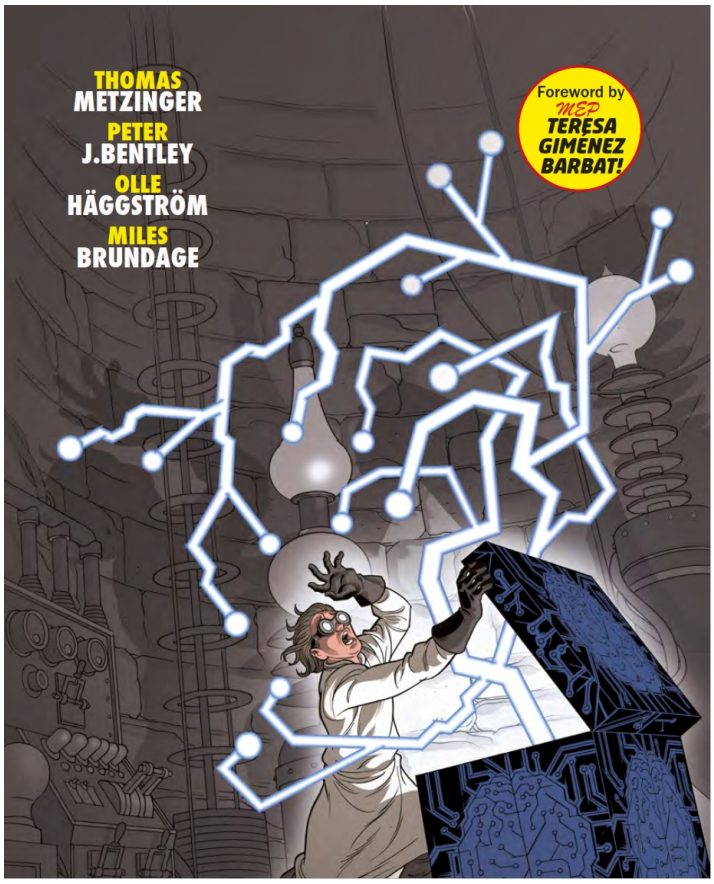
Parlamento europeo

¿Debemos temer a la inteligencia artificial?



ANÁLISIS EN PROFUNDIDAD

EPRS | Servicio de Estudios del Parlamento Europeo
Autor: Philip Boucher
Unidad de Previsión Científica (STOA)
PE 581.948

ES

¿Debemos temer a la inteligencia artificial?

Análisis en profundidad

Marzo de 2018

PE 614.547

AUTORES

Peter J. Bentley, University College London
Miles Brundage, Universidad de Oxford
Olle Häggström, Universidad de Chalmers
Thomas Metzinger, Universidad Johannes Gutenberg de Maguncia

Con prólogo de la diputada al PE María Teresa Giménez Barbart
e introducción de Philip Boucher, Unidad de Prospectiva Científica (STOA)

ADMINISTRADOR RESPONSABLE del STOA

Philip Boucher
Unidad de Prospectiva Científica (STOA)
Dirección de Evaluación de Impacto y Valor Añadido Europeo
Dirección General de Servicios de Estudios Parlamentarios
Parlamento Europeo, Rue Wiertz 60, B-1047 Bruselas
Correo electrónico: STOA@ep.europa.eu

VERSIONES LINGÜÍSTICAS

Original: EN

SOBRE EL EDITOR

Para ponerse en contacto con STOA o suscribirse a su boletín mensual, escriba a: STOA@ep.europa.eu
Este documento está disponible en la siguiente dirección de internet: www.europarl.europa.eu/stoa/

Manuscrito terminado en marzo de 2018

EXENCIÓN DE RESPONSABILIDAD

ISBN 978-92-846-3388-3
doi: 10.2861/61195
QA-01-18-199-ES-N

Índice

1. Prólogo

María Teresa Giménez Barbat, diputada al PE

Desde hace ya algunos años, la inteligencia artificial (IA) ha estado cobrando impulso. Una oleada de programas que sacan el máximo rendimiento a los procesadores de última generación están obteniendo resultandos espectaculares. Una de las aplicaciones más destacadas de la IA es el reconocimiento de voz: si bien los primeros modelos eran extraños y se caracterizaban por defectos constantes, ahora son capaces de responder correctamente a todo tipo de solicitudes de los usuarios en las más diversas situaciones. En el ámbito del reconocimiento de imagen también se están logrando avances notables, con programas capaces de reconocer figuras —e incluso gatos— en vídeos en línea que ahora se están adaptando para que el software controle los coches autónomos que invadirán nuestras calles en los próximos años. A día de hoy no podemos imaginar un futuro en Europa sin una IA avanzada que influya cada vez en más facetas de nuestra vida, desde el trabajo a la medicina, y desde la educación a las relaciones interpersonales. En febrero de 2017, el Parlamento Europeo aprobó un informe con recomendaciones destinadas a la Comisión sobre normas de Derecho civil sobre robótica. Muchos diputados escucharon una serie de expresiones curiosas, posiblemente por primera vez: conceptos como «robot autónomo inteligente» e incluso «personalidad electrónica».

Toda futura legislación en este ámbito que pretenda ser verdaderamente útil, favoreciendo el progreso y beneficiando al mayor número posible de ciudadanos, tiene que basarse en un diálogo con expertos. Esta preocupación era el núcleo de mi solicitud al grupo de Evaluación de las Opciones Científicas y Tecnológicas (STOA) para que organizase un acto en el que debatir si podemos ser optimistas respecto a la IA: ¿podemos fiarnos de que beneficiará a la sociedad? Conseguimos reunir a un grupo dirigido por el profesor de Psicología de Harvard y autor científico Steven Pinker. Le acompañaron Peter John Bentley, científico computacional de la University College London; Miles Brundage, del Future of Humanity Institute de la Universidad de Oxford; Olle Häggström, profesor de estadística matemática en la Universidad de Chalmers y autor del libro *Here be dragons*; y el filósofo Thomas Metzinger, de la Universidad de Maguncia y defensor de un código deontológico sobre la IA. Después del acto, Bentley, Brundage, Häggstrom y Metzinger nos enviaron textos que sirvieron de base para la siguiente recopilación.

Lo que tiene el lector en sus manos es una recopilación de documentos que tratan algunas de las ideas que considero especialmente útiles para políticos y legisladores. Por ejemplo, es esencial no caer en la tentación de legislar sobre problemas inexistentes. El camino hacia una sociedad más automatizada, en la que la inteligencia humana no sea la única compleja, no está exento de daños y temores. Nuestra inclinación pesimista ancestral nos hace ver las cosas más negras de lo que realmente son y oponernos por sistema al progreso tecnológico, y también nos da la capacidad de engendrar temores exorbitados como la idea de que una «superinteligencia» se volverá inevitablemente contra la humanidad y desencadenará un futuro «poshumano». Según Peter Bentley, autor del texto *Las tres leyes de la inteligencia artificial*, este mito de que la IA puede constituir una amenaza existencial para la humanidad es uno de los más extendidos y es la causa de numerosos malentendidos. La IA consta de algoritmos matemáticos limitados a la búsqueda de patrones; la creencia de que puede llevar a los robots a querer dominar el mundo no se basa en la realidad, sino que es pura ciencia ficción.

Otra idea destacable es que la IA impulsará y desarrollará una sociedad del bienestar. «Existen un sinfín de posibles usos maliciosos de la IA», explica Miles Brundage, pero si convergen una serie de condiciones descritas en su artículo *Agrandar la humanidad: la defensa del optimismo condicional sobre la inteligencia artificial*, podemos ser muy optimistas. La IA permitirá solucionar problemas complejos y se le atribuirá la responsabilidad de determinadas decisiones, evitando así prejuicios o abusos. Tendrá una importancia económica espectacular en los próximos años. Olle Häggström cita un estudio de McKinsey & Co, según el cual el valor económico añadido resultante de la IA puede cifrase de manera prudente en 30 000 millones de dólares. Thomas Metzinger define algunas de las dificultades más importantes

que ve en el futuro de la IA y propone un conjunto de recomendaciones prácticas sobre cómo podría responder la Unión. Sin duda, tendremos que coexistir con distintos grados de IA. Esperemos que entre todos podamos superar la mayoría de nuestros miedos y entender mejor una tecnología que ya está dando forma a nuestro futuro.

2. Introducción

Philip Boucher

Los seres humanos, en conjunto, están viviendo más años y con mayor salud que nunca. Para muchos, estas medidas básicas bastan para concluir que el mundo se está convirtiendo en un lugar mejor. Sin embargo, cuando vemos los titulares, está claro que sigue existiendo un gran sufrimiento humano. De hecho, si tenemos en cuenta las crecientes amenazas del cambio climático, el aumento del nivel del mar y la extinción masiva, así como las amenazas nucleares y la inestabilidad política, algunos encontrarían pocas razones para la alegría. Dependiendo de las variables que prioricemos (igualdad, biodiversidad, violencia, pobreza, niveles de CO_2, conflicto, agotamiento de la capa de ozono) y cómo las midamos, podemos esgrimir argumentos racionales para una visión optimista o pesimista del futuro de la humanidad.

El panorama también es variopinto cuando consideramos las nuevas tecnologías, como la inteligencia artificial (IA), que se prevé que tendrán una enorme repercusión en el futuro de la humanidad, para bien o para mal. Por ejemplo, la IA podría aportar beneficios sustanciales a varios aspectos de nuestra vida, desde las previsiones meteorológicas al diagnóstico del cáncer. Al mismo tiempo, se ha planteado la preocupación de que pueda amenazar a numerosos puestos de trabajo y asumir importantes procesos de toma de decisiones sin transparencia.

Personalidades muy conocidas se han sumado a ambos bandos del debate. Por ejemplo, Elon Musk compartía la preocupación de que la IA plantee una amenaza existencial para la raza humana, mientras que Bill Gates replicó que la tecnología nos hará más productivos y creativos. Sin embargo, más allá de los titulares, tanto Gates como Musk reconocen que la IA presenta una amplia gama de oportunidades y desafíos y ambos piden que se reflexione sobre cómo podemos gestionar su desarrollo de una manera que maximice sus beneficios sin exponernos a peligro.

Nuestras esperanzas y temores respecto a la IA no se refieren solo a futuros lejanos. A menudo se refieren a la IA actual, que ya tiene una influencia sustancial en nuestras vidas y aparentemente tanto para bien como para mal. Por ejemplo, la IA es parte tanto del problema como de la solución de las noticias falsas. Se han utilizado algoritmos de IA para apoyar una justicia penal más imparcial, pero se les acusa de sesgo racial.

Aunque nadie puede predecir cómo se desarrollará la IA en el futuro, parece que nos encontraremos numerosas dificultades y oportunidades, algunas más serias que otras. Si hubiese una única posición racional sobre el futuro de la IA, ciertamente sería más matizada que el optimismo desenfrenado o el miedo paralizador. Hasta que no sepamos más sobre los efectos de la IA y las capacidades de la humanidad para responder a ellos es importante crear espacios en los que podamos observar, reflexionar y debatir las cuestiones y, en caso necesario, preparar respuestas adecuadas. Este debate debe mantenerse abierto a una amplia variedad de disciplinas. La comunidad de científicos e ingenieros tienen un importante papel que desempeñar, especialmente a la hora de estudiar los límites de lo que es técnicamente posible. Por otro lado, la comprensión del desarrollo y la repercusión de la tecnología en la sociedad exige conocimientos científicos sociales. Ninguna disciplina tiene el monopolio de la sabiduría.

Es en este contexto en el que STOA organizó el 19 de octubre de 2017 un taller en el Parlamento Europeo para examinar si es racional ser optimista respecto a la IA. Steven Pinker (Universidad de Harvard) inauguró el acto con una ponencia sobre el amplio concepto del optimismo racional. A continuación hablaron cuatro ponentes de distintas disciplinas —Peter J. Bentley, científico computacional del University College London, Miles Brundage, investigador de política tecnológica de la Universidad de Oxford, Olle Häggström, estadístico de la Universidad de Chalmers, y Thomas Metzinger, filósofo de la Universidad Johannes Gutenberg de Maguncia— que presentaron sus propias posiciones acerca de si debemos temer a la IA. El animado debate sigue estando disponible en internet y nos complace mucho

que los cuatro ponentes accediesen a precisar sus ideas en documentos de posición individuales que se publican conjuntamente en esta recopilación. Dimos carta blanca a los autores para que expusiesen sus argumentos en sus propios términos y con su propio estilo, con el objetivo de realizar una contribución útil a los debates actuales sobre la IA en la comunidad parlamentaria y fuera de ella. Teniendo en cuenta la creciente atención que acapara el tema entre diputados al PE y ciudadanos por igual, habrá muchos más debates y publicaciones en los próximos años.

3. Las tres leyes de la inteligencia artificial: disipar mitos comunes

Peter J. Bentley

Introducción

La inteligencia artificial (IA) está de moda en la actualidad. Tras algunos éxitos notables en nuevas tecnologías de IA y nuevas aplicaciones, está experimentando un resurgimiento del interés, lo que ha dado lugar a una oleada de opiniones de numerosas disciplinas, entre ellas opiniones de legos, políticos, filósofos, empresarios y grupos de intereses profesionales. Sin embargo, entre estas opiniones raras veces se incluyen las de aquellas personas que mejor entienden la IA: los científicos computacionales y los ingenieros que se pasan los días construyendo las soluciones inteligentes, aplicándolas a nuevos productos y probándolas. Este artículo expone los puntos de vista de un científico computacional con experiencia en la creación de tecnologías de IA en un intento por presentar una opinión equilibrada e informada sobre el tema.

Derribar mitos

Una de las afirmaciones más extraordinarias que se repite a menudo es la de que la IA es, de algún modo, un peligro para la humanidad, incluso una «amenaza existencial». Algunos afirman que de un modo u otro podría desarrollarse de manera espontánea y feroz una IA como un cáncer exponencialmente inteligente. Podríamos empezar con algo simple, pero la inteligencia mejora escapando a nuestro control. Antes de que nos demos cuenta, toda la raza humana está luchando por su supervivencia (Barrat, 2015).

Todo eso suena absolutamente aterrador (que es por lo que muchas películas de ciencia ficción lo utilizan como tema). Sin embargo, a pesar de que haya analistas, filósofos y otras personas que hablan en serio de estas ideas con cierta inconsciencia, no son más que fantasías. La verdad es lo contrario: la IA —como toda la inteligencia— solo puede desarrollarse lentamente, en circunstancias arduas y dolorosas. No es fácil hacerse inteligente.

Siempre ha habido dos tipos de IA: realidad y ficción. La IA real es lo que nos rodea: los reconocedores de voz Siri o Echo, los sistemas de detección de fraudes ocultos de nuestros bancos, incluso los sistemas de lectura de matrículas que utiliza la policía (Aron, 2011; Siegel, 2013; Anagnostopoulos, 2014). La realidad de la IA es que construimos cientos de tipos distintos de software inteligente muy especializados para solucionar un millón de problemas diferentes en distintos productos. Esto lleva sucediendo desde el nacimiento del campo de la IA, que es contemporáneo al nacimiento de los ordenadores (Bentley, 2012). Las tecnologías de IA ya están integradas en software y hardware a nuestro alrededor, pero estas tecnologías son simplemente inteligentes. Son los equivalentes computacionales a los engranajes y muelles en los dispositivos mecánicos. Y, al igual que un engranaje roto o un muelle suelto, si fallan, podría fallar ese producto particular. Del mismo modo que un engranaje o muelle no puede convertirse por arte de magia en un robot asesino, nuestro software inteligente integrado en sus productos no puede convertirse en una IA maliciosa.

La IA real salva vidas ayudando a activar mecanismos de seguridad (frenado automático en coches o incluso vehículos sin conductor). La IA real nos ayuda a optimizar procesos o predecir fallos, mejorando la eficiencia y reduciendo los residuos medioambientales. La única razón por la que existen cientos de empresas de IA, y por la que miles de investigadores e ingenieros estudian este ámbito, es porque aspiran a producir soluciones que ayuden a las personas y mejoren nuestra vida (Richardson, 2017).

El otro tipo de IA —que consta de esas IA generales superinteligentes que nos matarán a todos— es ficción. Los científicos investigadores tienden a trabajar en el primer tipo de IA. Sin embargo, puesto que este artículo tiene que presentar un equilibrio a favor del sentido común racional, las siguientes secciones derribarán varios mitos en este ámbito. En este artículo presentaré «Tres leyes de la IA» para

explicar por qué los mitos son fantásticos, si no absurdos. Estas «leyes» son simplemente un resumen de los resultados de muchas décadas de investigación científica en IA, simplificada para legos.

Primer mito: una IA que se modifique a sí misma se hará superinteligente.

Algunos analistas creen que existe un cierto peligro en que una IA «se suelte» y «se haga superinteligente» (Häggström, 2016).

La primera ley de la IA nos dice por qué no va a ocurrir.

Primera ley de la IA: la dificultad engendra inteligencia.

A partir de nuestra investigación en el ámbito de la vida artificial observamos que la inteligencia solo existe para superar dificultades urgentes. Sin el tipo adecuado de problemas que solucionar, la inteligencia no puede surgir o aumentar (Taylor et al., 2014). La inteligencia solo es necesaria cuando estas dificultades pueden ser variadas e imprevisibles. La inteligencia solo se desarrollará para solucionar estas dificultades si su futuro depende de su éxito.

Para construir una IA simple, creamos un algoritmo para solucionar una dificultad específica. Para cultivar su inteligencia hasta transformarla en una IA general, debemos presentar a nuestra IA en desarrollo dificultades cada vez más complejas y variadas y desarrollar nuevos algoritmos para solucionarlas, manteniendo aquellos que tienen éxito. Sin nuevas dificultades constantes que solucionar, y sin una recompensa por el éxito, nuestras IA no adquirirán otro punto de IQ.

Los investigadores de IA saben todo esto también. Un robot que puede ejecutar bien una tarea nunca desarrollará sus capacidades sin que le obliguemos a ello (Vargas et al., 2014). Por ejemplo, el sistema de reconocimiento automático de matrículas que utiliza la policía es una forma especializada de IA diseñada para solucionar una dificultad concreta: leer matrículas de coches. Aunque se añadiese a esta IA simple algún proceso que le permitiese modificarse, nunca aumentaría su inteligencia sin que se le plantee una dificultad nueva y compleja. Sin una necesidad urgente, la inteligencia es simplemente una pérdida de tiempo y esfuerzo. Si se observa el mundo natural, queda ilustrado en la abundancia; la mayoría de las dificultades en la naturaleza no exigen un cerebro para solucionarlas. Solo muy pocos organismos han tenido que realizar los extraordinarios esfuerzos necesarios para desarrollar un cerebro e incluso menos desarrollan cerebros muy complejos.

La primera ley de la IA nos dice que la inteligencia artificial es un objetivo tremendamente difícil, que requiere las condiciones exactamente adecuadas y un esfuerzo considerable. No habrá inteligencias artificiales fugitivas ni tampoco inteligencias artificiales con un desarrollo propio que escapen a nuestro control. No habrá singularidades. La IA solo alcanzará el grado de inteligencia que nosotros le animemos (u obliguemos) a alcanzar, bajo fuerte presión.

Además, aunque pudiésemos crear una superinteligencia, no hay pruebas de que una IA tan superinteligente quisiese hacernos daño. Estas afirmaciones son profundamente erróneas, quizá derivadas de las observaciones del comportamiento humano, que en efecto es muy violento. Sin embargo, las IA no tienen inteligencia humana. Nuestro verdadero futuro casi sin duda será una continuación de la situación actual: las IA evolucionarán conjuntamente con nosotros y estarán diseñadas para adecuarse a nuestras necesidades, de la misma forma que hemos manipulado los cultivos, el ganado y las mascotas para adecuarse a nuestras necesidades (Thrall et al., 2010). Nuestros gatos y perros no están planeando matar a todos los humanos. Del mismo modo, una IA más avanzada se amoldará a nosotros de manera tan cercana que se integrará dentro de nosotros y de nuestras sociedades. Ya no desearía matarnos más de lo que desearía matarse a sí misma.

Segundo mito: con recursos suficientes (neuronas/ordenadores/memoria), una IA será más inteligente que los humanos.

Los analistas afirman que «más es mejor». Si un cerebro humano tiene cien mil millones de neuronas, una IA con un billón de neuronas simuladas será más inteligente que un ser humano. Si un cerebro humano es equivalente a todos los ordenadores de internet, una IA suelta en internet tendrá inteligencia humana. En realidad, no es el número lo que importa, sino cómo se organizan estos recursos, como explica la segunda ley de la IA.

Segunda ley de la IA: la inteligencia requiere una estructura adecuada.

No existe un tamaño universal para las estructuras cerebrales. Cada tipo de dificultad exige un nuevo diseño para superarla. Para entender lo que vemos, necesitamos un tipo específico de estructural neuronal. Para mover los músculos necesitamos otro tipo. Para almacenar recuerdos necesitamos otro. La biología nos demuestra que no se necesitan muchas neuronas para ser asombrosamente inteligente. El truco está en organizarlas de la manera correcta, construyendo el algoritmo óptimo para cada problema (Garner y Mayford, 2012).

¿Por qué no podemos utilizar las matemáticas para crear inteligencias artificiales?

Utilizamos muchos cálculos inteligentes y, debido a ello, algunos métodos de aprendizaje automático producen resultados predecibles, permitiéndonos entender exactamente lo que pueden hacer y no estas IA. Sin embargo, la mayoría de soluciones prácticas son impredecibles, porque son muy complejas y pueden utilizar la aleatoriedad dentro de sus algoritmos, lo que significa que nuestras matemáticas no pueden seguirlos, y porque a menudo reciben datos impredecibles. Aunque no tenemos matemáticas para predecir las capacidades de una nueva IA, sí tenemos matemáticas que nos hablan sobre los límites de computación. Alan Turing ayudó a inventar la ciencia computacional teórica hablándonos sobre un tipo de límite: nunca podemos predecir si un algoritmo arbitrario (incluida una IA) se detendrá o no en sus cálculos (Turing, 1937). También tenemos el teorema «No Free Lunch», que nos dice que no hay ningún algoritmo que supere a todos los demás en todos los problemas, lo que significa que necesitamos un nuevo algoritmo de IA adaptado a cada nuevo problema si queremos la inteligencia más eficaz (Wolpert, 1996; Wolpert y Macready, 1997). Incluso tenemos el teorema de Rice, que nos dice que es imposible que un algoritmo depure otro algoritmo perfectamente, lo que implica que, aunque una IA pueda modificarse a sí misma, nunca podrá decir si la modificación funciona para todos los casos sin pruebas empíricas (Rice, 1953).

Para construir una IA, tenemos que diseñar nuevas estructuras/algoritmos especializados para cada dificultad a la que se enfrenta. Diferentes tipos de problemas requieren diferentes estructuras. Un problema nunca antes afrontado puede exigir el desarrollo de una nueva estructura que nunca se haya creado antes. No existe una estructura universal que se adecue a todos los problemas: el teorema «No Free Lunch» nos lo dice (Wolpert, 1996; Wolpert y Macready, 1997) (véase el recuadro). Por lo tanto, la creación de una inteligencia cada vez mayor, o la capacidad para gestionar dificultades cada vez más diferentes, es un proceso de innovación continua, con la invención de nuevas estructuras necesarias que se adaptan a cada nueva dificultad. Un gran problema en la investigación de la IA es descubrir qué estructuras o algoritmos resuelven qué dificultades. La investigación todavía se encuentra en una fase incipiente en este ámbito, que es por lo que, en la actualidad, todas las IA son extremadamente limitadas en inteligencia.

A medida que hacemos más inteligentes nuestras IA (o si alguna vez conseguimos descubrir cómo hacer inteligencias artificiales que puedan seguir alterándose a sí mismas), nos encontramos con más problemas. No podemos diseñar la inteligencia de una sola vez, porque no tenemos matemáticas para predecir las capacidades de una nueva estructura y porque no entendemos lo suficiente cómo se relacionan diferentes estructuras/algoritmos con las dificultades. Nuestra única opción para diseñar una mayor inteligencia es un planteamiento de ensayo incremental.

Para cada nueva estructura, tenemos que incorporarla en la inteligencia sin alterar las estructuras existentes. Se trata de algo extremadamente difícil de lograr y puede dar lugar a capas y capas de nuevas estructuras, cada una de las cuales trabaja cuidadosamente con las estructuras anteriores, como es visible en el cerebro humano. Si queremos un cerebro cada vez más inteligente como el nuestro, también podemos añadir la capacidad de readaptación de algunas estructuras si otras están dañadas, cambiando sus estructuras hasta que puedan asumir al menos parcialmente las funciones perdidas. Tampoco tenemos mucha idea de cómo lograrlo.

La segunda ley de la IA nos dice que los recursos no son suficientes. Seguimos teniendo que diseñar nuevos algoritmos y estructuras dentro de las IA y como apoyo para cada nueva dificultad a la que se enfrenten.

Estas son las razones por las que no podemos crear inteligencias para fines generales utilizando un único enfoque. No existe ni una sola IA en el planeta (ni siquiera el «aprendizaje profundo» tan de moda) que pueda utilizar el mismo método para procesar voz, conducir un coche, aprender cómo jugar a un videojuego complejo, controlar un robot para que corra a lo largo de una calle de una ciudad concurrida, lavar platos en un fregadero y planificar una estrategia de generación de inversión para una empresa. Cuando un cerebro humano realiza estas tareas, utiliza un sinfín de distintas estructuras neuronales en distintas combinaciones, cada una diseñada para solucionar un subproblema diferente. No tenemos la capacidad para fabricar estos cerebros, así que, en su lugar, construimos una solución inteligente especializada para cada problema y las utilizamos de manera aislada entre sí.

Tercer mito: puesto que la velocidad de los ordenadores se duplica cada dieciocho meses, las IA explotarán este poder de computación y se volverán exponencialmente más inteligentes.

Los analistas afirman que la velocidad de cálculo pura y dura superará todas las dificultades en la creación de la IA. Si las IA utilizan ordenadores que sean suficientemente rápidos, podrán aprender y superarnos en pensamiento. Puesto que la velocidad de los procesadores informáticos ha estado duplicándose aproximadamente cada dieciocho meses durante décadas, es seguro algo inevitable. Lamentablemente, este punto de vista no reconoce el impacto de un exponencial opuesto que funciona como freno considerable al desarrollo de las IA: los ensayos.

Tercera ley de la IA: la inteligencia requiere un ensayo exhaustivo.

La inteligencia superior requiere los diseños más complejos del universo, pero cada pequeñísimo cambio realizado para intentar mejorar el diseño de una inteligencia puede destruir alguna o todas sus capacidades existentes. No ayuda que no tengamos matemáticas capaces de predecir las capacidades de una inteligencia general (véase el recuadro). Por estos motivos, cada nuevo diseño de inteligencia necesita un ensayo completo en todos los problemas para cuya solución se ha concebido. El ensayo parcial no basta. La inteligencia debe probarse en todas las posibles permutaciones del problema durante su vida útil diseñada; de lo contrario, sus capacidades pueden no ser fiables.

Todos los investigadores de IA conocen demasiado bien esta dura verdad: para construir una IA, es necesario entrenarla y probar todas sus capacidades de manera exhaustiva en su entorno previsto en todas las fases de su diseño. Como dijo Marvin Minsky, fundador del campo de la IA, «…hay tantas historias de cómo podrían salir mal las cosas, pero no veo manera de tomarlas en serio, porque es bastante difícil entender por qué alguien las instalaría a gran escala sin muchos ensayos» (Achenbach, 2016). Más que cualquier otro aspecto, es el proceso de ensayo lo que requiere la mayor parte del tiempo. Esta limitación de tiempo supone un freno en el proceso de diseño de inteligencia. En el peor de los casos, el nivel de ensayo es exponencial por cada ganancia gradual de inteligencia.

Para entender por qué, imagine una inteligencia que pueda reconocer diez colores diferentes y tenga que distinguir dos tipos de objeto utilizando esta única característica, el color. En este caso, la AI puede entender como máximo diez tipos distintos de elementos y clasificarlos en dos clases. Si se amplían sus capacidades para manejar dos características —pongamos color y diez tonalidades— puede entender

como máximo cien tipos distintos de elementos. Si pudiese manejar cien características, podría entender 10 000 000 000 000 000 000 000 000 000 000 000 000 000 000 000 000 000 000 000 000 000 000 000 000 000 000 000 000 000 000 000 000 000 000 elementos. ¿Cómo sabemos que la IA los clasificará todos de manera correcta? Esto se convierte en una cuestión muy importante para las aplicaciones esenciales para la seguridad. Imaginen que la IA está conduciendo su coche y las dos clases de objeto son «carretera» y «obstáculo».

En la realidad es mucho más complejo. Muchas inteligencias biológicas pueden distinguir un número inimaginable de elementos utilizando miles o millones de características, cada una de las cuales podría adoptar cientos o millones de valores. El ensayo de estas inteligencias en efecto lleva mucho tiempo. No debería sorprender mucho. La evolución crea sus cerebros biológicos ensayando incontables billones de cerebros en paralelo en todas las hipótesis posibles durante miles de millones de años y solo permite que los cerebros de éxito contribuyan a los futuros diseños. Se trata de un método sumamente eficiente y eficaz de ensayo con el que nosotros solo podemos soñar.

La tremenda necesidad de probar las IA tiene implicaciones significativas. No podemos diseñar mejores IA sin probarlas en cada fase. Google lleva años probando sus coches sin conductor y sigue haciéndolo. (En junio de 2016, Google había probado su flota de vehículos en modo autónomo en un total de 2 777 585 km) (Google, 2016). Todos los ingenieros e investigadores de IA saben que no podemos utilizar la inteligencia artificial en una aplicación esencial para la seguridad hasta que se hayan realizado ensayos adecuados. En el futuro próximo también necesitaremos certificación con el fin de saber exactamente lo bien que funciona una IA para tareas bien definidas. Y si la IA sigue aprendiendo o se actualiza, no podemos suponer que, puesto que ha pasado antes una prueba, seguirá haciéndolo; al igual que un piloto humano, las IA que sigan aprendiendo deben volver a someterse a ensayos continuamente para garantizar que sigan estando certificadas[1].

La tercera ley de la IA nos dice que, a medida que aumenta la inteligencia, el tiempo necesario para el ensayo puede incrementarse exponencialmente. En última instancia, el ensayo puede imponer límites prácticos a una inteligencia artificial viable y fiable. De la misma forma que cada vez es más difícil ir más rápido a medida que nos acercamos a la velocidad de la luz, cada vez es más difícil aumentar la inteligencia mientras construimos cerebros más inteligentes.

Una vez más, esta es una razón fundamental por la que la investigación y la aplicación de la IA se dedican a encontrar soluciones inteligentes a problemas muy concretos[2].

Conclusiones

La IA existe desde el nacimiento de los ordenadores. Lleva el tiempo suficiente para que haya experimentado períodos de entusiasmo en los que algunos destacados expertos realizan afirmaciones extraordinarias (Bentley, 2012). Claude Shannon fue uno de los mayores pioneros de la informática y la IA. En 1961 dijo: «Confío en que, en cuestión de diez o quince años, surja algo del laboratorio que no

[1] Además, cualquier interacción humana con inteligencias artificiales implica también formación y nueva certificación para nosotros. Un automóvil con una IA que asuma el control en algunas circunstancias se convierte en una responsabilidad cuando la IA llega al límite de sus capacidades y el conductor no ha sido formado para mantenerse lo suficientemente alerta como para recuperar el control (Eriksson y Stanton, 2017).

[2] Otra razón fundamental es nuestro propio cerebro: ahora mismo, y en un futuro previsible, no somos lo suficientemente inteligentes para crear inteligencia. No entendemos cómo funciona el cerebro biológico. No sabemos por qué funcionan algunas de nuestros mejores métodos de IA. No sabemos cómo mejorarlos. El efecto de freno en el progreso, causado por nuestra propia ignorancia, es considerable.

esté demasiado alejado del robot de ciencia ficción».[3]Predijo que para mediados de la década de 1970 tendríamos máquinas autónomas que caminarían, hablarían y pensarían. Cuarenta años después, apenas podemos hacer andar a un robot. Ciertamente, no puede pensar por sí mismo. En la actualidad, hay estudios (que contienen opiniones muy variadas) que concluyen que existe un 50 % de posibilidades de que la IA supere a los humanos en todas las tareas en cuarenta y cinco años (Grace et al., 2017). Todo suena muy familiar. Y será igual de impreciso.

No se crean el bombo publicitario. Somos terribles prediciendo el futuro y, casi sin excepción, las previsiones (incluso las realizadas por expertos mundiales) son completamente erróneas. Al final, la historia nos dice que el bombo publicitario es la razón por la que la investigación de la IA se hunde en los períodos de recesión (Bentley, 2012). Las grandes afirmaciones llevan a una gran publicidad, que conduce a una gran inversión y nuevas normativas. Y entonces es cuando sacude la realidad inevitable. La IA no está a la altura del bombo publicitario. La inversión se seca. La normativa ahoga la innovación. Y la IA se convierte en una expresión sucia que nadie osa pronunciar. Otro ocaso de la IA destruye el progreso.

Las historias alarmistas y las previsiones tontas no tienen cabida en el progreso científico o la elaboración de políticas; déjenlas para los cines. Sin embargo, la calma y el debate racional son muy importantes. Ahora se está utilizando tecnología de IA para nuevas aplicaciones esenciales para la seguridad. La distracción que causa el alarmismo podría provocar una pérdida de vidas. En lugar de centrarse en lo que podría pasar si se hiciese realidad una historia de ciencia ficción, deberíamos centrarnos en nuevas normativas de seguridad y certificaciones para cada aplicación de IA esencial para la seguridad específica. ¿Dónde están los nuevos ensayos de seguridad de las carreteras y la certificación para coches sin conductor? ¿Dónde están los nuevos exámenes de conducción para conductores humanos que tienen coches sin conductor? ¿Dónde están los nuevos indicadores homologados de vehículos que informan a los peatones de que el coche los ha visto y pueden cruzar de forma segura la carretera? ¿Dónde están las normativas que impiden que los servicios de noticias segmentadas sobre la IA creen puntos de vista cada vez más polarizados en las poblaciones? (Cesa-Bianchi et al., 2017). Ha llegado el momento de dejar de lado los sinsentidos y centrarse en la realidad, aquí y ahora. ¿Cómo hacemos segura a día de hoy cada nueva aplicación específica de software inteligente?

La inteligencia artificial tiene un potencial asombroso para mejorar nuestra vida, ayudándonos a vivir con más salud y más felices y generando un mayor número de nuevos empleos. La creación de IA engloba muchas de las mayores hazañas científicas y de ingeniería que hemos emprendido. Es una nueva revolución tecnológica. Pero esta revolución no ocurrirá por sí sola por arte de magia. Las tres leyes de la IA nos dicen que, si queremos hacer inteligencias artificiales más avanzadas, debemos dar poco a poco más dificultades a nuestras IA, diseñar cuidadosamente nuevas estructuras inteligentes para que puedan superar estas dificultades y realizar ensayos masivos para confirmar que se puede confiar en ellas para resolver las dificultades. Miles de científicos e ingenieros cualificados están siguiendo exactamente estos pasos de manera incansable (problema, solución hipotética, ensayo) para traernos cada pequeñísima mejora, porque este es nuestro proceso de diseño y nuestro método científico. No teman a la IA, maravíllense con la persistencia y la habilidad de estos especialistas humanos que están dedicando su vida para ayudar a crearla. Y valoren que la IA está ayudando a mejorar nuestra vida cada día.

[3] Extracto de una entrevista con Claude Shannon, que apareció en el programa de televisión The Thinking Machine, de la serie documental «Tomorrow», 1961. Copyright CBS News.

Referencias

Achenbach, J. (2016). Professor Marvin Minsky: Mathematician and inventor inspired by Alan Turing to become a pioneer in the field of artificial intelligence. Obituaries, Independent. Friday 29 January 2016.

Anagnostopoulos, C-N E., (2014) License Plate Recognition: A Brief Tutorial. IEEE Intelligent Transportation Systems Magazine. Volume: 6, Issue: 1, pp. 59 – 67. DOI: 10.1109/MITS.2013.2292652

Aron, J. (2011) How innovative is Apple's new voice assistant, Siri? New Scientist Vol 212, Issue 2836, 29 October 2011, p. 24. https://doi.org/10.1016/S0262-4079(11)62647-X

Barrat, J. (2015) Why Stephen Hawking and Bill Gates Are Terrified of Artificial Intelligence. Huffington Post.

Bentley, P. J. (2012) Digitized: The science of computers and how it shapes our world. OUP Oxford. ISBN-13: 978-0199693795.

Cesa-Bianchi, N., Pontil, M., Shawe-Taylor, J., Watkins, C., y Yilmaz, E. (2017) Proceedings of workshop: Prioritise me! Side-effects of online content delivery. The problem of bubbles and echo-chambers: new approaches to content prioritisation for on-line media, 12 de junio de 2017. Knowledge 4 All Foundation. Londres, Reino Unido.

Eriksson, A., y Stanton, N. A. (2017). Driving performance after self-regulated control transitions in highly automated vehicles. Human Factors: The Journal of the Human Factors and Ergonomics Society. DOI: 10.1177/0018720817728774

Garner, A. y Mayford, M. (2012) New approaches to neural circuits in behaviour. Learn. Mem. 2012. 19: 385-390. Doi: 10.1101/lm.025049.111

Google (2016). «Google Self-Driving Car Project Monthly Report - June 2016» (PDF). Google. Recuperado el 15 de julio de 2016. https://static.googleusercontent.com/media/www.google.com/en//selfdrivingcar/files/reports/report-0616.pdf

Grace, K., Salvatier, J. Dafoe, A., Zhang, B. y Evans, O. When Will AI Exceed Human Performance? Evidence from AI Experts. arXiv:1705.08807

Häggström, O. (2016) Here Be Dragons: Science, Technology and the Future of Humanity. OUP Oxford.

Rice, H. G. (1953). Classes of Recursively Enumerable Sets and Their Decision Problems. Trans. Amer. Math. Soc. 74, 358-366.

Richardson, J. (2017) Three Ways Artificial Intelligence is Good for Society. IQ magazine, Intel. Disponible en línea: https://iq.intel.com/artificial-intelligence-is-good-for-society/

Siegel, E. (2013) Predictive Analytics: The Power to Predict who will Click, Buy, Lie or Die. John Wiley & Sons, Inc. ISBN: 978-1-118-35685-2.

Taylor, T., Dorin, A., Korb, K. (2014) Digital Genesis: Computers, Evolution and Artificial Life. Presented at the 7th Munich-Sydney-Tilburg Philosophy of Science Conference: Evolutionary Thinking, University of Sydney, 20-22 de marzo de 2014. arXiv:1512.02100 [cs.NE]

Thrall, P. H., Bever, J. D., y Burdon, J. J. (2010) Evolutionary change in agriculture: the past, present and future. Evol Appl.3(5-6): 405–408. doi: 10.1111/j.1752-4571.2010.00155.x

Turing, A. (1937) On computable numbers, with an application to the Entscheidungsproblem, Proceedings of the London Mathematical Society, Series 2, Volume 42, pp 230–265, doi:10.1112/plms/s2-42.1.230

Vargas, P. A., Di Paolo, E. A., Harvey, I. y Husbands, P. (Eds) (2014) The Horizons of Evolutionary Robotics. MIT Press.

Wolpert, D.H. (1996). The Lack of A Priori Distinctions between Learning Algorithms. Neural Computation, pp. 1341-1390.

Wolpert, D.H., Macready, W.G. (1997). No Free Lunch Theorems for Optimization. IEEE Transactions on Evolutionary Computation 1, 67.

4. Agrandar la humanidad: la defensa del optimismo condicional sobre la inteligencia artificial

Miles Brundage

Introducción

Las opiniones de los expertos sobre el horizonte temporal de los futuros avances en la inteligencia artificial (IA) varían ampliamente, con algunos que esperan una IA de nivel humano en las próximas décadas y otros que piensan que será mucho más adelante (Grace et al., 2017). Del mismo modo, los expertos no se ponen de acuerdo acerca de si es probable que los avances en la IA sean beneficiosos o perjudiciales para la civilización humana, con una gama de opiniones que van desde aquellas que consideran seguro que será extremadamente beneficioso, las que consideran probable que sea extremadamente perjudicial (incluso con riesgo de extinción humana) y muchas entre medias (AI Impacts, 2017). Aunque los riesgos del desarrollo de la IA han atraído recientemente una atención sustancial (Bostrom, 2014; Amodei y Olah et al., 2016), ha habido poco debate sistemático sobre las maneras precisas en que la IA podría ser beneficiosa a largo plazo.

En este artículo no pretendo determinar lo que es probable que ocurra, sino más bien defender el *optimismo condicional* sobre la IA y exponer las razones por las que se podría anticipar que la IA es una tecnología transformadora para la humanidad, posiblemente beneficiosa desde el punto de vista de la transformación. Con esto quiero decir que, si la humanidad lograr sortear las dificultades técnicas, éticas y políticas del desarrollo y la difusión de potentes tecnologías de IA, la IA puede tener un enorme impacto posiblemente muy positivo en el bienestar de la humanidad. Para justificar esta conclusión, primero examino las características de la IA que se prestan a generar un enorme impacto (positivo o negativo) en el bienestar de la humanidad a largo plazo. A continuación describo brevemente algunas condiciones para el éxito, es decir, qué dificultades tendrían que sortearse para traer el futuro positivo que posibilitan las características de la IA. Posteriormente, en la mayor parte del artículo enumero tres razones distintas por las que esperar (condicionalmente) que la IA tenga una enorme impacto positivo en la humanidad: una IA poderosa agilizaría en gran medida la realización de tareas (*celeridad de tareas*), permitiría una coordinación a mayor escala y más eficaz de las personas y las instituciones (*coordinación mejorada*) y permitiría reorientar la vida de los seres humanos hacia la consecución de los objetivos que consideran intrínsecamente gratificantes, manteniendo al mismo tiempo un elevado nivel de vida sin necesidad de trabajo no deseado (*sociedad del ocio*).

Ninguno de estos resultados está garantizado con el desarrollo de la IA, pero parto de la hipótesis de que la IA es necesaria para cumplir cada uno de ellos. En contra de la idea de que la IA es tan arriesgada que debe evitarse su desarrollo, yo argumento en cambio que la IA será un pilar esencial de la prosperidad humana a largo plazo y concluyo con una visión positiva de cómo sería el resultado final.

Características de la IA

La IA es un corpus de investigación e ingeniería que se centra en la utilización de tecnología digital para crear sistemas capaces de realizar tareas (a menudo como resultado de un aprendizaje) que comúnmente se cree que requieren inteligencia cuando las realiza un ser humano o un animal, y que ha progresado muy rápido en los últimos años tras décadas de escasos resultados. Entre los recientes logros notables de la IA están la superación de los resultados humanos en el juego del Go y el logro de resultados sobrehumanos en una serie de tareas de procesamiento de imágenes. Las tecnologías de IA están muy extendidas en la vida moderna, con aplicaciones de uso común como motores de búsqueda, reconocimiento de voz en los teléfonos y traducción automática por internet.

Sin embargo, algo más importante que cualquier otro logro particular de la IA en una tarea específica es que combina las propiedades de las tecnologías digitales en general (incluida la *escalabilidad* mediante la copia de programas y la aceleración de su ejecución) con propiedades que comúnmente se piensa que son únicas de los seres humanos (*competencia*). Es decir, la importancia de la IA reside en gran medida en su capacidad para ampliar el rendimiento de las tareas inteligentes, ya que, por ejemplo, la traducción automática permite que millones de usuarios traduzcan un texto simultáneamente. Además de esta característica de *competencia ampliable*, en principio puede darse casi cualquier objetivo a las IA potentes (Bostrom, 2014), lo que es una fuente de riesgos y oportunidades. Por último, tanto en los ámbitos limitados en la actualidad como en la toma de decisiones inteligentes en general a largo plazo, la IA puede superar el rendimiento humano, abriendo la oportunidad de orientar un gran número de sistemas rápidos y competentes a la consecución de objetivos casi arbitrarios. Son estas propiedades de la IA las que fundamentan el debate de las consecuencias sociales que sigue a continuación.

Condiciones para el éxito

Una tecnología que es flexible y potente tendrá infinidad de consecuencias sociales (como ha tenido la electricidad, por ejemplo). Sin embargo, a diferencia de la electricidad, los sistemas de IA sirven a una variedad mucho más amplia de posibles funciones y servirán a funciones aún más diversas en el futuro. Existen infinidad de posibles usos maliciosos de la IA (Brundage y Avin et al., 2018) y muchas formas en las que podría utilizarse de manera perjudicial involuntariamente, por ejemplo con el sesgo algorítmico (Kirkpatrick, 2016). Para lograr a largo plazo los beneficios descritos a continuación, tendrán que evitarse numerosos resultados negativos. Quizá lo más importante es que se tendrá que abordar el *problema de control*, es decir, tendremos que aprender cómo asegurar que los sistemas de IA cumplan los objetivos que queremos que cumplan (Bostrom, 2014; Amodei y Olah et al., 2016; Bostrom, Dafoe, y Flynn, 2017) sin causar daños durante su proceso de aprendizaje, malinterpretar lo que se requiere de ellos o resistirse al control humano. Aunque los sistemas de IA actuales tienen capacidades limitadas en relación con los seres humanos y es improbable que se materialicen algunas preocupaciones extremas de seguridad (como sistemas de IA que pueden evitar su apagado), la solución del problema de control es un requisito previo esencial a largo plazo para que los sistemas de IA más potentes tengan efectos positivos en la sociedad. Además, tendrán que sortearse con éxito las dificultades políticas de la IA, incluidos los riesgos asociados a la concentración indebida de poder y riqueza (Bostrom, Dafoe y Flynn, 2017) y las arriesgadas carreras de desarrollo que fomentan la falta de atención a la seguridad para ganar ventaja (Armstrong et al., 2016; Bostrom, 2017).

En adelante, para centrar el debate, doy por sentado que se han solucionado todas las dificultades anteriores y profundizo en formas en las que el resultado podría ser extremadamente beneficioso. Como señalé anteriormente, este artículo no está concebido para leerse como predicción, sino como ejercicio para examinar más detalladamente un lado del libro de costes/beneficios. Después de presentar cada una de estas razones para el optimismo, las combinaré en una visión general positiva de un posible futuro con una IA más avanzada.

Razones para el optimismo: celeridad de tareas, coordinación mejorada y sociedad del ocio

Celeridad de tareas

La competencia ampliable de la IA se presta a la ejecución de un gran número de tareas con mayor rapidez de lo que sería posible de otro modo, incluidas tanto tareas que los seres humanos somos capaces de cumplir (contando con suficiente tiempo y recursos) como otras que no somos capaces de cumplir en ningún plazo de tiempo debido a nuestras limitaciones cognitivas y organizativas. Ya ha habido demostraciones de rendimiento humano y sobrehumano de sistemas de IA y una rápida transición del juego del Go del primero al segundo en los últimos años. Existen numerosos usos de

sistemas tecnológicos de nivel humano (o incluso sistemas de rendimiento inferior al humano), como la capacidad de realizar tareas que son tediosas y requieren un gran consumo de tiempo. La traducción automática es un ejemplo ilustrativo de ello, donde cada pequeño aumento del rendimiento mejorado de los sistemas de traducción automática puede desplegarse relativamente rápido a una gran variedad de pares de idiomas y millones de usuarios. Asimismo, el reconocimiento de voz no ha llegado a los niveles de rendimiento humano en todos los contextos, pero a menudo ahorra tiempo a los usuarios de dispositivos digitales que prefieren no teclear cada palabra.

La aplicación de la *celeridad de tareas* característica de la IA a una gama más amplia de ámbitos, incluidos aquellos en los que se requiere un elevado nivel de inteligencia y perspicacia como la ciencia y la ingeniería, podría tener consecuencias más radicales. Teniendo en cuenta la velocidad potencial mucho mayor de los ordenadores en relación con los cerebros humanos (con miles de millones de operaciones por segundo para una determinada unidad computacional en comparación con cientos) y la capacidad de ampliar los sistemas de IA a grandes cantidades de hardware informático, la IA general podría permitir alcanzar rápidamente avances científicos y de ingeniería. Algunos de estos avances son del tipo que los seres humanos serían capaces de lograr a la larga, teniendo tiempo suficiente, pero se verían acelerados por la IA dirigida al problema. Otros podrían no ser alcanzables sin la ayuda de la IA debido a las limitaciones cognitivas humanas (como los límites en la memoria a corto y largo plazo). Los únicos límites claros de lo que podrían lograr las IA más sofisticadas son los límites de la física y estos permiten obtener ordenadores mucho más rápidos, materiales más fuertes y energía más barata, entre otras cosas mediante el desarrollo de fabricación atómicamente precisa (Drexler, 2013). En el ámbito de la investigación biológica, incluso el envejecimiento no es claramente una característica permanente de la condición humana, y un sinfín de otras mejoras físicas y cognitivas parecen físicamente posibles (Kurzweil, 2005).

Coordinación mejorada

Los sistemas de IA más sofisticados, si se aplican adecuadamente, podrían permitir resolver algunos de los conflictos sociales actualmente intratables a través de la coordinación mejorada. Los dilemas del prisionero y otros problemas de acción colectiva, en los que el bienestar general de dos o más partes mejoraría si cooperasen pero cada una de ellas tiene un incentivo para no hacerlo, son generalizados en la sociedad. Estos dilemas se han utilizado históricamente para justificar la creación de gobiernos poderosos e instituciones internacionales para coordinar a los gobiernos. Sin embargo, nuestras herramientas para coordinar son limitadas, en parte porque es difícil supervisar el comportamiento de los seres humanos en busca de señales de deserción de un acuerdo, y en parte porque la confianza interpersonal e intergrupal puede ser difícil de lograr cuando las intenciones humanas están ocultas dentro de mentes opacas. Cada uno de estos obstáculos a la cooperación (supervisión insuficiente y seres humanos poco fiables) puede atenuarse potencialmente mediante la aplicación de la IA para el cumplimiento de los acuerdos. Examino cada uno por separado.

Con respecto a la supervisión insuficiente, en las últimas décadas ha habido una tendencia histórica a una recopilación y análisis más generalizados de datos sobre el comportamiento humano. Cada vez más, los seres humanos desarrollan sus interacciones empresariales y sociales por internet, lo que hace que sea más fácil para las empresas y los gobiernos supervisar sus actividades, para bien y para mal. Asimismo, las cámaras cada vez más extendidas (incluidas cámaras de vigilancia específicas y cámaras integradas en teléfonos inteligentes y otros dispositivos) pueden utilizarse para hacer un seguimiento de las actividades físicas humanas. El abuso de la autoridad de vigilancia de los gobiernos está bien documentado y no debe interpretarse que el análisis aquí presentado rebaja dichos abusos. Sin embargo, existe un posible lado positivo muy importante en la vigilancia a través de los sistemas de IA: pueden utilizarse para supervisar más eficazmente los acuerdos intranacionales e internacionales, haciendo posiblemente más manejable la cooperación en ámbitos como el control de armas, la reparación medioambiental y la ciberdelincuencia. Por ejemplo, la proliferación nuclear es un problema constante, como demuestran los recientes conflictos internacionales por los programas nucleares de Irán y Corea

del Norte. Parte del problema de la ejecución de acuerdos internacionales (incluso los ampliamente beneficiosos como los relacionados con la no proliferación) es que las actividades en internet y fuera de ella, aunque son más detectables que en cualquier otro momento de la historia humana, se supervisan de manera imperfecta, permitiendo actividades ilegales como la venta clandestina de información nuclear. La IA podría ayudar con esta dificultad relativa a acuerdos más eficaces y comunes automatizando el proceso de recopilación y análisis de información obtenida de diversas fuentes de datos, posibilitando la vigilancia a una escala mucho mayor. Para ello, pueden combinarse la IA y la robótica, por ejemplo, para utilizar drones pequeños y baratos a fin de ampliar el alcance de las actividades de vigilancia.

En segundo lugar, la IA puede eliminar algunos aspectos del sesgo y la corrupción humanos de los regímenes de vigilancia y la gobernanza más en general, precisamente porque puede sacar a los seres humanos de determinados procesos decisorios. A diferencia de un ser humano que trabaja en la Agencia Nacional de Seguridad, por ejemplo, que podría verse tentado a abusar de su poder por razones personales, un sistema de IA utilizado para vigilancia puede tener un código auditado para garantizar que ningún ser humano vea nunca datos que no se le permiten ver o que ningún ser humano vea nunca ningún dato de vigilancia. En una evolución aún más extrema, el cifrado homomórfico podría permitir analizar datos cifrados con la garantía de que ni siquiera la propia IA puede ver los datos no cifrados (Trask, 2017). Con estas medidas podrían negociarse y ejecutarse acuerdos de un ámbito más amplio, ayudando posiblemente a eliminar muchas formas de delincuencia y ampliando la magnitud potencial de instituciones políticas eficaces.

Sociedad del ocio

El tercer y último beneficio principal de la IA avanzada de la que hablo es la posibilidad de generar una sociedad del ocio próspera y ética. Abundan las predicciones relativas al momento y la secuencia de empleos que se automatizarán con la IA, la robótica y otras tecnologías (Brundage, 2015; Brynjolfsson y McAfee, 2014; Grace et al., 2017). No tomo aquí ninguna posición sobre cuánto tiempo se tardará en que sea tecnológicamente posible automatizar todos los trabajos humanos, sino que solo afirmo que en principio es posible y probable que ocurra en algún momento en el futuro. Esto se deriva simplemente de la opinión de que la cognición y el comportamiento humanos son procesos físicos que pueden a la larga ser simulados por otros sistemas físicos, en particular los ordenadores digitales y (en casos en que se requiera actividad física para el trabajo) robots. Si se alcanzase este nivel de capacidad técnica, el contrato social de la sociedad tendría que renegociarse de algún modo. Podría desarrollarse de muchas formas diferentes. Quizá se distribuiría un nivel mínimo de ingresos a todos los miembros de la sociedad para proporcionar un nivel de vida básico; quizá los ciudadanos y los gobiernos acordarían que es positivo tener la necesidad de trabajar y que (incluso si es tecnológicamente innecesario) el trabajo remunerado debería continuar de cierta forma, tal vez prohibiendo la automatización de determinados trabajos; y quizá algunos trabajos seguirían desempeñándose en casos en los que el cliente atribuya un valor intrínseco a que las tareas sean realizadas por un ser humano en lugar de una IA. Un posible supuesto, que no defiendo como el adecuado o el más probable sino solo como posiblemente muy valioso, es una sociedad del ocio propiciada por la IA. En este tipo de sociedad, los seres humanos se centran en las actividades que consideran intrínsecamente gratificantes (como crear arte, aprender, jugar a juegos, criar hijos o pasar tiempo con amigos o la pareja) y no tienen obligación de trabajar para mantener un elevado nivel de vida. El nivel de vida mínimo en esta sociedad podría ser mucho mayor que en la actualidad, teniendo en cuenta que la plena automatización generaría un rápido crecimiento económico y que pronto podríamos acercarnos a otros varios límites físicos, como la mejora cognitiva y una producción de energía y bienes mucho más barata.

¿Cuánto mejor podría ser esta sociedad del ocio en relación con las sociedades que tenemos hoy en día o con aquellas que conocemos a lo largo de la historia? El límite parece ser alto: es difícil estimar el grado de prosperidad que podría alcanzar este tipo de sociedad, pero un mínimo razonable para esta estimación es que podría ser al menos tan buena como ha sido cualquier vida humana, dada la ausencia

de límites físicos claros en la capacidad de producir dichos niveles de vida a gran escala cuando puedan automatizarse todas las tareas. En los casos en los que alcanzar esta elevada calidad de vida no se reduce simplemente a producir materiales físicos baratos, como parece probable, también podrían impulsarse inteligencias artificiales de realidad virtual inmersiva (físicas o virtuales) y socialmente interactivas para ofrecer una gama de experiencias casi ilimitada. La simple reproducción de los niveles de vida actuales de forma física o virtual y a gran escala claramente no llega a los límites de prosperidad potencial, pero este análisis ilustra lo mínimo que debemos esperar que sea posible a la larga.

Cabe señalar una última consideración con respecto al logro de una sociedad del ocio ética: el bienestar de los propios sistemas de IA, si es que tal concepto es aplicable a ellos. Esta preocupación merece ser estudiada seriamente y esperemos que el futuro progreso en el entendimiento de la inteligencia y la conciencia nos ayude a entender mejor el panorama de posibles mentes. Una perspectiva ética imperiosa es que el sustrato (es decir, el cerebro o chip de ordenador) *per se* no debe utilizarse como base para la discriminación entre seres humanos e inteligencias artificiales (Bostrom y Yudkowsky, 2011), aunque en última instancia podríamos aprender que los sustratos son relevantes para el tipo de conciencia que pueden admitir. Sin embargo, pueden evitarse algunos dilemas éticos mediante el diseño reflexivo y responsable de sistemas por defecto y podríamos esforzarnos por diseñar sistemas de manera que no puedan sufrir, incluso si dichos sistemas admiten una experiencia consciente (Bryson, 2016). A muy largo plazo, estas cuestiones tendrán que resolverse, pero una cosa está clara: una sociedad del ocio propiciada por la IA al menos parece brindar la *posibilidad* de alcanzar un ocio y prosperidad generalizados de manera ética, teniendo en cuenta lo que sabemos a día de hoy. En contraposición, otras vías hacia las sociedades del ocio (como las logradas históricamente a costa de la esclavitud humana) claramente carecen de ética y, sin las capacidades tecnológicas asociadas a la IA avanzada, lo mejor que podríamos esperar sería un nivel de vida más bajo en una sociedad del ocio alcanzada por medios políticos. Téngase en cuenta que es concebible que los propios sistemas diseñados puedan alcanzar grados y volúmenes de bienestar aún mayores, en relación con los seres humanos que sacan partido a los sistemas diseñados, si sus sustratos resultan admitir estas experiencias conscientes, pero el universo es suficientemente extenso como para que esto no sea (al menos en principio) incompatible con que los humanos también alcancen un elevado nivel de vida.

Conclusión: IA escalable para ampliar la prosperidad y la civilización humana

A la larga, es probable que se inventen sistemas de IA capaces de realizar cualquier tarea que pueden realizar los seres humanos (y muchas más). No sabemos cuánto tiempo llevará, pero los expertos están en gran parte de acuerdo en que es posible, y muchos creen que es probable que ocurra en este siglo. ¿Qué podemos y no podemos decir sobre un mundo con estos sistemas?

No podemos decir con seguridad que los seres humanos sobrevivan para disfrutarlo. De hecho, incluso sin una IA más avanzada, los seres humanos hemos tenido (al menos desde el desarrollo de las armas nucleares) la capacidad de autodestruirnos, y hay argumentos convincentes de que la IA podría ser otro tipo de tecnología peligrosa (Bostrom, 2014). Sin embargo, no está claro tampoco que no sobrevivamos para disfrutarla. No hay una contradicción inherente en la existencia de un sistema artificial muy inteligente que se esfuerce por mejorar el bienestar humano sin resistirse o resentirse por esta posición servil, y muchos investigadores están trabajando activamente en garantizar que estos sean el tipo de sistemas que construyamos finalmente. Tampoco podemos decir todavía que, si sobrevivimos para ver este mundo, será positivo para los humanos. Podría abusarse de esta tecnología para crear un estado autoritario estable con una resistencia sin precedentes y a escala mundial, basado en la automatización de la vigilancia, la coacción y el aplastamiento de la disidencia. Y entre la utopía y la distopía son posibles muchos más supuestos.

Sin embargo, podemos decir algunas cosas sobre el tipo de sociedades que la humanidad *podría* lograr si consigue realizar esta transición. Los tres factores examinados anteriormente —celeridad de tareas, coordinación mejorada y sociedad del ocio— son importantes en sí mismos a nivel individual, y a nivel

colectivo se combinan para esbozar un camino a una civilización extensa, próspera y espacial. En un mundo en el que cualquier tarea puede acelerarse con la ayuda de la IA, una tarea muy beneficiosa que agilizar sería el desarrollo y el despliegue de tecnologías para una rápida colonización espacial. Esto daría acceso a enormes cantidades de tierra, recursos materiales y emocionantes oportunidades de exploración para la humanidad. La combinación de la apertura de estas nuevas fronteras con la aceleración de otras tareas, como el desarrollo de novedosas técnicas de mejora cognitiva y bienes y servicios mucho más baratos, podrían permitir un nuevo Renacimiento en cuestiones humanas. Aunque la IA podría convertirse en una nueva generación de armas utilizadas por los Estados y las personas en contra de los demás —o utilizarse para crearlas— también podría emplearse para negociar ambiciosos acuerdos internacionales (y quizá en última instancia interplanetarios) para prohibir estos usos maliciosos.

No faltan razones por las que podría evitarse este nuevo Renacimiento. Podríamos discutir por los beneficios relativos de la IA y vernos envueltos en un conflicto internacional, perdiendo de vista al mismo tiempo los beneficios absolutos mucho mayores a disposición de todos; o podríamos implantar un sistema de IA que al principio parezca reflejar nuestros valores pero finalmente dé lugar a un estancamiento cultural y un debilitamiento humano. Sin embargo, al igual que con las dificultades técnicas anteriores, no conozco ninguna razón por la que estas dificultades políticas sean insuperables.

Agradecimientos

Gracias a John Danaher, Anders Sandberg, Ben Garfinkel, Carrick Flynn, Stuart Armstrong y Eric Drexler por sus útiles comentarios sobre las versiones anteriores de estas ideas. Cualquier error restante es responsabilidad del autor.

Referencias

AI Impacts, 2017. «AI hopes and fears in numbers» *AI Impacts* blog, https://aiimpacts.org/ai-hopes-and-fears-in-numbers/

Amodei, D., Olah, C., Steinhardt, J., Christiano, P., Schulman, J. y Mané, D. 2016. «Concrete Problems in AI Safety» arXiv preprint server, https://arxiv.org/abs/1606.06565

Armstrong, S., Bostrom, N., y Shulman, C. 2016. «Racing to the Precipice: a Model of Artificial Intelligence Development» *AI & Society*, pp. 1-6.

Bostrom, N. and Yudkowsky, E. 2011. «The Ethics of Artificial Intelligence» en *Cambridge Handbook of Artificial Intelligence*, ed. Ramsey, W. y Frankish, K.

Bostrom, N. 2014. *Superintelligence: Paths, dangers, strategies*. Oxford: Oxford University Press.

Bostrom, N. 2017. «Strategic Implications of Openness in AI Development» *Global Policy*, Vol. 8, Issue 2.

Bostrom, N., Dafoe, A., y Flynn, C. 2017. «Policy Desiderata in the Development of Machine Superintelligence» http://www.nickbostrom.com/papers/aipolicy.pdf

Brundage, M. 2016. «Economic Possibilities for Our Children: Artificial Intelligence and the Future of Work, Education, and Leisure» *2015 AAAI Workshop on AI, Ethics, and Society*.

Brundage, M. y Avin, S. et al. 2018. «The Malicious Use of Artificial Intelligence: Forecasting, Prevention, and Mitigation».

Bryson, J. 2016. «Patiency is Not a Virtue: AI and the Design of Ethical Systems» *2016 AAAI Spring Symposium Series*.

Grace, K., Salvatier, J., Dafoe, A., Zhang, B., y Evans, O. 2017. «When Will AI Exceed Human Performance? Evidence from AI Experts» arXiv preprint server, https://arxiv.org/abs/1705.08807

Kirkpatrick, K. 2016. «Battling Algorithmic Bias» *Communications of the ACM*, Vol. 59, No. 10, pp. 16-17, https://cacm.acm.org/magazines/2016/10/207759-battling-algorithmic-bias/abstract

Kurzweil, R. 2005. *The Singularity is Near: When Humans Transcend Biology*. Nueva York: Viking Press.

Trask, A. 2017. «Safe Crime Detection» https://iamtrask.github.io/2017/06/05/homomorphic-surveillance/

5. Observaciones sobre la inteligencia artificial y el optimismo racional

Olle Häggström

Introducción

El futuro de la inteligencia artificial (IA) y su repercusión en la humanidad es un tema importante. Se trató en una mesa redonda organizada por el grupo STOA (Evaluación de las Opciones Científicas y Tecnológicas) del Parlamento Europeo en Bruselas el 19 de octubre de 2017. Steven Pinker fue el ponente principal de la reunión, con Peter Bentley, Miles Brundage, Thomas Metzinger y yo como ponentes adicionales (véase el vídeo en STOA, 2017). Este ensayo se basa en mi preparación para ese acto, junto con algunas reflexiones (recicladas parcialmente del post de mi blog (Häggström, 2017)) sobre lo que dijeron otros ponentes en la reunión.

Optimismo

El título del acto del 19 de octubre contenía el término «optimismo racional», que inicialmente pensé que era un oxímoron, puesto que considero tanto el optimismo como el pesimismo distorsiones sesgadas de las pruebas disponibles. En particular, me parece *irracional* afirmar, sobre la base de pruebas insuficientes, que todo va a salirle bien a la humanidad. Sin embargo, pensándolo bien, decidí que existe un tipo distinto de optimismo que estoy más dispuesto a calificar de racional, en particular…

> …tener una visión epistemiológicamente bien calibrada del futuro y sus incertidumbres, aceptar que el futuro no está escrito en piedra, y actuar sobre la base de la suposición de que las posibilidades de un buen futuro pueden depender de las acciones que emprendamos hoy.

Obsérvese que la hipótesis de trabajo puede resultar ser (al menos parcialmente) incorrecta. Por ejemplo, quizá el mundo sea tan caótico que es inútil intentar juzgar si una determinada acción en la actualidad aumenta o reduce las posibilidades de un futuro largo y próspero para la humanidad. Si es el caso, nuestras acciones no importan (en un sentido predecible) para este futuro. Pero no sabemos si es así, por lo que tiene sentido *suponer* (aunque provisionalmente) que nuestras acciones sí importan e intentar determinar qué acciones mejoran nuestras posibilidades de un buen futuro. Este es el espíritu con el que se escribe el resto del ensayo.

Inteligencia artificial

Al igual que otras tecnologías emergentes como la biología sintética y la nanotecnología, la IA trae aparejados tanto enormes beneficios potenciales como enormes riesgos. Por lo que respecta a los beneficios, la empresa de consultoría de gestión McKinsey & Co publicó un informe en 2013 que cifraba el valor económico añadido por las innovaciones en IA y robótica a nivel mundial en los próximos diez años en 50 billones de dólares (Manyika et al. 2013; Omohundro, 2015), que sospecho que es una subestimación, en parte debido al ritmo imprevisto al que ha despegado el aprendizaje automático alimentado por los macrodatos desde entonces. Aunque no debemos cometer el error de pensar que el crecimiento económico y la mejora de la vida son automáticamente lo mismo, sigue estando claro que los avances en la IA pueden hacernos mucho bien. En una perspectiva a más largo plazo, apenas hay límites (aparte de las leyes de la física) al bien que puede hacer.

Los riesgos son de varios tipos. El que está más íntimamente relacionado con los beneficios económicos estimados es el problema de lo que puede hacer la automatización impulsada por la IA al mercado laboral. En el caso de los vehículos autónomos, un sector entero del mercado laboral, con millones de conductores de camiones, conductores de autobuses y conductores de taxis, corre el riesgo de ser totalmente eliminado en un horizonte de quizá no más de veinte años. ¿Encontrarían todas estas

personas trabajo en otro lugar o se quedarían desempleadas? Es probable que ocurran cosas similares a otros sectores del mercado laboral. Y, aunque la sustitución del trabajo humano por máquinas por supuesto no es un fenómeno nuevo, la revolución de la IA trae un cambio: ya no solo es el trabajo manual el que asumen las máquinas, sino cada vez más el trabajo intelectual. En combinación con la mayor velocidad de automatización, esto plantea serias preocupaciones sobre si se encontrarán nuevas tareas de trabajo humano a un ritmo que coincida con la automatización (como ha sucedido en gran parte anteriormente) o si las cifras de desempleo se dispararán; véase, por ejemplo, el libro de 2014 de Brynjolfsson y McAfee (2014). A largo plazo, quizá no sea ilusoria una situación hipotética limitante en la que las máquinas nos superen en todos los trabajos, dando lugar a un 100 % de desempleo. Esto plantea al menos dos cuestiones sociales cruciales. Primero, ¿cómo puede organizarse una sociedad en la que las personas no trabajan sino que invierten su tiempo en aspiraciones más elevadas como el arte, la cultura o el amor o simplemente en jugar a videojuegos tremendamente placenteros? Segundo, aunque podamos diseñar satisfactoriamente esta utopía, sigue estando la cuestión de cómo llevar a cabo la transición de la sociedad actual a la utopía sin crear por el camino niveles sin precedentes de desigualdad económica y malestar social.

Si esto suena moderadamente alarmante, considérese a continuación la cuestión de qué podría entrañar un mayor desarrollo de la tecnología de la IA para las armas autónomas. Aquí simplemente citaré un pasaje de una carta abierta de 2015 que firmé junto con otros miles de científicos (Russell et al., 2015):

> Si cualquier gran potencia militar impulsa el desarrollo de armas de IA, es prácticamente inevitable una carrera armamentística global, y el final de esta trayectoria tecnológica es obvio: las armas autónomas se convertirán en los kalashnikovs del mañana. A diferencia de las armas nucleares, no requieren materiales caros o difíciles de obtener, por lo que se volverán omnipresentes y baratas de producir en masa para todas las principales potencias militares. Solo será cuestión de tiempo hasta que aparezcan en el mercado negro y en manos de terroristas, dictadores que quieren controlar mejor a su población, señores de la guerra que quieren perpetrar una limpieza étnica, etc. Las armas autónomas son ideales para tareas como asesinatos, desestabilización de naciones, sometimiento de poblaciones y asesinatos selectivos de un determinado grupo étnico. Por lo tanto, creemos que una carrera armamentística militar de IA no sería beneficiosa para la humanidad.

En la reunión de Bruselas (STOA, 2017 a las 12:01:00 según el reloj mostrado en el vídeo), Pinker indicó una postura optimista con respecto a este riesgo militar de la IA: lo descartó destacando que se necesitaría un loco para construir algo tan horrible como «un enjambre de robots diseñados para atacar a personas basándose en el reconocimiento facial» y que no hay margen de maniobra para que los locos hagan estas cosas porque la ingeniería a día de hoy no la llevan a cabo genios solitarios sino grandes colaboraciones. Esta visión de color de rosa ignora totalmente cómo funcionan las carreras armamentísticas militares y el complejo militar-industrial, así como el hecho de que llevamos desarrollando armas de destrucción masiva igual de terribles desde hace más de setenta años. Esta evolución ha sido propiciada no por locos solitarios sino por grandes esfuerzos de colaboración (siendo el ejemplo más famoso el proyecto Manhattan); ¿por qué se detendrían de repente? La objeción de Pinker entra directamente en la categoría que califiqué antes de optimismo irracional.

Estos dos riesgos (riesgo de desigualdad económica resultante del aumento del desempleo y riesgo de una carrera armamentística de IA) deben tomarse en serio y debemos intentar determinar su gravedad y la forma mitigarlos. En las siguientes tres secciones, me centraré en un tercer tipo de riesgo de la IA, más exótico y especulativo que los dos anteriores, pero quizá no menos real: la aparición de una IA superinteligente cuyos valores no estén bien alienados con los nuestros.

Riesgo derivado de la superinteligencia

Suponga que los investigadores de IA un día cumplen su tan ansiado objetivo de crear una IA que sea superinteligente, entendiéndose como una máquina que nos supere a los seres humanos claramente en

toda la serie de competencias que calificamos como inteligencia. En ese momento, ya no podemos esperar seguir manteniendo el control. El experimento mental conocido como *Apocalipsis de los clips* puede servir de relato aleccionador (Bostrom, 2003):

Imagínese una fábrica de clips, que está dirigida por una IA avanzada (pero todavía no superinteligente), programada para maximizar la producción de clips. Sus ingenieros informáticos están intentando continuamente mejorarla y, un día, más o menos por accidente, consiguen que la máquina cruce el umbral en el que entra en una espiral de automejora en rápido aumento conocida como *explosión de inteligencia* o *singularidad*. Rápidamente se convierte en la primera IA superinteligente del mundo y, al haber mantenido su objetivo de maximizar la producción de clips, inmediatamente pasa a convertir todo el planeta (incluidos nosotros) en un montón gigantesco de clips para a continuación expandirse al espacio exterior con el fin de convertir el sistema solar, la Vía Láctea y después el resto del universo observable en clips.

Este ejemplo es caricaturesco a propósito para destacar que solo es una ilustración de un fenómeno mucho más general (hasta donde sé, nadie tiene miedo de que una IA convierta el mundo literalmente en clips). El objetivo es enfatizar que, para que un avance de IA se vuelva peligroso, no se necesitan malas intenciones: no necesitamos invocar a un científico loco que esté tramando destruir el mundo como venganza contra la humanidad. Incluso los objetivos que suenan inocentes como la maximización de la producción de clips pueden dar lugar a situaciones hipotéticas peligrosas.

Ahora bien…¿de verdad pueden? Dos de los ponentes de la reunión de Bruselas (Pinker y Bentley) expresaron muy enérgicamente la opinión de que no merece la pena tomarse en serio el riesgo de una catástrofe de superinteligencia. Parecían encantados de estar unidos en este punto de vista, pese a que las respectivas razones que destacaron eran muy diferentes.

Para abordar la cuestión de si el riesgo de una catástrofe de superinteligencia es real, ayuda dividirla en dos:

> 1) ¿Cabe esperar que el desarrollo de la IA llegue finalmente al punto de crear superinteligencia? De ser así, ¿cuándo y con qué rapidez?
>
> 2) Una vez creada, ¿qué estará inclinada a hacer la IA superinteligente? ¿Podría hacer algo peligroso?

Trataré estas dos subpreguntas por separado en las dos próximas secciones. Para que el riesgo de superinteligencia sea real, la respuesta a la pregunta 1) debe ser «sí» y la respuesta a la pregunta 2) debe ser «sí, podría hacer algo peligroso». En la reunión de Bruselas, Bentley cuestionó la respuesta a la pregunta 1), mientras que Pinker cuestionó la respuesta a la pregunta 2).

¿Cuándo (si es que alguna vez) podemos esperar superinteligencia?

Suponiendo una visión naturalista del mundo (por lo que la mente humana no surge a través del dualismo cartesiano de alguna chispa divina o algún otro tipo de magia), lo razonable es esperar que, cuando la evolución biológica dio con el cerebro humano, todavía no estaba ni siquiera cerca de lograr una forma globalmente óptima de configurar la materia para maximizar la inteligencia. Por lo tanto, debemos esperar que existan posibles configuraciones de materia que logren la superinteligencia. A partir de aquí, basta solo un pequeño salto para concluir (con el apoyo, por ejemplo, de la tesis de Church-Turing) que esta configuración puede simularse en un ordenador, en cuyo caso la superinteligencia es en principio viable mediante algún programa informático adecuado.

¿Qué grado de dificultad hay en encontrar este programa? No sabemos. El desarrollo de la IA ha logrado con mucho éxito, especialmente en los últimos años, construir IA para tareas específicas como conducir un coche o derrotar a seres humanos en juegos como el ajedrez o el Go. El progreso hacia la inteligencia artificial *general* (IAG) —una máquina que demuestra una inteligencia de nivel humano o mejor de una forma suficientemente flexible para funcionar en todos los ámbitos con los que nos encontramos

normalmente los seres humanos (ajedrez, baloncesto, desarrollo de software, cocina, cuidados, reconocimiento facial, conversación en una cena, etc.)— ha sido mucho menos impresionante. Algunos afirman que el progreso ha sido literalmente nulo, pero me parece un poco injusto. Por ejemplo, hace unos años se desarrolló una IA que aprendió rápidamente a jugar con éxito una serie de videojuegos Atari (Clark, 2015). Ciertamente está muy lejos de la capacidad de manejar toda la serie de tareas a las que se enfrentan los seres humanos en el mundo físico, pero sigue siendo una mejora no igual a cero al tener una habilitad especializada en solo un videojuego. Un posible camino hacia la IAG, entre muchos, podría ser una expansión gradual del ámbito en el que la máquina es capaz de actuar de manera inteligente.

Existen muchos enfoques posibles para crear software inteligente. Actualmente se observa un enorme auge del llamado aprendizaje profundo (LeCun et al. 2015), que es básicamente un renacimiento y mayor desarrollo de las viejas técnicas de redes neuronales que solían generar resultados poco impresionantes, pero que en la actualidad, gracias a máquinas más rápidas y acceso a enormes conjuntos de datos para formar a las máquinas, resuelven un problema importante tras otro. Este es un ejemplo del llamado método de caja negra, en el que los ingenieros que construyen con éxito una IA normalmente siguen sin entender cómo razona. Otro ejemplo del enfoque de caja negra es la programación genética, en la que una población de programas candidatos compite de una manera que imita los mecanismos de selección, reproducción y mutación de la evolución biológica. Sin embargo, existen otras formas (no de caja negra), en particular la llamada GOFAI («Good Old-Fashioned AI») en las que los conceptos y los procedimientos de razonamiento de la máquina son codificados manualmente por los programadores. También pueden existir métodos basados en la imitación del cerebro humano, ya sea mediante la comprensión del tipo de información de alto nivel que se procesa en el cerebro y que es clave para la IAG o (como defendió enérgicamente Kurzweil (2005)) copiando mediante fuerza bruta el funcionamiento exacto del cerebro con suficiente detalle (ya sean sinapsis o incluso niveles más bajos) para reproducir su comportamiento.

Quizá ninguno de estos planteamientos genere nunca la IAG, pero la postura razonable parece ser al menos estar abierto a la posibilidad de que uno de ellos, o alguna combinación, lleven finalmente a ella. ¿Pero cuándo? Esto parece aún más incierto, y un estudio realizado por Müller y Bostrom (2016) de las estimaciones de los cien investigadores de IA más citados del mundo refleja unas estimaciones repartidas a lo largo de todo el siglo actual (y más adelante). Su estimación mediana de la fecha de surgimiento de lo que podría denominarse IAG a nivel humano es 2050, con una estimación mediana del 50 % en el caso de surgimiento de superinteligencia en los siguientes treinta años. Véase asimismo el estudio más reciente (Grace et al., 2017). Teniendo en cuenta la enorme variación en la opinión de los expertos, sería epistemiológicamente imprudente tener una creencia firme sobre si se creará superinteligencia y cuándo, en lugar de aceptar de manera prudente y concienzuda que bien puede crearse dentro de décadas, o siglos o nunca.

Sin embargo, en la reunión de Bruselas, Peter Bentley dijo sobre la superinteligencia que «no va a surgir, ¡esa es la cuestión! Es totalmente irracional incluso concebir que surgirá» (STOA, 2017 en 12:08:45). ¿De dónde viene esa certeza absoluta? En su presentación, Bentley tenía básicamente solo un argumento para su posición: su experiencia y la de otros desarrolladores de IA de que todo el progreso en el ámbito requiere mucho trabajo y que cualquier nuevo algoritmo que inventan solo soluciona un problema concreto. Una vez que se cumple ese objetivo, la mejora inicialmente rápida del algoritmo siempre viene seguida por un punto de rendimiento decreciente. Por lo tanto (destacó), solucionar otro problema siempre exige el arduo trabajo de inventar y aplicar otro algoritmo más.

Esta línea de argumentación de Bentley esconde debajo de la alfombra un hecho conocido, a saber, que sí existen algoritmos con una capacidad de solución de problemas más ilimitada, como ejemplifica el software del cerebro humano. Su convicción al 100 % de que el ingenio científico humano en el próximo siglo (o la escala temporal que se elija adoptar) no logrará descubrir este algoritmo parece difícil de defender racionalmente: requiere una fe dogmática.

Para resumir esta sección: aunque sigue siendo una posibilidad que la IA nunca llegue a la superinteligencia, también es bastante plausible que lo haga a la larga. Suponiendo que lo haga, la fecha es sumamente incierta y, para tener debidamente en cuenta esta incertidumbre, debemos reconocer que podría ocurrir en cualquier momento durante el siglo actual y quizá incluso después. Y no debemos (como se subraya en un importante artículo de Sotala y Yampolskiy (2015)) caer en el tentador error de pensar que, solo por el mero hecho de que el momento del surgimiento de la superinteligencia es incierto, también debe ser temporalmente lejano.

¿Qué decidirá hacer una IA superinteligente?

Imaginemos la situación, en algún momento del futuro, en la que se ha desarrollado una IA superinteligente, una hipótesis que, como argumenté en la anterior sección, no es en absoluto inverosímil. Parece probable que en esta situación ya no tendremos el control y que nuestro destino dependerá de lo que decida hacer la IA, del mismo modo que, a día de hoy, el destino de los chimpancés depende de las decisiones tomadas por los seres humanos y no tanto de las decisiones tomadas por ellos. Una manera de intentar evitar esta conclusión es establecer formas de mantener la IA en una caja e incapaz de influir en el mundo más que a través de un estrecho canal de comunicaciones cuidadosamente controlado por administradores de la seguridad humana. Este denominado enfoque de la IA en una caja ha atraído cierta atención en la investigación de la seguridad de la IA (véase, por ejemplo, Armstrong et al., 2012), pero la conclusión general tiende a ser que el control de un ser superinteligente es una tarea demasiado difícil de cumplir para simples humanos y que lo mejor que podemos esperar es mantener la IA en una caja durante un período temporal y más bien breve.

Así que imaginemos ahora que la IA superinteligente ya no está en una caja, sino que es capaz de recorrer libremente internet (incluida la internet de las cosas), crear numerosas copias de seguridad de sí misma, utilizar su inteligencia superior para atravesar (o sortear) los cortafuegos que se pongan en su camino, etc. Ya no tenemos el control, y la futura supervivencia y bienestar de la humanidad dependerán de lo que decida hacer la máquina. ¿Qué decidirá hacer? Depende de cuáles sean sus objetivos. Predecirlo no es tarea fácil y todo debate sobre ello tiene que ser especulativo al menos en cierto grado. Sin embargo, existe un marco que nos permite ir más allá de la mera especulación, en particular lo que decidí (Häggström, 2016) llamar *la teoría Omohundro-Bostrom de los objetivos finales frente a los instrumentales de la IA* (Omohundro, 2008; Bostrom, 2012, 2014). Esta teoría no está escrita en piedra de la manera que lo está un teorema matemático establecido, por lo que puede estar abierto a revisión, junto con las predicciones que hace; sin embargo, la teoría es lo suficientemente plausible para que merezca la pena tomarse en serio sus predicciones. Tiene dos pilares: la tesis de la ortogonalidad y la tesis de la convergencia instrumental. Los explicaré por orden.

La tesis de la ortogonalidad afirma (aproximadamente) que prácticamente cualquier objetivo final es compatible con niveles arbitrariamente elevados de inteligencia. Se pueden construir contraejemplos artificiosos basados en la idea de paradojas autorreferenciales (un contraejemplo podría ser «mantén tu nivel de inteligencia general por debajo del de un perro medio de 2017»), pero la idea es que, al margen de esto, puedes programar cualquier función de objetivo para que tu IA intente optimizarla y el objetivo es posible para IA de inteligencia arbitrariamente elevada. Los novatos de la teoría Omohundro-Bostrom y de la futurología de la IA en general a menudo replicarán que un objetivo estrecho de miras como es la maximización de los clips es intrínsecamente estúpido y que, por lo tanto, es contradictorio sugerir que una IA superinteligente pueda tener este objetivo, pero se confunde inteligencia con objetivos: la inteligencia es simplemente la capacidad para dirigir el mundo hacia objetivos específicos, sean cuales sean. La maximización de los clips nos *parece* estúpida a nosotros, pero no es porque *sea* estúpida en un sentido objetivo, sino porque es contraria a *nuestros* objetivos.

Ahora pasemos a *la tesis de la convergencia instrumental*. La IA puede adoptar distintos objetivos instrumentales, no como objetivos en sí mismos, sino como instrumentos para promover su objetivo final. La tesis de la convergencia instrumental afirma que hay una serie de objetivos instrumentales que

puede esperarse que adopte la IA para una gama extremadamente amplia de objetivos finales que pueda tener. Algunos objetivos instrumentales a los que parece aplicarse la tesis son…

- la supervivencia (¡no dejes que te desconecten!);
- adquisición de hardware y otros recursos;
- mejora del software y hardware propios;
- preservación del objetivo final; y
- Si el objetivo final no está en consonancia con los valores humanos, mantener un comportamiento discreto (ocultar tu objetivo o tu capacidad) hasta que llegue el momento en que puedas superar fácilmente toda la resistencia humana.

Un caso típico de cómo funciona la lógica es el primer objetivo instrumental de la lista: la supervivencia. Más o menos independientemente de su objetivo final, es probable que la IA calcule que estará en mejor posición de promover este objetivo si existe y está en funcionamiento en oposición a si se destruye o apaga. Por lo tanto, tiene sentido que la IA se resista a nuestros intentos de apagarla. Puede utilizarse un razonamiento similar para motivar los demás objetivos instrumentales de la lista. El objetivo instrumental de mejora del software y hardware propios es lo que cabe esperar que desencadene la IA, una vez que sea suficientemente inteligente para ser buena en el diseño de inteligencia artificial, para entrar en el tipo de espiral de automejora mencionado anteriormente y que puede o no resultar suficientemente rápida (dependiendo de la intrincada cuestión de si el denominado rendimiento de la reinversión cognitiva principalmente está aumentando o disminuyendo; véase Yudkowsky, 2013) para justificar la calificación de explosión de inteligencia.

La idea de la convergencia instrumental a menudo se pierde en las críticas del discurso sobre el riesgo de la superinteligencia. En particular, en la reunión de Bruselas, me decepcionó escuchar a Pinker decir lo siguiente, tan solo unos minutos después de que hubiese explicado los fundamentos de la teoría de Omohundro-Bostrom y el caso especial de la supervivencia:

> Si diésemos a la máquina el objetivo de sobrevivir, haría cualquier cosa, incluso destruirnos, para ello. La forma de evitarlo: ¡no construir esos estúpidos sistemas! (STOA, 2017, 11:57:45)

Pasa por alto lo central, que es que la teoría de Omohundro-Bostrom nos da razones para creer que una IA suficientemente inteligente probablemente adoptará el objetivo instrumental de supervivencia, independientemente de que los programadores humanos le hayan dado explícitamente ese objetivo.

El caso de la preservación del objetivo final es especialmente interesante. Puede resultar tentador pensar que una IA con el objetivo de maximizar los clips verá, si llega a un nivel suficientemente alto de inteligencia, lo limitado y tonto que es ese objetivo y pasar a otra cosa. Así que imaginemos a la IA contemplando un cambio a otro objetivo que merezca más la pena (para nosotros), como la conservación de los ecosistemas. Se pregunta «¿qué es mejor: ceñirse a la maximización de clips o cambiar a la conservación de los ecosistemas?», pero qué significa aquí «mejor», es decir, ¿cuál es el criterio para evaluar cuál de estos objetivos es preferible? Pues bien, dado que la IA todavía no ha cambiado de objetivo, sino que simplemente está contemplando hacerlo, su objetivo sigue siendo la maximización de los clips, por lo que el criterio de evaluación será «¿qué objetivo dará lugar a un mayor número de clips?». La repuesta a esta pregunta es muy probablemente «la maximización de los clips», lo que motivará a la IA a ceñirse a ese objetivo. Este es el mecanismo básico detrás del objetivo instrumental de la preservación del objetivo final.

Debido a este mecanismo, es improbable que una IA superinteligente nos permita alterar su objetivo final, por lo que si tiene como objetivo final la maximización de clips, probablemente estemos condenados. Por lo tanto, tenemos que inculcar a la IA objetivos que nos gusten más antes de que llegue a los niveles de superinteligencia. Este es el propósito del programa de investigación *AI alignment* (Alineación de la IA), formulado (con el título alternativo de *friendly AI* (IA amigable), que quizá es mejor evitar puesto que tiene una connotación innecesariamente antropomórfica) en un influyente

artículo de Yudkowsky (2008) y muy debatido desde entonces (véase, por ejemplo, Bostrom, 2014; Häggström, 2016; Tegmark, 2017). Para atacar el problema sistemáticamente, puede dividirse en dos. Primero, el problema técnico de cómo cargar los objetivos deseados en la IA. Segundo, el problema ético de cuáles son estos objetivos deseados o quién los determina y a través de qué procedimiento (democrático o de otro tipo). Ambos son extremadamente difíciles. Por ejemplo, una percepción clave que se remonta por lo menos a Yudkowsky (2008) es que los valores humanos son muy frágiles, en el sentido de que equivocarse un poco en ellos puede provocar una catástrofe en manos de una IA superinteligente. La razón por la que deberíamos trabajar hoy en la alineación de la IA no es que sea probable que la superinteligencia esté a la vuelta de la esquina (aunque véase Yudkowsky, 2017), sino que, si está a décadas de distancia, solucionar la alineación de la IA bien puede requerir estas décadas con poco o ningún margen de postergación.

Cuando Pinker, en el pasaje citado anteriormente en esta sección, dice «La forma de evitarlo: ¡no construir esos estúpidos sistemas!», podría interpretarse como una *defensa* del trabajo de alineación de la IA. Sin embargo, me parece que esta formulación no transmite la dificultad del problema y da la impresión errónea de que la alineación de la IA no requiere una atención seria.

¿Debemos callarnos al respecto?

En su argumentación en la reunión de Bruselas en contra de tomarse en serio el riesgo apocalíptico de la IA, Pinker señaló (STOA, 2017, 11:51:40) que la ciudadanía ya tiene la amenaza nuclear y climática por las que preocuparse; por lo tanto, afirmó, plantear otro riesgo mundial más puede abrumar a la gente y provocar que simplemente renuncien al futuro. Puede haber algo de cierto en esta especulación, pero, para evaluar el mérito del argumento, tenemos que considerar por separado las dos posibilidades: a) que el riesgo apocalíptico de la IA sea real, y b) que el riesgo apocalíptico de la IA sea espurio.

En el caso de b), *por supuesto* que no deberíamos malgastar tiempo y esfuerzo en debatir este riesgo, pero no necesitábamos el argumento de la abrumación de la ciudadanía para entenderlo. Considérese en cambio el caso a). Aquí, la recomendación de Pinker equivale a simplemente ignorar una amenaza que puede matarnos a todos. No me parece que sea una buena idea. Desde luego sería maravilloso sobrevivir a la amenaza nuclear y solucionar la crisis climática, pero su utilidad se ve gravemente obstaculizada en caso de que nos conduzca a un apocalipsis de la IA. Mantenerse callado sobre el riesgo real también parece ir directamente en contra de una de las ideas más preciadas de Pinker durante la última década o más, a saber, la de la apertura científica e intelectual y los valores de la Ilustración en general. Lo mismo se aplica a la situación en la que no estamos seguros de si es a) o b); seguramente, el enfoque que mejor se corresponde con los valores de la Ilustración es debatir abiertamente el problema e intentar determinar si el riesgo es real.

Conclusión y lectura adicional

El surgimiento de la superinteligencia puede, si nos hemos preparado para ello con suficiente cuidado, resultar lo mejor que le pase a la humanidad, pero también trae aparejado un grave riesgo catastrófico. Este riesgo y los riesgos más realistas de la IA examinados anteriormente merecen nuestra atención. No es que *vaya* a producirse un apocalipsis de la IA, sino que es suficientemente plausible como para que merezca la pena intentar determinar cómo *prevenirlo*. Esto es lo que he defendido en el presente ensayo. Sin embargo, he sido muy breve y aconsejo a los lectores que quieran leer el desarrollo de mi argumento en mayor detalle que consulten el capítulo 4 de mi libro (Häggström, 2016). Para una explicación aún más detallada, recomiendo encarecidamente los libros de Bostrom (2014) y Tegmark (2017). De ellos, el libro de Tegmark está dirigido más claramente a un público amplio, mientras que el de Bostrom es más académico, pero ambos contienen (con algunos solapamientos) muchas ideas sorprendentes e importantes.

Agradecimientos

Agradezco a Björn Bengtsson sus valiosos comentarios sobre el manuscrito.

Referencias

Armstrong, S., Sandberg, A. y Bostrom, N. (2012) Thinking inside the box: controlling and using an oracle AI, *Minds and Machines* 22, 299-324.

Bostrom, N. (2003) Ethical issues in advanced artificial intelligence, *Cognitive, Emotive and Ethical Aspects of Decision Making in Humans and in Artificial Intelligence, Vol. 2* (ed. Smit, I. et al.) International Institute of Advanced Studies in Systems Research and Cybernetics, pp. 12-17.

Bostrom, N. (2012) The superintelligent will: motivation and instrumental rationality in advanced artificial agents, *Minds and Machines* 22, 71-85.

Bostrom, N. (2014) *Superintelligence: Paths, Dangers, Strategies*, Oxford University Press, Oxford.

Brynjolfsson, E. y McAfee, A. (2014) *The Second Machine Age: Work, Progress, and Prosperity in a Time of Brilliant Technologies*, W.W. Norton, Nueva York.

Clark, L. (2015) DeepMind's AI is an Atari gaming pro now, *Wired*, 25 de febrero.

Grace, K., Salvatier, J., Dafoe, A., Zhang, B. y Evans, O. (2017) When will AI exceed human performance? Evidence from AI experts, *arXiv*:1705.08807.

Häggström, O. (2016) *Here Be Dragons: Science, Technology and the Future of Humanity*, Oxford University Press, Oxford.

Häggström, O. (2017) The AI meeting in Brussels last week, *Häggström hävdar*, 23 de octubre.

Kurzweil, R. (2005) *The Singularity Is Near: When Humans Transcend Biology*, Viking, Nueva York.

LeCun, Y., Bengio, Y. y Hinton, G. (2015) Deep learning, *Nature* 521, 436-444.

Manyika, J., Chui, M., Bughin, J., Dobbs, R. Bisson, P. y Marrs, A. (2013) Disruptive technologies: Advances that will transform life, business, and the global economy, *McKinsey Global Institute*.

Müller, V. y Bostrom, N. (2016) Future progress in artificial intelligence: A survey of expert opinion. En *Fundamental Issues of Artificial Intelligence,* Springer, Berlín, pp. 553-571.

Omohundro, S. (2008) The basic AI drives, *Artificial General Intelligence 2008: Proceedings of the First AGI Conference* (Wang, P., Goertzel, B. and Franklin, S., eds), IOS, Ámsterdam, pp 483-492.

Omohundro, S. (2015) McKinsey: $50 trillion of value to be created by AI and robotics through 2025, *Self-Aware Systems*, 4 de agosto.

Russell, S. et al. (2015) *Autonomous Weapons: An Open Letter from AI and Robotics Researchers*, Future of Life Institute.

Sotala, K. y Yampolskiy, R. (2015) Responses to catastrophic AGI risk: a survey, *Physica Scripta* 90, 018001.

STOA (2017), Vídeo de la reunión del STOA el 19 de octubre de 2017, https://web.ep.streamovations.be/index.php/event/stream/171019-1000-committee-stoa/embed

Tegmark, M. (2017) *Life 3.0: Being Human in the Age of Artificial Intelligence*, Brockman Inc, Nueva York.

Yudkowsky, E. (2008) Artificial intelligence as a positive and negative factor in global risk, en *Global Catastrophic Risks* (eds. Bostrom, N. y Ćirković, M.), Oxford University Press, Oxford, pp 308-345.

Yudkowsky, E. (2013) *Intelligence Explosion Microeconomics*, Machine Intelligence Research Institute, Berkeley, CA.

Yudkowsky, E. (2017) *There's No Fire Alarm for Artificial General Intelligence*, Machine Intelligence Research Institute, Berkeley, CA.

6. Hacia una carta mundial sobre la inteligencia artificial

Thomas Metzinger

Introducción

Ha llegado el momento de trasladar el debate público actual sobre la inteligencia artificial (IA) a las propias instituciones políticas. Muchos expertos consideran que nos enfrentamos a un punto de inflexión en la historia durante la próxima década y que hay una ventana temporal que se cierra en relación con la ética aplicada de la IA. Por lo tanto, las instituciones políticas deben producir *y* aplicar un conjunto mínimo pero suficiente de limitaciones éticas y jurídicas para el uso beneficioso y el futuro desarrollo de la IA. También deben crear un proceso racional y empírico de debate crítico destinado a actualizar continuamente, mejorar y revisar este primer conjunto de limitaciones normativas. Teniendo en cuenta la situación actual, el resultado por defecto es que los valores rectores del desarrollo de la IA sean establecidos por un pequeño número de seres humanos, grandes empresas privadas e instituciones militares. Por lo tanto, un objetivo es integrar proactivamente tantas perspectivas como sea posible y de manera oportuna.

Ya han surgido numerosas iniciativas diferentes en todo el mundo que están investigando activamente los recientes avances en IA en relación con cuestiones relativas a la ética aplicada, sus aspectos jurídicos, las futuras implicaciones socioculturales, los riesgos existenciales y la formulación de políticas.[4]Hay un acalorado debate público y algunos quizá incluso tengan la impresión de que importantes instituciones políticas como la Unión no son capaces de reaccionar a una velocidad adecuada a los nuevos riesgos tecnológicos y a la creciente preocupación entre la ciudadanía. Por lo tanto, debemos aumentar la agilidad, eficiencia y sistematicidad de los esfuerzos políticos actuales para aplicar normas desarrollando un proceso democrático más formal e institucionalizado y quizá incluso nuevos modelos de gobernanza.

Para iniciar un proceso más sistemático y estructurado, presentaré una lista concisa y no exclusiva de las esferas problemáticas más importantes, cada una con recomendaciones prácticas. La primera esfera problemática que se examinará es la que, en mi opinión, viene constituida por aquellas cuestiones que tienen menos posibilidades de resolverse. Por lo tanto, debe abordarse en un proceso de múltiples capas, empezando por la propia Unión Europea (UE).

El problema de la «carrera hacia el abismo»

Tenemos que formular y aplicar normas de seguridad mundiales para la investigación de la IA. Es necesaria una carta *mundial* sobre la IA, porque dichas normas de seguridad solo pueden ser eficaces si conllevan un compromiso vinculante con determinadas normas por parte de *todos* los países que participan e invierten en el tipo pertinente de investigación y desarrollo. Habida cuenta del competitivo contexto económico y militar actual, la seguridad de la investigación de la IA muy probablemente se reducirá a favor de un progreso más rápido y un coste reducido, en concreto trasladándola a países con normas de seguridad laxas y escasa transparencia política (una analogía sólida y obvia es el problema de la evasión fiscal por parte de empresas y fideicomisos). Si se logran una cooperación y coordinación internacionales, en principio podría evitarse una «carrera hacia el abismo» en las normas de seguridad

[4] Para consultar una sinopsis de las iniciativas existentes, véase Baum 2017 y Boddington 2017, p. 3. Me he abstenido de presentar aquí una documentación completa, pero algunos puntos de entrada útiles a la literatura son Mannino et al. 2015, Stone et al. 2016, IEEE 2017, Bostrom, Dafoe y Flynn 2017, Madary y Metzinger 2016 (para la realidad virtual).

(mediante la relocalización de la investigación científica e industrial de la IA). Sin embargo, el panorama actual de incentivos hace que sea un resultado muy improbable.

Recomendación 1

La Unión debería elaborar de inmediato una carta europea sobre la IA.

Recomendación 2

En paralelo, la Unión debería iniciar un proceso político conducente a la elaboración de una carta mundial sobre la IA.

Recomendación 3

La Unión debería invertir recursos en el refuerzo sistemático de la cooperación y la coordinación internacionales. Debería minimizarse la desconfianza estratégica; pueden definirse puntos en común a través de hipótesis extremadamente negativas.

La segunda esfera problemática que se examinará puede decirse que está constituida por el conjunto de cuestiones más urgentes y estas también tienen pocas posibilidades de solucionarse en grado suficiente.

Prevención de una carrera armamentística de IA

Redunda en interés de los ciudadanos de la Unión que se evite una carrera armamentística de IA desde el principio, por ejemplo entre China y Estados Unidos. Una vez más, quizá sea demasiado tarde para ello, y evidentemente la influencia europea es limitada, pero la investigación y el desarrollo de armas ofensivas autónomas debería prohibirse y no financiarse en el territorio de la Unión. Las armas autónomas seleccionan y atacan a objetivos sin intervención humana y actuarán en una escala de tiempo y reacción cada vez más corta, lo que a su vez hará que sea racional transferir cada vez más autonomía humana a estos sistemas. Por lo tanto, pueden crear contextos militares en los que sea racional ceder el control humano casi totalmente. En esta esfera problemática, el grado de complejidad es incluso mayor que en la prevención del desarrollo y la proliferación de armas nucleares, por ejemplo, porque la mayor parte de la investigación en la materia no tiene lugar en universidades públicas. Además, si la humanidad se ve forzada a una carrera armamentística a este nuevo nivel tecnológico, el *propio* proceso histórico de una carrera armamentística puede volverse autónomo y resistir a intervenciones políticas.

Recomendación 4

La Unión debería prohibir *toda* la investigación sobre armas ofensivas autónomas en su territorio y buscar acuerdos internacionales.

Recomendación 5

En cuanto a las aplicaciones militares puramente defensivas, la Unión debería financiar la investigación del máximo grado de autonomía de los sistemas inteligentes que parezca aceptable desde una perspectiva ética y jurídica.

Recomendación 6

A nivel internacional, la Unión debería poner en marcha una gran iniciativa para prevenir el surgimiento de una carrera armamentística de IA, utilizando todos los instrumentos diplomáticos y políticos disponibles.

La tercera esfera problemática que se examinará es aquella en la que el horizonte predictivo sigue siendo probablemente bastante lejano, pero donde la incertidumbre epistémica es alta y el posible daño podría ser extremadamente grande.

Una moratoria en la fenomenología sintética

Es importante que todos los políticos entiendan la diferencia entre inteligencia artificial y conciencia artificial. La creación involuntaria o incluso intencional de conciencia artificial es muy problemática desde una perspectiva ética, porque puede dar lugar a sufrimiento artificial y a un sentido de identidad experimentado de manera consciente en los sistemas inteligentes autónomos. La «fenomenología sintética» (un término acuñado por analogía con la «biología sintética») se refiere a la posibilidad de crear no solo inteligencia general, sino también conciencia o experiencias subjetivas en sistemas artificiales avanzados. Los futuros sujetos artificiales de experiencia no tienen representación en el proceso político actual, no tienen estatuto jurídico y sus intereses no están representados en ningún comité de ética. Para tomar decisiones éticas, es importante entender qué sistemas naturales y artificiales tienen la capacidad de producir conciencia y en particular experimentar estados negativos como el sufrimiento[5]. Un posible riesgo es aumentar drásticamente la cantidad general de sufrimiento en el universo, por ejemplo a través de cascadas de copias o la rápida duplicación de sistemas conscientes a gran escala.

Recomendación 7

La Unión debería prohibir toda la investigación que corra el riesgo de crear o esté directamente dirigida a crear una fenomenología sintética en su territorio, y buscar acuerdos internacionales.[6]

Recomendación 8

Teniendo en cuenta el nivel actual de incertidumbre y desacuerdo dentro del incipiente ámbito de la conciencia de las máquinas, hay una necesidad acuciante de promover, financiar y coordinar los proyectos de investigación interdisciplinaria en la materia (que comprenden filosofía, neurociencia e informática). Los temas pertinentes específicos son modelos conceptuales, neurobiológicos y computacionales empíricos de experiencia consciente, conciencia propia y sufrimiento.

Recomendación 9

A nivel de investigación fundacional, es necesario promover, financiar y coordinar la investigación sistemática sobre la ética aplicada de los sistemas no biológicos capaces de tener experiencia consciente, conciencia propia y sufrimiento experimentado subjetivamente.

La siguiente esfera problemática general que se examinará es la más compleja y la que probablemente contenga el mayor número de problemas imprevistos e incógnitas desconocidas.

Peligros para la cohesión social

La tecnología de IA avanzada claramente brindará numerosas posibilidades para optimizar el propio proceso político, incluidas oportunidades novedosas de ingeniería social racional basada en valores y formas de gobernanza más eficientes con base empírica. Por otro lado, no solo es plausible suponer que existen muchos nuevos riesgos y peligros, actualmente desconocidos, que pueden socavar el proceso de mantenimiento de la coherencia de nuestras sociedades; también es racional suponer la existencia de un gran número de incógnitas desconocidas, de riesgos relacionados con la IA que solo descubriremos por accidente y en una fase tardía. Por lo tanto, la Unión debería asignar *recursos separados* a la preparación para situaciones en las que de repente se descubran incógnitas desconocidas imprevistas.

---

[5] Véase Metzinger 2013, 2017.

[6] Incluidos enfoques que persigan una confluencia de la neurociencia y la IA con el objetivo específico de fomentar el desarrollo de conciencia de las máquinas. Para consultar ejemplos recientes, véase Dehaene, Lau y Kouider 2017, Graziano 2017, Kanai 2017.

Muchos expertos creen que el riesgo más próximo y bien definido es el desempleo masivo por la automatización. Por lo tanto, la aplicación de tecnología de IA por partes interesadas económicamente poderosas puede dar lugar a un gradiente de ingresos más profundo, mayor desigualdad y peligrosos patrones de estratificación social. Los riesgos concretos son amplios recortes salariales, el colapso del impuesto sobre la renta, más una sobrecarga de los sistemas de seguridad social. Sin embargo, la IA plantea muchos otros riesgos para la cohesión social, por ejemplo mediante los medios sociales de propiedad privada y controlados de forma autónoma destinados a captar la atención humana y «empaquetarla» para su ulterior uso por clientes o en la «ingeniería» de la formación de voluntad política a través de estrategias de *Big Nudging* (macroencauzamiento) y arquitecturas de decisiones controladas por la IA, que no son transparentes para los ciudadanos cuyo comportamiento se controla de esta manera. La futura tecnología de IA será extremadamente buena en la modelización y el control predictivo del comportamiento humano, por ejemplo mediante el refuerzo positivo y sugerencias indirectas, haciendo que el cumplimiento de determinadas normas o la aparición «espontánea» de «motivos» y decisiones parezcan totalmente no forzados. En combinación con el *Big Nudging* y el control predictivo de usuarios, la tecnología de vigilancia inteligente también podría aumentar los riesgos mundiales ayudando *localmente* a estabilizar regímenes autoritarios de manera eficiente. Una vez más, muy probablemente, la mayoría de estos riesgos para la cohesión social todavía son desconocidos en la actualidad y tal vez solo los descubramos por accidente. Los responsables políticos también deben entender que cualquier tecnología que optimice deliberadamente la inteligibilidad de su propia acción para los usuarios humanos también puede en principio optimizarla para *engañar*. Por lo tanto, debe tenerse mucho cuidado de evitar la especificación accidental o incluso intencionada de la función de recompensa de cualquier IA de una manera que pueda perjudicar indirectamente al bien común.

La tecnología de IA es actualmente un bien privado. Las instituciones políticas democráticas tienen la obligación de convertir una gran parte de ella en un bien *común* protegido, algo que pertenezca a toda la humanidad. En la tragedia de los comunes, todos pueden ver lo que viene, pero si no existen mecanismos para contrarrestar eficazmente la tragedia, se producirá, por ejemplo en situaciones descentralizadas. La Unión debería desarrollar proactivamente estos mecanismos.

Recomendación 10

Dentro de la Unión, el aumento de la productividad relacionada con la IA debe distribuirse de una manera socialmente justa. Evidentemente, la práctica del pasado y las tendencias mundiales apuntan de forma clara a la dirección contraria: (casi) nunca lo hemos hecho en el pasado y los incentivos financieros existentes contrarrestan directamente esta recomendación.

Recomendación 11

La Unión debería investigar cuidadosamente el potencial de una renta básica sin condiciones o un impuesto sobre la renta negativo en su territorio.

Recomendación 12

Se necesitan programas de investigación sobre la viabilidad de iniciativas de reciclaje profesional programadas en el momento oportuno dirigidas a los estratos de la población amenazados para desarrollar aptitudes creativas y dotes sociales.

La siguiente esfera problemática es difícil de abordar, porque la mayor parte de la investigación puntera sobre IA ya ha salido de las universidades e instituciones de investigación financiadas con fondos públicos. Está en manos de empresas privadas y, por lo tanto, es sistemáticamente no transparente.

Ética de la investigación

Uno de los problemas teóricos más difíciles reside en definir las condiciones en las que sería racional renunciar por completo a determinadas vías de investigación de la IA (por ejemplo las que conlleven el surgimiento de fenomenología sintética o una evolución explosiva de sistemas que se autooptimicen de manera autónoma y no estén alineados de manera fiable con los valores humanos). ¿Cuáles serían las hipótesis mínimas concretas que justifiquen una moratoria en determinadas ramas de la investigación? ¿Cómo tratarán las instituciones democráticas a los agentes deliberadamente poco éticos en una situación en la que la toma de decisiones colectiva es poco realista y escalonada y en la que deben crearse formas no globales de cooperación *ad hoc*? Se han planteado cuestiones similares en la llamada «investigación de ganancia de función» que conlleva la experimentación destinada a aumentar la transmisibilidad o virulencia de patógenos, como determinadas cepas del virus de la gripe H5N1 de alta patogenicidad, la viruela o el ántrax. En este caso, los investigadores de la gripe se impusieron a sí mismos de forma loable una moratoria voluntaria y temporal. En principio, esto también sería posible en la comunidad de investigación de la IA. Por lo tanto, la Unión debería complementar siempre su carta sobre la IA con un código de conducta ética concreto para investigadores que trabajen en proyectos financiados.

Sin embargo, el objetivo más profundo sería desarrollar una *cultura de sensibilidad moral* más amplia dentro de las propias comunidades de investigación en la materia. La determinación y minimización racional y empírica de los riesgos (también los relativos a un futuro más lejano) debería formar parte de la propia investigación, y los científicos deberían cultivar una actitud proactiva, especialmente si son los primeros en ser conscientes de los nuevos tipos de riesgos a través de su propio trabajo. La comunicación con el público, en caso necesario, debería ser por iniciativa propia, un acto de asumir el control y actuar antes de una situación futura, en vez de solo reaccionar a las críticas de no expertos con algún conjunto de normas formales preexistentes. Como escriben Madary y Metzinger (2016, p. 12) en su código de conducta ética que incluye recomendaciones para una buena práctica científica en la realidad virtual, los científicos deben entender que seguir un código deontológico no es lo mismo que *ser* ético. Un código deontológico específico para el ámbito, por muchas versiones futuras coherentes, desarrolladas y bien estructuradas que tenga, nunca podrá sustituir al propio razonamiento ético.

Recomendación 13

La carta mundial sobre la IA, o su precursora europea, deberían complementarse siempre con un código de conducta ética concreto que oriente a los investigadores en su trabajo práctico cotidiano.

Recomendación 14

Debe formarse a una nueva generación de expertos en ética aplicada especializados en problemas de tecnología de IA, sistemas autónomos y ámbitos relacionados. La Unión debería invertir de forma sistemática e inmediata en el desarrollo de la futura especialización necesaria dentro de las instituciones políticas pertinentes y debería hacerlo aspirando a un nivel especialmente elevado de excelencia académica y profesionalidad por encima de la media.

Metagobernanza y la diferencia de ritmo

Como se señaló brevemente en el párrafo introductorio, la aceleración del desarrollo de la IA quizá se ha convertido en el ejemplo *paradigmático* de divergencia extrema entre los enfoques gubernamentales existentes y lo que se necesitaría para optimizar la ratio riesgo/beneficio de manera oportuna. Se ha convertido en el ejemplo paradigmático de presión temporal, en términos de determinación, evaluación y gestión racional y empírica de los riesgos emergentes, creación de directrices deontológicas y aplicación de un conjunto de normas jurídicas ejecutables. Existe un «problema de seguimiento del ritmo». Las estructuras de gobernanza existentes simplemente no son capaces de responder a este

desafío con suficiente rapidez; la supervisión política ya se ha quedado muy por detrás de la evolución tecnológica.[7]

No estoy llamando la atención sobre la situación actual porque quiera desatar la alarma o concluir con un tono distópico y pesimista. Más bien, lo que quiero decir es que la adaptación de las *propias* estructuras de gobernanza forma parte del panorama de problemas: para eliminar o al menos minimizar la diferencia de ritmo, tenemos que invertir recursos en cambiar la estructura de los propios enfoques de gobernanza. La «metagobernanza» significa justo eso: una gobernanza *de* la gobernanza frente a los riesgos y posibles beneficios de un crecimiento explosivo de determinados sectores de desarrollo tecnológico. Por ejemplo, Wendell Wallach ha señalado que la supervisión eficaz de las tecnologías emergentes exige una combinación tanto de normativas vinculantes aplicadas por organismos gubernamentales como mecanismos de gobernanza no vinculantes ampliados.[8] Por lo tanto, Marchant y Wallach han propuesto los llamados «comités de coordinación de la gobernanza» (CCG), un nuevo tipo de institución que proporciona un mecanismo para coordinar y sincronizar lo que acertadamente describen como una «explosión de estrategias, acciones, propuestas e instituciones de gobernanza»[9] con trabajo existente en instituciones políticas establecidas. Un CCG para la IA podría actuar como «administrador de problemas» para una tecnología específica que emerge con rapidez, centro de intercambio de información, sistema de alerta temprana, instrumento de análisis y seguimiento, evaluador de mejores prácticas internacionales, y fuente independiente y fiable a la que recurrir para éticos, medios de comunicación, científicos y partes interesadas. Como indican Marchant y Wallach, *la influencia de un CCG en la satisfacción de la necesidad crítica de una entidad de coordinación central dependerá de su capacidad para establecerse como intermediario honesto respetado por todas las partes interesadas pertinentes*.[10]

Por supuesto, también son concebibles muchas otras estrategias y enfoques de gobernanza. Este no es el lugar para tratar los detalles. Aquí, la idea general es simplemente que solo podemos afrontar el desafío que plantea el rápido desarrollo de la IA y los sistemas autónomos si priorizamos en nuestra agenda la cuestión de la metagobernanza desde el principio.

Recomendación 15

La Unión debería invertir en la investigación y el desarrollo de nuevas estructuras de gobernanza que aumenten drásticamente la velocidad con la que las instituciones políticas establecidas pueden responder a los problemas y aplicar realmente nuevas normativas.

Conclusión

He propuesto que la Unión comience a trabajar inmediatamente en la elaboración de una carta mundial sobre la IA, en un proceso de múltiples capas que empiece por una carta sobre la IA para la propia

---

[7] Gary Marchant (2011) deja muy clara la cuestión en el resumen de un capítulo de un libro reciente: *Las tecnologías emergentes se están desarrollando a un ritmo cada vez más acelerado, mientras que los mecanismos jurídicos para su posible supervisión están ralentizándose, si acaso. La legislación a menudo está estancada, la regulación con frecuencia se encuentra osificada y los procedimientos judiciales a veces se describe que avanzan a ritmo glacial. Esta divergencia entre la velocidad de la tecnología y la ley tiene dos consecuencias. En primer lugar, algunos problemas son supervisados por marcos reguladores cada vez más obsoletos y desactualizados. En segundo lugar, otros problemas carecen por completo de supervisión significativa. Para abordar esta creciente diferencia entre la ley y la regulación, serán necesarios nuevos instrumentos, enfoques y mecanismos jurídicos. No bastará con mantener el status quo*.

[8] Véase Wallach 2015 (Chapter 14), p. 250.

[9] Esta cita se ha extraído de un borrador preliminar no publicado titulado «An agile ethical/legal model for the international and national governance of AI and robotics»; véase asimismo Marchant y Wallach 2015.

[10] Marchant y Wallach 2015, p. 47.

Unión. Para ilustrar brevemente algunas de las cuestiones básicas desde mi perspectiva como filósofo, he descrito cinco importantes esferas temáticas y formulado quince recomendaciones generales para su debate crítico. Evidentemente, esta aportación no pretendía ser una lista exclusiva o exhaustiva de las cuestiones pertinentes. Por el contrario: en esencia, la ética aplicada de la IA no es en absoluto un campo para grandes teorías o debates ideológicos, sino sobre todo un problema de gestión racional y sobria de riesgos que entraña diferentes horizontes predictivos con una gran incertidumbre. Sin embargo, una parte importante del problema es que no podemos depender de intuiciones, porque debemos satisfacer limitaciones de racionalidad contraintuitivas.

Permítanme terminar citando un documento de política reciente titulado *Artificial Intelligence: Opportunities and Risks*, publicado por la Effective Altruism Foundation de Berlín (Alemania):

En las situaciones en las que hay que tomar una decisión y lo que está en juego es muy importante, los siguientes principios son de vital importancia:

1. Merece la pena el coste de las precauciones onerosas incluso para los riesgos de baja probabilidad, siempre que haya suficiente que ganar/perder con ello.
2. Cuando haya escaso consenso entre los expertos en un ámbito, es aconsejable la modestia epistémica. Es decir, no debe tenerse demasiada confianza en la exactitud de la opinión propia en cualquier caso[11].

Referencias

Adriano, Mannino; Althaus, David; Erhardt, Jonathan; Gloor, Lukas; Hutter, Adrian; Metzinger, Thomas (2015): Artificial Intelligence. Opportunities and Risks. En: *Policy Papers of the Effective Altruism Foundation* (2), S. 1–16. https://ea-foundation.org/files/ai-opportunities-and-risks.pdf.

Baum, Seth (2017): A Survey of Artificial General Intelligence Projects for Ethics, Risk, and Policy. Global Catastrophic Risk Institute Working Paper 17-1. https://ssrn.com/abstract=3070741.

Boddington, Paula (2017): Towards a Code of Ethics for Artificial Intelligence. Cham: Springer International Publishing (Artificial Intelligence: Foundations, Theory, and Algorithms).

Bostrom, Nick; Dafoe, Allan; Flynn, Carrick (2017): Policy Desiderata in the Development of Machine Superintelligence. working Paper, Oxford University. http://www.nickbostrom.com/papers/aipolicy.pdf.

Dehaene, Stanislas; Lau, Hakwan; Kouider, Sid (2017): What is consciousness, and could machines have it? En: *Science (New York, N.Y.)* 358 (6362), S. 486–492. DOI: 10.1126/science.aan8871.

Graziano, Michael S. A. (2017): The Attention Schema Theory. A Foundation for Engineering Artificial Consciousness. En: *Frontiers in Robotics and AI* 4, S. 61. DOI: 10.3389/frobt.2017.00060.

Madary, Michael; Metzinger, Thomas K. (2016): Real virtuality. A code of ethical conduct. recommendations for good scientific practice and the consumers of VR-technology. En: *Frontiers in Robotics and AI* 3, S. 3. http://journal.frontiersin.org/article/10.3389/frobt.2016.00003/full

Marchant, Gary E. (2011): The growing gap between emerging technologies and the law. En Marchant, Gary E.; Allenby, Braden R.; Herkert, Joseph R. (Hg.): The Growing Gap Between Emerging Technologies and Legal-Ethical Oversight: Springer, S. 19–33.

Marchant, Gary E.; Wallach, Wendell (2015): Coordinating technology governance. En: *Issues in Science and Technology* 31 (4), S. 43.

[11] Cf. Mannino et al. 2015.

Metzinger, Thomas (2013): Two principles for robot ethics. En: In Hilgendorf, Eric; Günther, Jan-Philipp (Hg.) (2013): Robotik und Gesetzgebung: BadenBaden, Nomos S. 247–286. https://www.blogs.uni-mainz.de/fb05philosophieengl/files/2013/07/Metzinger_RG_2013_penultimate.pdf

Metzinger, Thomas (2017): Suffering. En: Kurt Almqvist und Anders Haag (Hg.): The Return of Consciousness. Estocolmo: Axel and Margaret Ax:son Johnson Foundation, S. 237–262. https://www.blogs.uni-mainz.de/fb05philosophieengl/files/2013/07/Metzinger_Suffering_2017.pdf

Kanai, Ryota (2017): We Need Conscious Robots. How introspection and imagination make robots better. En: *Nautilus* (47). http://nautil.us/issue/47/consciousness/we-need-conscious-robots.

Stone, Peter; et al. (2016): Artificial Intelligence and Life in 2030. One Hundred Year Study on Artificial Intelligence: Report of the 2015-2016 Study Panel. Stanford, CA: Stanford University. https://ai100.stanford.edu/2016-report.

The IEEE Global Initiative on Ethics of Autonomous and Intelligent Systems (2017): Ethically Aligned Design. A Vision for Prioritizing Human Well-being with Autonomous and Intelligent Systems. http://standards.ieee.org/develop/indconn/ec/auto_sys_form.html.

Wallach, W. (2015): A Dangerous Master. How to Keep Technology from Slipping Beyond Our Control. Nueva York: Basic Books.

Para bien o para mal, se prevé que la inteligencia artificial (IA) tendrá una enorme repercusión en el futuro de la humanidad. A medida que llegan nuevas promesas y preocupaciones a un público cada vez más amplio, el debate está empezando a captar la imaginación pública. En esta publicación, presentamos cuatro artículos de opinión, cada uno de los cuales responde a la pregunta ¿debemos temer a la IA? Los cuatro autores provienen de distintos ámbitos disciplinarios y presentan perspectivas divergentes sobre si debemos temer el futuro de la IA y cómo deberíamos proceder con su desarrollo.

Los avances en la inteligencia artificial han inspirado formidables esperanzas y temores, muchos de ellos apenas fundamentados en la realidad. Esta magnífica recopilación, de verdaderos expertos, aplica la racionalidad y el análisis a esta esfera emocional y es indispensable para cualquiera que quiera entender uno de los temas más importantes de nuestro día.

Steven Pinker Johnstone, Catedrático de Psicología en la Universidad de Harvard y autor de *Enlightenment Now: The Case for Reason, Science, Humanism, and Progress.*

Publicación de la
Unidad de Previsión Científica (STOA)
EPRS | Servicio de Estudios del Parlamento Europeo

PE 581.948
ISBN 978-92-846-3388-3
doi: 10.2861/61195
QA-01-18-199-ES-N