Si esto ya no está garantizado, el control se vuelve omnipresente. Los macrodatos aumentan aún más y los datos se adquieren sin comprensión ni explicación. Nos convertimos en parte de un sistema predictivo interconectado y afinado que se cierra dinámicamente sobre sí mismo. La capacidad humana de enseñar a otros lo que sabemos y hemos experimentado comienza a parecerse a la de una máquina que puede enseñarse a sí misma e inventar las reglas. Las máquinas no tienen empatía ni sentido de la responsabilidad. Sólo los humanos pueden rendir cuentas y sólo los humanos tienen la libertad de asumir responsabilidades.

Por fortuna, todavía no hemos llegado a ese extremo. Todavía podemos preguntarnos: ¿de verdad queremos vivir en un mundo completamente previsible donde el análisis predictivo invada y guíe nuestros pensamientos y deseos más íntimos? Eso significaría renunciar a la incertidumbre inherente del futuro y reemplazarla con la peligrosa ilusión de tener el control. ¿O estamos dispuestos a reconocer que nunca se puede lograr un mundo previsible del todo? Entonces tendríamos que reunir el valor para asumir los riesgos de un mundo falsamente determinista. Este libro se ha escrito como un argumento contra la ilusión de un mundo totalmente previsible y a favor de la sabiduría y el valor necesarios para vivir con incertidumbre.

Por supuesto, mi viaje no acaba ahí. «La vida sólo puede entenderse mirando atrás, pero debe vivirse hacia delante». Esta cita de Søren Kierkegaard espera ser contrastada con nuestros movimientos entre los mundos *online* y *offline*, entre el yo virtual, el yo imaginado y el yo real. ¿Cómo se puede vivir en las condiciones vigentes, con todas sus oportunidades y limitaciones? La cita implica una disyunción entre la Vida como abstracción que trasciende lo personal y el vivir como experiencia consciente que llena

cada momento de nuestra existencia. Con el estupendo conocimiento que ahora tenemos sobre la Vida en toda su diversidad, formas y niveles, sobre sus orígenes en el pasado profundo y su continua evolución, ¿no es ahora el momento de hacer que este conocimiento influya sobre cómo vivir hacia delante? La especie humana ha superado la evolución biológica –cuyo producto todavía somos–. La ciencia y la tecnología nos han permitido avanzar a una velocidad acelerada por los caminos de una evolución cultural que cada vez somos más capaces de moldear.

Y, sin embargo, aquí estamos, lidiando con una crisis de sostenibilidad global llena de consecuencias nefastas y crecientes tensiones geopolíticas. Mientras escribo, estamos en las garras de una pandemia, a la que seguirán otras si se continúan erosionando los hábitats naturales de los animales que portan virus zoonóticos propagables entre humanos. Las deficiencias de nuestras instituciones, creadas en siglos anteriores y diseñadas para afrontar retos distintos al nuestro, nos miran a la cara. El espectro del malestar social y las sociedades polarizadas ha regresado, cuando lo que se necesita es una mayor coherencia social, igualdad y justicia social si queremos escapar de nuestro estado actual.

Nos hemos embarcado en un viaje para seguir adelante con algoritmos predictivos que nos permiten ver más allá. Afortunadamente, somos cada vez más conscientes de lo crucial que es el acceso a datos de calidad del tipo correcto. Somos cautelosos acerca de la erosión adicional de nuestra privacidad y reconocemos que la circulación de mentiras deliberadas y discursos de odio en las redes sociales representan una amenaza para la democracia. Confiamos en la IA y, al mismo tiempo, desconfiamos de ella. Es probable que esta ambivalencia perdure, ya que por inteligentes que sean los algoritmos cuando avanzamos hacia el futuro en la era digital, no van más allá de encontrar correlaciones.

Incluso las redes neuronales más sofisticadas, que son versiones simplificadas del cerebro, sólo pueden detectar regularidades e identificar patrones basados en datos que provienen del pasado. No está involucrado ningún razonamiento causal, ni una IA pretende que lo sea. ¿Cómo podemos seguir adelante si no entendemos la vida tal como ha evolucionado en el pasado? Algunos in-

formáticos, como Judea Pearl y otros, deploran la ausencia de una búsqueda de relaciones causa-efecto. La «inteligencia real», argumentan, implica comprensión causal. Para que la IA llegue a tal etapa debe poder razonar de una manera contrafáctica. No es suficiente ajustar simplemente una curva a lo largo de una línea de tiempo indicada. Hay que abrir el pasado para entender una frase como «qué hubiera pasado si...». La acción humana consiste en lo que hacemos, pero comprender lo que hicimos en el pasado para poder hacer predicciones sobre el futuro siempre debe involucrar el contrafactual de que podríamos haber actuado de manera diferente. Al transferir un proceso humano a una IA debemos asegurarnos de que tenga la capacidad de discernir esta cualidad que es básica para la comprensión y el razonamiento humanos (Pearl y Mackenzie 2018).

El poder de los algoritmos es tan grande que olvidamos con facilidad la importancia del vínculo entre comprensión y predicción. Los usamos para hacer previsiones prácticas y calculables que son útiles en nuestra vida diaria, ya sea en la gestión de los sistemas de salud, en el comercio financiero automatizado, para hacer negocios más rentables o para expandir las industrias creativas. Pero no debemos ceder a la conveniencia de la eficiencia y abandonar el deseo de comprender, ni la curiosidad y la perseverancia que sustentan tal deseo (Zurn y Shankar 2020).

Hace tiempo que existen dos modos diferentes de pensar sobre cómo avanzar. Una línea de pensamiento remonta su linaje a la antigua fascinación por los autómatas y, de manera más general, al buen funcionamiento de las máquinas que han impulsado las revoluciones tecnológicas, con sus líneas de producción automatizadas dedicadas a aumentar la eficiencia y reducir los costos. Aquí es donde entran todas las promesas de la automatización, expresadas en sueños desorbitados e imaginarios tecnológicos. Los algoritmos de aprendizaje profundo continuarán equipando a los ordenadores con una comprensión estadística del lenguaje y, por tanto, ampliarán su capacidad de «razonamiento». Existe confianza entre los profesionales de la IA en que el trabajo en IA ética está progresando. La suposición tácita es que el lado oscuro de las tecnologías digitales y todos los proble-

mas hasta ahora no resueltos también serán resueltos por una inteligencia definitiva, una especie de Leviatán benigno y con visión de futuro, apto para manejar nuestras preocupaciones y guiarnos a través de los conflictos y desafíos a los que se enfrenta la humanidad del siglo xxi.

La otra línea de pensamiento insiste en que la comprensión teórica es necesaria y urgente, no sólo para los matemáticos y científicos informáticos, sino para desarrollar herramientas que evalúen el rendimiento y la calidad de salida de los algoritmos de, aprendizaje profundo, optimizando su entrenamiento. Semejante línea exige valentía para abordar las difíciles preguntas de por qué y cómo, y reconocer tanto los usos como las limitaciones de la IA. Dado que los algoritmos tienen enormes implicaciones para los humanos, será importante que sean justos y alinearlos con los valores humanos. Aunque podemos predecir con seguridad que los algoritmos darán forma al futuro, la cuestión de qué tipos de algoritmos darán esa forma sigue abierta todavía (Wigderson 2019).

La comprensión también incluye la expectativa de que podamos aprender cómo funcionan las cosas. Si un sistema de inteligencia artificial pretende resolver problemas al menos tan bien como un humano, entonces no hay razón para no esperar y exigir transparencia y responsabilidad de él. En realidad, estamos muy lejos de recibir respuestas satisfactorias sobre cómo funcionan las representaciones internas de la IA en detalle, y todavía más lejos de resolver las preguntas sobre causa y efecto. Empezamos a darnos cuenta de que estamos a punto de perder algo vinculado a nuestra condición humana, por complicado que nos resulte saber exactamente de qué se trata.

Quizá ha llegado el momento de admitir que no tenemos el control de todo, de admitir con humildad que el frágil y arriesgado viaje de coevolución con las máquinas que hemos construido será más fecundo si renovamos los intentos de comprender nuestra humanidad y nuestra comunidad. De saber cómo podríamos vivir mejor juntos. Tenemos que continuar nuestra exploración para avanzar en la vida, mientras tratamos de mirar atrás hacia lo que hemos vivido, y unir ambas visiones. En tal caso, la predic-

ción dejará de trazar únicamente las trayectorias hacia nuestro futuro, y se convertirá en una parte integral de la comprensión sobre *cómo* avanzar y vivir mejor. En lugar de predecir lo *que* sucederá, nos ayudará a comprender *por qué* suceden las cosas.

Después de todo, lo que nos hace humanos es nuestra capacidad única de hacernos la pregunta: *¿por qué suceden las cosas... por qué y cómo?*