Gobernanza y control institucional para el futuro de la inteligencia artificial

Antonio Diéguez

Un mecanismo que persiga metas no buscará necesariamente nuestras metas, a menos que lo diseñemos para ese propósito, y en ese diseño debemos prever todos los pasos del proceso para el que está diseñado…
(Wiener, 1964: 63-64)

Si existe siquiera una pequeña probabilidad de que se dé la singularidad, haríamos bien en pensar sobre las formas que podría tomar y si hay algo que podamos hacer para influir en los resultados en la dirección positiva.
(Chalmers, 2010: 10)

La situación en la que estamos

Los desarrollos recientes en inteligencia artificial (en adelante IA) están empezando a tener una influencia decisiva en nuestras formas de vida. Sus aplicaciones en la biomedicina, en el

transporte, en el procesamiento del lenguaje natural, en la economía y las finanzas, en la tecnología militar, en la vigilancia y control de los ciudadanos, en la robótica, etc., son tan imponentes como desconcertantes para muchos, y su aparente inexorabilidad e inmanejabilidad, en sintonía con lo que ha sucedido con los avances de la biotecnología, ha suscitado la preocupación de numerosos analistas. Hay aquí, en efecto, problemas éticos, sociales y políticos implicados que reclaman un examen profundo. Por si esto no fuera suficiente, los posibles efectos negativos de estos avances en IA y en biotecnología han sido exacerbados en la imaginación popular en no poca medida por el discurso transhumanista.

No es sorprendente, por ello, que la gobernanza de la tecnología se haya convertido en un asunto prioritario en la agenda política (Dafoe, 2018), y quizás ya no suene tan melodramático como antes, sino incluso realista, decir que se trata de un asunto en el que se juega la propia supervivencia de nuestra especie. Pese a ello, seguimos con instituciones y sistemas regulatorios que, a lo sumo, muestran su funcionalidad en relación con la tecnología de la tercera revolución industrial (revolución digital), pero que parecen obsoletos para regular las tecnologías de la cuarta (unión de tecnologías digitales, particularmente la IA y las redes de sistemas inteligentes, la robótica, el internet de las cosas, las tecnologías de nuevos materiales, la nanotecnología y las biotecnologías). Una revolución que ha comenzado ya.[1]

1 Tal como suele emplearse hoy la expresión 'cuarta revolución industrial', se la entiende con el significado que hemos mencionado, que es a su vez el que le dio en 2016 el economista, ingeniero y fundador del Foro Económico Mundial Klaws Schwab. Para Schwab, la primera revolución industrial habría sido la de la máquina de vapor y fue desde mediados del XVIII hasta mediados del XIX, la segunda, a finales del XIX y principios del XX, habría sido la del acero, el petróleo, la electricidad y la cadena de montaje, la tercera habría sido la de la industria de la computación, la revolución digital, la revolución de las tecnologías de la información y la comunicación (TICs), surgida en los años 60 del siglo XX y todavía vigente, y la cuarta, de la síntesis de varias tecnologías emergentes, como las citadas. No obstante, otros autores habían empleado anteriormente la expresión 'cuarta revolución', como el filósofo oxoniense Luciano Floridi, quien se centra más bien en cambios culturales y de conocimiento y se basa en su clasificación en la que ya hiciera hace años Bruce Mazlish, basada a su vez en las tres heridas de Freud. La primera, según Floridi, habría sido la revolución copernicana, con la que surge la ciencia moderna, la segunda habría sido la revolución darwiniana, con la que se fundamenta una visión naturalista del ser humano, la tercera habría sido la freudiana, con la que comienza el estudio de los aspectos inconscientes de nuestra mente, y la cuarta sería la revolución de las TICs, que habría desplazado al ser humano como ocupante único de la infosfera, y habría tenido su punto de origen en Alan Turing y el desarrollo de las ciencias de la computación (Floridi, 2014).

Como bien explica Luciano Floridi, en un libro que lleva precisamente por título *The Fourth Revolution* (2014), el reto que tenemos delante no es tanto el que puedan presentar las innovaciones tecnológicas como tales, sino el que plantea la propia gobernanza de lo digital. Algunos legisladores y expertos son conscientes de la magnitud del desafío ante el que estamos, pero hay dudas razonables de que puedan ejercer una influencia decisiva en el plano legal e institucional al ritmo que sería exigible. Así describe la apremiante situación el conocido empresario y economista Klaus Schwab, fundador del Foro Económico Mundial:

> Con el rápido ritmo de cambio provocado por la cuarta revolución industrial, los reguladores están siendo desafiados en un grado sin precedentes. Las autoridades políticas, legislativas y regulatorias de hoy se ven a menudo superadas por los acontecimientos y son incapaces de lidiar con la velocidad del cambio tecnológico y la importancia de sus implicaciones. El ciclo de noticias de veinticuatro horas ejerce presión sobre los líderes para comentar o actuar de forma inmediata frente a los acontecimientos, lo cual reduce el tiempo disponible para llegar a respuestas calibradas, medidas y razonadas. Hay un peligro real de pérdida del control sobre lo que importa, sobre todo en un sistema mundial con casi doscientos estados independientes y miles de culturas y lenguas diferentes. […].
>
> Muchos de los avances tecnológicos que actualmente vemos no son correctamente tenidos en cuenta en el actual marco regulatorio y podrían incluso causar una ruptura del contrato social que los gobiernos han establecido con sus ciudadanos. Un gobierno ágil significa que los reguladores deben encontrar formas de adaptarse continuamente a un nuevo entorno de rápidos cambios y reinventarse para entender mejor lo que están regulando. Para ello, los gobiernos y los organismos reguladores necesitan colaborar estrechamente con las empresas y la sociedad civil con el fin de diseñar las transformaciones necesarias en los planos global, regional e industrial.
>
> […] En la era de la cuarta revolución industrial, lo que se necesita no es necesariamente más legislación o que esta sea más rápida, sino más bien un ecosistema regulatorio y legislativo que pueda producir estructuras más resistentes. Este enfoque podría mejorarse mediante la creación de un espacio de mayor

> sosiego que permita reflexionar sobre las decisiones importantes. El reto es hacer de esta deliberación algo mucho más productivo de lo que lo es hoy en día e incluir la previsión para crear el máximo espacio para la innovación. (Schwab, 2016: 59-60)

En lo que sigue expondré, en primer lugar, las líneas fundamentales del debate actual sobre los posibles efectos de la IA. A continuación, argumentaré acerca de la necesidad de promover instituciones que faciliten la defensa de los derechos de los ciudadanos frente a los desafíos actuales o potenciales de la IA y defenderé la idea de que lo que importa a la hora de discutir sobre esta necesidad no es si los sistemas empleados pueden ser considerados como realmente inteligentes, sino la importancia de las decisiones que vayan a ser puestas en sus manos. Esbozaré asimismo algunas directrices que creo que podrían ser de ayuda en la clarificación de lo que exigiría un eventual control efectivo de la IA.

Algunas notas sobre las posibilidades de la IA

¿Qué pasaría –se pregunta Stuart Russell (2019) al comienzo de su libro *Human Compatible*– si los científicos que trabajan en el campo de la IA tuviesen éxito en los próximos años? ¿Lo han pensado con detenimiento? ¿Saben qué implicaría conseguir *todos* los objetivos principales de su trabajo, que se resumen en crear máquinas inteligentes? Russell considera que esta es la mayor cuestión a la que se enfrenta la humanidad, puesto que de todos los futuros que se nos anuncian, más o menos distópicos, él estima que el de la creación de una IA superinteligente es el más probable.

Puede parecer una pregunta exagerada, pero no carece por completo de base. A finales de 2015 Elon Musk y Sam Altman fundaron la empresa OpenAI, ante la preocupación de que el desarrollo futuro de una inteligencia artificial general (en adelante IAG) represente un riesgo real para la supervivencia de nuestra especie. El propósito de OpenAI, que ahora cuenta con el apoyo de Microsoft, es crear una IAG amigable, es decir, una inteligencia artificial versátil y compleja, superior a la humana en su capacidad para resolver problemas y alcanzar objetivos muy distintos, pero que sea en cualquier circunstancia beneficiosa para los seres humanos y esté sometida a sus intereses. La empresa DeepMind, hoy perteneciente a Alphabet, cuya principal filial es

Google, tiene como uno de sus lemas, tal como consta en su página web, el de "resolver la inteligencia", un propósito que se concreta en la tarea de despejar el camino para la aparición de esa IAG amigable.

Al parecer, la expresión inteligencia artificial general, así como su acrónimo (en inglés es AGI), fueron acuñados por Shane Legg, el cofundador de DeepMind, en torno a 2007, fecha en la que apareció el primer libro dedicado expresamente a este tema, editado por Ben Goertzel y Cassio Pennachin (Heaven, 2020; Goertzel y Pennachin, 2007).[2] En dicha obra, sus editores definen la IAG de la siguiente manera: son "sistemas de IA que poseen un grado razonable de autocomprensión y autocontrol autónomo, y tienen la capacidad para resolver una variedad de problemas complejos en una variedad de contextos, y para aprender a resolver nuevos problemas que no conocían en el momento de su creación" (Goertzel y Pennachin, 2007: VI).

En realidad, podemos decir que la creación de una IAG fue en sus orígenes, de forma más o menos implícita, el objetivo último del campo disciplinar de la Inteligencia Artificial (recuérdese la motivación que inspiró el Solucionador General de Problemas, creado por Herbert Simon, J. C. Shaw y Allen Newell en 1957), aunque la confianza en la posibilidad de que esto se consiguiera en un plazo previsible ha variado mucho según los momentos y según los autores. En palabras de Russell, la meta permanente ha sido construir un sistema "que aprenda lo que necesite aprender de todos los recursos disponibles, que haga preguntas cuando sea necesario, y que comience a formular y a ejecutar planes que funcionen" (Russell, 2019: 46). Y esto, después de todo, es lo que consideramos que hace la inteligencia humana.

Hasta el momento, sin embargo, todos los logros en este campo han sido en el desarrollo de lo que se conoce como inteligencia artificial particular, específica o estrecha. Se trata de sistemas computacionales que despliegan una gran capacidad, superior incluso a la humana, para realizar tareas muy específicas y bien definidas, como jugar a un juego con reglas fijas (ajedrez, go, damas, videojuegos, etc.), responder a preguntas de cultura general, realizar diagnósticos médicos precisos (enfermedades infecciosas, tipos de cáncer, medicina personalizada, etc.), diseñar medicamentos, identificar caras y otras imágenes, reconocer, procesar e interpretar la voz humana, traducir de un

2 Una buena caracterización de la inteligencia artificial general y de las capacidades que debería poseer puede encontrarse en Voss (2007).

idioma a otro, etc. En realidad, buena parte de lo que consideramos hoy como inteligencia artificial son sistemas de minería de datos, es decir, sistemas computacionales muy potentes que analizan cantidades masivas de datos y obtienen a partir de ellos patrones desconocidos y lo que podríamos calificar como conocimiento nuevo sobre esos datos.

Ahora bien, por extraordinarios que sean estos logros, estos sistemas no alcanzan la complejidad y flexibilidad de la inteligencia humana. Los dispositivos o sistemas más inteligentes de los que disponemos pueden ser utilizados con eficacia en tareas muy diferentes a aquellas para las que fueron programados (a lo sumo pueden aprender a jugar varios juegos), ni pueden ejercer un control autónomo. Hay incluso quienes piensan que ni siquiera deberíamos llamarles inteligentes, puesto que la única inteligencia real e innegable que aparece en ellos es la del programador humano o la de los seres humanos en cuyo contexto social estos sistemas cumplen alguna función (Julia, 2019; Marcos, 2021; Sánchez-Migallón, 2021).

En general, se asume que una máquina es inteligente cuando es capaz de realizar tareas tales que decimos que requieren de inteligencia cuando las lleva a cabo un ser humano. Esta es una definición operativa aceptada por muchos de los que trabajan en ese campo. Establece que los sistemas de inteligencia artificial se consideran como inteligentes por sus resultados, no por su naturaleza. En un sentido parecido los define un informe de la OCDE sobre el desarrollo de la IA: "Un sistema de IA –dice– es un sistema basado en una máquina que puede, para un conjunto dado de objetivos humanamente definidos, hacer predicciones, recomendaciones o tomar decisiones que influyen en ambientes reales o virtuales. Los sistemas de IA están diseñados para funcionar con distintos niveles de autonomía" (OEDC, 2020). Por su parte, el grupo de expertos sobre IA creado por la Comisión Europea en 2018 se atreve con una definición mucho más detallada:

> Los sistemas de inteligencia artificial (IA) son sistemas de *software* (y en algunos casos también de *hardware*) diseñados por seres humanos que, dado un objetivo complejo, actúan en la dimensión física o digital mediante la percepción de su entorno a través de la obtención de datos, la interpretación de los datos estructurados o no estructurados que recopilan, el razonamiento sobre el conocimiento o el procesamiento de la información derivados de esos datos, y decidiendo la acción o acciones óptimas que deben llevar a cabo para lograr

> el objetivo establecido. Los sistemas de IA pueden utilizar normas simbólicas o aprender un modelo numérico; también pueden adaptar su conducta mediante el análisis del modo en que el entorno se ve afectado por sus acciones anteriores. (Grupo de expertos de alto nivel sobre inteligencia artificial, 2019a)

No obstante, como veremos después, la propia caracterización de la inteligencia es un viejo problema que está lejos de haber generado un consenso. Y como no es fácil dirimir esta cuestión, no debe extrañar que tampoco haya acuerdo sobre cómo definir la inteligencia artificial (Wang, 2019).[3]

Aceptemos, sin embargo, que en un sentido no meramente metafórico podemos hablar de inteligencia artificial. ¿Debemos entonces temer a la IAG, en caso de que se logre su creación en el futuro? ¿Tendremos máquinas superinteligentes que tomarán el control de todo el planeta? ¿Podremos unirnos a las máquinas alguna vez y llegar a ser cíborgs, o podremos volcar nuestra mente en un ordenador, fundiéndonos así con la inteligencia artificial, y conseguir de ese modo un soporte imperecedero, y, por tanto, la inmortalidad, como los transhumanistas aseguran que sucederá inevitablemente? Son preguntas que se repiten a menudo cuando se discute sobre el futuro de la IA en los medios de comunicación y en los libros de divulgación, y creo que, por extrañas que parezcan, deben ser tomadas en serio, aunque, como diré a continuación, no deben ser el centro del debate.

El peligro es lo que hagan las máquinas, no cómo lo hagan

Se han publicado varios libros en los últimos años insistiendo en la idea de que la IA será la tecnología dominante en el futuro, la que configurará todo nuestro entorno. En ellos, después de que se nos presente una imagen optimista o sombría de ese futuro, según los casos, el autor suele explicarnos cómo cree que podremos controlar y dominar dicha tecnología para que no termine dañando a los seres humanos (*e.g.*, Bostrom, 2014; Tegmark, 2017; Russell, 2019). El asunto, que se conoce ya como "el problema del control", se ha tornado ineludible para el especialista, pero también para el filósofo. Su discusión

3 Es interesante a este respecto ver las réplicas al artículo de Wang en el número especial *11*(2) de la revista *Journal of General Artificial Intelligence*, publicado en febrero de 2020.

es interesante, pero tiene el desafortunado efecto de dejar en un lugar menor otros problemas ligados al desarrollo de la IA que son, sin embargo, mucho más urgentes.

No conviene olvidar que, con independencia de si el desarrollo futuro de una IAG superior a la humana puede representar un peligro para la supervivencia de nuestra especie, lo que por el momento constituye un desafío desde el punto de vista de la salvaguarda de los derechos de las personas son ciertas aplicaciones de la IA cuyas consecuencias negativas sobre esos derechos se están viendo ya. Los ejemplos son múltiples. El más notorio sea quizás el uso de nuestros datos personales por parte de sistemas de IA pertenecientes a las grandes empresas tecnológicas, cuyo poder es cada vez mayor. Pero también hay que señalar los sesgos y la opacidad de los algoritmos usados en la toma de decisiones importantes para la vida de las personas, ya sea en la contratación de personal, en la concesión de créditos, o en otras muchas circunstancias laborales y sociales. También resulta muy preocupante la utilización de la IA en la identificación de rostros en lugares públicos y en la búsqueda de delincuentes y la prevención del delito, y no digamos ya su uso para la vigilancia y represión de disidentes políticos. A todos estos peligros se han añadido otros últimamente con repercusiones geoestratégicas, como la creación de armas autónomas, la extensión de los ciberataques, de las noticias falsas y de la desestabilización política mediante la desinformación. Acerca de estos peligros reales y actuales ya empiezan a aparecer también libros muy recomendables, como *Weapons of Math Destruction*, de Cathy O'Neil (2016), *Rebooting AI*, de Gary Marcus y Ernest Davis (2019), *AI Ethics*, de Mark Coeckelbergh (2020), y *Privacy is Power*, de Carissa Véliz (2020), por citar solo algunos. Mientras que el problema del control se refiere a cómo evitar que nos dañe una futura IAG autónoma, estas amenazas provienen del *uso* que los seres humanos están dando a los sistemas de inteligencia artificial hoy disponibles.

Con todo, para no dejar una imagen marcadamente negativa, es pertinente añadir que la IA está siendo también un instrumento muy eficaz, y se espera que lo sea aún más en el futuro, en la persecución de delitos financieros, en la protección de la seguridad de las personas, en la potenciación del progreso biomédico, en el logro de una creciente eficiencia energética y en la protección el medio ambiente. Estos beneficios potenciales, todo sea dicho, justifican asumir algunos riesgos inevitables (Agar, 2016).

Creo que, para analizar las consecuencias posibles, tanto favorables como desfavorables, del desarrollo futuro de la inteligencia artificial, discutir si se trata de inteligencia genuina, similar a la humana, con posibilidad de ser consciente o no, o, como se suele decir, si es verdadera inteligencia o solo simulación de inteligencia, es desviar el foco del auténtico problema.[4] Quizás tenga razón Luciano Floridi cuando, parafraseando a Clausiewitz, dice que la IA es la continuación de la inteligencia humana por medios estúpidos. Quizás acierte también Erik J. Larson cuando afirma que lo que llamamos inteligencia artificial opera solo mediante inferencias inductivas, triturando conjuntos de datos para hacer predicciones, mientras que la inteligencia humana lo hace sobre todo mediante inferencias abductivas, capaces de elaborar de forma creativa conjeturas diferentes en función del contexto (Larson, 2021). Quizás haya que hacer caso a Adriana Braga y Robert K. Logan (2021) cuando sostienen que la inteligencia artificial carece (y carecerá siempre) de atributos esenciales en la inteligencia humana, como la autoconsciencia, la voluntad, la finalidad, la curiosidad, la pasión, el deseo, la imaginación, la intuición, la moralidad, la sabiduría, el humor, etc. Puede incluso que la mejora de los sistemas de IA para resolver de forma rápida y precisa más y mejores problemas lleve a la creación de máquinas con una inteligencia cada vez menos análoga a la de los humanos (Signorelli, 2019).

No se debería, pues, ser muy categórico en la aceptación de la posibilidad futura de una inteligencia artificial similar a la humana, dada las dificultades bien conocidas para dotar a los sistemas computacionales de algunas de las cualidades mencionadas y, sobre todo, de eso que en el ser humano llamamos "sentido común", que en el fondo es la capacidad para comprender las situaciones en contextos cambiantes. Entre los propios especialistas no hay acuerdo al respecto. Algunos parecen pensar que es cuestión de tiempo el que se consiga mientras que otros creen que se trata de un objetivo quimérico.

Ahora bien, tampoco parece que se pueda ser muy categórico en el rechazo de la posibilidad de una IA de este tipo, especialmente si la inteligencia humana es caracterizada de una forma no demasiado estrecha, como hace Gottfredson (1997: 13); por ejemplo, cuando la define como "la capacidad mental que, entre otras cosas, implica la capacidad para razonar, planear,

4 En todo caso, el libro de Marcus y Davis (2019) sitúa esta discusión en términos muy sensatos. Una pregunta interesante es si una IAG, fuera consciente o no, debería tener derechos. Pero la mera consideración de esta posibilidad de la consciencia artificial es demasiado especulativa por el momento.

resolver problemas, pensar de forma abstracta, comprender ideas complejas, aprender rápidamente y aprender de la experiencia". Todas esas capacidades, con la posible excepción de la de pensar de forma abstracta y la de comprender ideas complejas, la realizan ya las máquinas, y ni siquiera estas dos excepciones pueden considerarse *a priori* como imposibles de alcanzar, en el supuesto de que podamos en algún momento caracterizarlas con precisión. Debe tenerse en cuenta además que la inteligencia posee grados. No es lo mismo la inteligencia de un insecto que la de un gran simio o la de un ser humano. La IA futura podría ser muy diferente de la inteligencia humana; aun así, cabría considerarla como una forma de inteligencia en algún sentido relevante. Como escribe Pei Wang, "«inteligencia humana», «inteligencia artificial/de computador/de máquina», e «inteligencia» deben tomarse como tres conceptos diferentes, siendo el último el que proporciona una generalización adecuada de los dos primeros" (2019: 27).

Pero dejando de lado este asunto, como digo, lo que me parece que debería preocuparnos ante todo es qué podrán hacer con nosotros las máquinas que creemos en el futuro si es que estas tienen capacidades y poder para tomar decisiones por nosotros, decisiones que se consideren en la práctica como inapelables en su autoridad. No es cómo piensen esas máquinas (si es que a lo que hagan le podemos llamar pensamiento), ni si su inteligencia es como la nuestra, lo que importa, sino cómo actúen, puesto que serán agentes que interactuarán con los humanos y no sabemos cómo se comportarán. Lo relevante no es si esas máquinas entienden o no sus decisiones, en el sentido de entender que utiliza Searle en su conocido argumento de la habitación china, ni tampoco si esas decisiones surgen de un libre albedrío genuino o de interacciones complejas e imprevisibles para nosotros de los algoritmos que ellas implementan; lo que importa son las decisiones que tomarán y sus efectos prácticos sobre nosotros.

En otras palabras, incluso si estas máquinas no son propiamente máquinas inteligentes, pero son sistemas adaptativos y autónomos, dirigidos a fines, y sus decisiones nos afectan vitalmente, entonces el problema del control sigue siendo un problema ineludible. La cuestión es que tendremos que bregar con dispositivos capaces de generar conocimiento a partir de datos, de resolver problemas en contextos diferentes, de tener el control de una gran cantidad de situaciones y de modificar la realidad, incluyendo en ella la concerniente a los seres humanos, y podrían en principio llegar a hacer esto con independencia de

los deseos que podamos tener, pero siendo ya tan absoluta nuestra dependencia de ellos que sería impensable su desconexión. Ese es el verdadero problema que conviene plantearse.

Sobre la necesidad del control institucional de la investigación en IA

Muchos investigadores destacados en el campo de la IA son ciertamente muy críticos con las visiones radicales que difunden Elon Musk y otros defensores del transhumanismo y consideran que la preocupación por una IAG capaz de convertirse en una superinteligencia fuerte, con el poder y los medios adecuados para amenazar nuestra existencia, carece de fundamentos sólidos. Aun así, puesto que no se trata de algo imposible, y dado el enorme riesgo que encerraría la creación de una tal superinteligencia, no sería excesivo reclamar que hasta que no haya ciertas garantías de que el problema del control pueda resolverse, cualquier investigación encaminada a la creación de una IAG debería regularse de forma rigurosa y vigilarse con cuidado. A sabiendas de que no cabe esperar una completa seguridad en este desarrollo, dado que, como ha argumentado Bostrom (2014: 117ss.), todo puede ir muy bien en el diseño de sistemas de IA particulares o estrechos, e incluso en las primeras fases de la creación de una IAG, y en cualquier momento, de forma inesperada, irreversible y fatal, el camino podría torcerse. Pero, como dije antes, la importancia de la regulación institucional en el campo de la IA es independiente de que se acepte o no la plausibilidad de esta situación extrema y de que se vea o no en la IA un riesgo existencial real. No debería, al menos, estar centrada en esa posibilidad, ni condicionada por ella.

Philip Pettit ha defendido, de forma convincente en mi opinión, que el control institucional es la mejor forma de control democrático frente a otras dos alternativas, que serían la influencia causal popular directa sobre el gobierno (en este caso sobre las empresas y laboratorios que desarrollan la IA) y la dirección intencional popular mediante asociaciones o grupos (Pettit, 2008). A nadie se le oculta que la influencia causal popular directa sobre las empresas tecnológicas es minúscula por lo general y la dirección intencional popular es poco efectiva en estos casos y difícil de llevar a cabo. Si trasladamos esto a la cuestión que nos ocupa, puede decirse que se ha vuelto ya imprescindible la creación de agencias o comités independientes, aunque con un sólido

implante institucional público, encargadas de la regulación de la investigación en IA. ¿Pero cómo darle contenido a todo esto? Ese es el reto.

Aun a riesgo de acercarse demasiado al pensamiento desiderativo, sabiendo, por ejemplo, que la segunda potencia mundial en IA es China, y quizás pronto la primera, y que Rusia también pretende ser una potencia en este ámbito, creo que entre las exigencias básicas para poder intentar un control efectivo en los desarrollos futuros de la IA deberían estar las siguientes:[5]

5 En todo caso, estas directrices que señalo no están muy lejos de otras que ya se han hecho, como las cinco recomendaciones sobre la IA realizadas en 2020 por la OCDE ((i) crecimiento inclusivo, desarrollo sostenible y bienestar; (ii) valores centrados en el ser humano y equidad; (iii) transparencia y explicabilidad; (iv) robustez, seguridad y protección; y (v) responsabilidad), o las siete recomendaciones que realizó en 2019 el Grupo de Expertos de la Comisión Europea ((i) agencia y supervisión humana, (ii) robustez técnica y seguridad, (iii) privacidad y gobernanza de datos, (iv) transparencia, (v) diversidad, no discriminación y equidad, (vi) bienestar ambiental y social, y (vii) responsabilidad), o las doce recomendaciones, más detalladas, que hizo el grupo The Public Voice (2018), que son las siguientes:

1. Derecho a la transparencia. Todas las personas tienen derecho a conocer la base de una decisión de IA que les concierna. Esto incluye el acceso a los factores, la lógica y las técnicas que produjeron el resultado.
2. Derecho a la determinación humana. Todas las personas tienen derecho a una determinación final hecha por una persona.
3. Obligación de identificación. La institución responsable de un sistema de IA debe darse a conocer al público.
4. Obligación de equidad. Las instituciones deben garantizar que los sistemas de inteligencia artificial no reflejen prejuicios injustos ni tomen decisiones discriminatorias inadmisibles.
5. Obligación de evaluación y rendición de cuentas. Un sistema de IA debe implementarse solo después de una evaluación adecuada de su propósito y objetivos, y de sus beneficios tanto como de sus riesgos. Las instituciones deben ser responsables de las decisiones tomadas por un sistema de IA.
6. Obligaciones de precisión, confiabilidad y validez. Las instituciones deben garantizar la precisión, confiabilidad y validez de las decisiones.
7. Obligación de calidad de los datos. Las instituciones deben establecer la procedencia de los datos y garantizar la calidad y relevancia de los datos introducidos en los algoritmos.
8. Obligación de seguridad pública. Las instituciones deben evaluar los riesgos de seguridad pública que surgen del despliegue de sistemas de IA que dirigen o controlan dispositivos físicos e implementan controles de seguridad.
9. Obligación de ciberseguridad. Las instituciones deben proteger los sistemas de IA contra las amenazas de ciberseguridad.
10. Prohibición de perfiles secretos. Ninguna institución establecerá o mantendrá un sistema de perfil secreto.
11. Prohibición de la puntuación unitaria. Ningún gobierno nacional establecerá o mantendrá una puntuación de propósito general para sus ciudadanos o residentes.
12. Obligación de rescisión. Una institución que ha establecido un sistema de inteligencia artificial tiene la obligación positiva de terminar con el sistema si el control humano del mismo ya no es posible.

1. *Seguridad en el diseño*. Este requisito es el mínimo exigible, porque sin esa seguridad falla todo lo demás. Como dice Joanna J. Bryson, "la falta de ciberseguridad debería considerarse un riesgo significativo para la IA y para la economía digital, en especial en el internet de las cosas. Si no podemos confiar en los dispositivos inteligentes, o ni si quiera en los dispositivos conectados, no deberían ser bienvenidos en casa ni en los lugares de trabajo" (2019: 149). Las investigaciones en IA deben preocuparse por anticipar en la medida de lo posible qué peligros probables podría generar un mal funcionamiento del sistema que diseñan y qué medidas serían efectivas en tal caso para evitar las consecuencias. Es obvio que esto no puede hacerse más que hasta un cierto punto, pero al menos ese grado básico debe ser requerido.
2. *Conocimiento suficiente de los procesos y de los resultados*. En el desarrollo de sistemas de IA deben primar en todo momento la transparencia y la información. En la situación ideal habría que exigir transparencia en los procesos que llevan a las decisiones que tomen las máquinas. Los sistemas de IA no deberían ser cajas negras, cuyos errores solo puedan ser corregidos *a posteriori*, y si lo son, no deberían ponerse en sus manos decisiones fundamentales sobre la vida de los seres humanos, ni permitir que la responsabilidad de los resultados sea eludida. Es imprescindible abrir esas cajas negras en la medida de lo posible y buscar en el interior los posibles fallos en las decisiones. No se trata de que debamos conocer exactamente todos los procesos computacionales por los cuales estos sistemas han llegado a una decisión, del mismo modo que no necesitamos conocer los procesos neurofisiológicos subyacentes a una decisión humana, pero al menos deberíamos poder reconstruir de forma inferencial el proceso total, de forma que podamos precisar las razones de un resultado defectuoso o inaceptable.
3. *Enfoque axiológico*. Las prioridades axiológicas deben estar bien establecidas, teniendo en cuenta los intereses de los ciudadanos y, en especial, de los que pueden resultar más perjudicados. Obviamente, no es cuestión de establecer mediante consulta popular una lista de deseos, sino de incorporar a los programas políticos los fines concretos que se buscarán mediante el desarrollo tecnológico y las prioridades en su financiación. Los valores no pueden ser impuestos por el mero éxito económico obtenido con ese desarrollo tecnológico. Los objetivos perseguidos por las

máquinas deben estar bien precisados y delimitados, y deben servir al bien común, evitando y corrigiendo sesgos en la consecución de los mismos.

4. *Posibilidad de incorporar en las decisiones de las máquinas los objetivos y valores humanos*. La incorporación de objetivos y valores humanos en las decisiones de las máquinas es fundamental, si es que queremos que estas decisiones sean beneficiosas, aunque hay enormes dificultades para conseguirlo. ¿Cómo hacerlo de forma que no se generen resultados paradójicos o contraproducentes (conseguir, por ejemplo, resolver el problema del calentamiento global podría llevar a las máquinas a exterminar a la mayoría de los seres humanos)? ¿Qué valores incorporar si hay discrepancia entre los propios humanos? ¿Los de qué tradición moral o cultural? ¿Debería poder incorporarse cualquier tipo de valores, incluyendo los de asesinos y terroristas? ¿Cuánto tiempo mantendrían las máquinas esos valores y objetivos antes de asumir los suyos propios? Una propuesta habitual es que ellas mismas vayan aprendiéndolos observando nuestras elecciones (Bostrom, 2014; Russell, 2019). Pero esto exigiría que tuvieran sensibilidad al contexto social, cultural, religioso, etc., lo cual parece bastante difícil por el momento. La incorporación de valores humanos, entre otras cosas, debería llevar a la incapacidad de las máquinas para prestarse a la explotación de seres humanos, al abuso, la discriminación o el daño explícito a otras personas. Pese a las dificultades señaladas, este desiderátum es lo suficientemente importante como para ser recogido en uno de los documentos sobre directrices éticas para la IA preparados por la Unión Europea. Según el documento, la IA debería ser desarrollada desde un enfoque centrado en el ser humano y en sus valores. Así lo expresa el mencionado documento:

> Una IA con un enfoque centrado en la persona se esfuerza por asegurar que los valores humanos ocupen un lugar central en el desarrollo, despliegue, utilización y supervisión de los sistemas de IA, garantizando el respeto de los derechos fundamentales, incluidos los recogidos en los Tratados de la Unión Europea y en la Carta de los Derechos Fundamentales de la Unión Europea; todos ellos constituyen una referencia unitaria a un fundamento común arraigado en el respeto de la dignidad humana, en el que el ser humano disfruta de una condición moral única e inalienable. Esto requiere asimismo tener en cuenta el entorno natural y el resto de seres vivos que forman parte del ecosis-

> tema humano, así como un enfoque sostenible que permita la prosperidad de las generaciones futuras. (Grupo de expertos de alto nivel sobre inteligencia artificial, 2019b: 49)

Una sugerencia interesante a este respecto es la que ha realizado Carissa Véliz (2019): sean cuales sean las normas éticas concretas que haya de establecer cualquier normativa futura, hay un principio que parece central, y es, al igual que sucede en la bioética, el respeto por la autonomía de los seres humanos.

5. *Necesidad de introducir en el diseño mecanismos que permitan limitar el control que las máquinas puedan ejercer sobre los seres humanos*. La decisión última debería poder estar siempre en manos humanas. Esto es una consecuencia del principio anterior de respeto por la autonomía de las personas.
6. *Capacidad para intervenir en la agenda investigadora y participación en el diseño*. Como apunta el filósofo de la ciencia Philip Kitcher (2001 y 2011), en una ciencia "bien ordenada" la determinación de la agenda investigadora debe contar con la opinión y los intereses de los expertos, pero también, y fundamentalmente, de los ciudadanos. Debe favorecerse el debate cívico sobre los objetivos prioritarios en dicha agenda. Esto implica la necesidad de una evaluación previa en relación con los posibles efectos de los nuevos sistemas de IA y la protección de derechos de las personas que puedan verse afectadas por los efectos negativos. Una posibilidad de llevar a la realidad en cierto grado este ideal sería establecer controles éticos antes de la publicación de los artículos. Algunas asociaciones ya lo han hecho con respecto a ponencias presentadas en congresos (Hutson, 2021). También sería útil el análisis de impactos medioambientales de los desarrollos en IA.
7. Órganos regulatorios ágiles y con c*apacidad conminatoria*. Estos órganos institucionales deberían promover legislaciones efectivas. Debe existir la posibilidad legal de exigir responsabilidades ante los tribunales a los fabricantes y programadores. Que las máquinas tomen sus decisiones de forma autónoma no los exime por completo de responsabilidad. Sin embargo, en la actualidad, la defensa a ultranza de las compañías tecnológicas de mantener control exclusivo sobre su propia producción

hace muy complejo cualquier control externo. Estas compañías ejercen ya presiones en el campo de la ética de la IA, especialmente en la que estas mismas empresas desarrollan con el objetivo inconfesado de lavar la propia imagen. Como ha mostrado un estudio (Abdalla y Abdalla, 2020), más de la mitad de los investigadores con financiación conocida en ética de la IA de cuatro grandes universidades norteamericanas habían aceptado dinero de grandes compañías tecnológicas, lo cual puede introducir sesgos claros en los resultados de su trabajo. Tal como explican sus autores:

> Una forma importante en la que las grandes compañías tecnológicas consiguen influir sobre los que se dedican a la investigación en ética de la IA es actuando como falsos donantes de becas o subvenciones. Esto es, al proporcionar una gran cantidad de dinero a los investigadores, las grandes compañías tecnológicas pueden decidir qué se investigará y qué no. Mostramos que la mayoría (58%) de los profesores de ética de IA buscan dinero de esas compañías. Esto significa que estas pueden influir en lo que trabajan. Esto se debe a que, para conseguir esa financiación para la investigación, los académicos se verán presionados para modificar su trabajo de modo que sea más receptivo a los puntos de vista de las grandes compañías tecnológicas. Esta influencia puede ocurrir incluso sin la intención explícita de manipulación, si quienes solicitan premios y aquellos que deciden quién merece financiación no comparten los mismos puntos de vista subyacentes sobre qué es lo ético o cómo "debería resolverse". (Abdalla y Abdalla, 2020: 5-6)

8. *Educación de los ciudadanos*. Es imprescindible promover la educación tecnológica de los ciudadanos. Estos deben alcanzar, a través del sistema educativo, conocimientos básicos suficientes sobre el uso de las nuevas tecnologías, particularmente sobre la IA. Solo así podrán hacer un uso crítico y adecuado de estas tecnologías.

Conclusiones

La discusión sobre el futuro y la gobernanza de la IA, que tanta atención atrae en los medios de comunicación, puede estar siendo desviada de los problemas reales que su desarrollo suscita en el momento presente debido al énfasis puesto en la creación de una IAG y en las posibilidades más peligrosas que esta creación abriría. Prestamos demasiada atención a los problemas que plantea el transhumanismo y a la posibilidad de que máquinas superinteligentes acaben tomando el control y tendemos a perder de vista lo esencial. Los peligros no están ahí, o al menos no lo estarán por mucho tiempo, sino en cosas mucho más perentorias, como la creciente presión para ceder nuestra privacidad y nuestro poder de decisión, no tanto a las máquinas como a los propietarios de las máquinas. No solo los bancos, las instituciones financieras, las empresas, los hospitales, los tribunales de justicia, sino incluso los gobiernos nacionales confían cada vez más en las decisiones tomadas por algoritmos (Tsamados *et al*., 2021). Es necesario un control democrático institucionalmente establecido de la calidad de esos algoritmos y sistemas, así como del impacto de esas decisiones. No se trata de dejar de pensar por completo sobre las posibilidades extremas, sino de darnos cuenta de qué es ahora lo importante. Muchas personas podrían estar dispuestas a renunciar a esa capacidad de control mientras se mantenga el espectáculo de la tecnología a pleno rendimiento, con sus innumerables distracciones y satisfacciones a corto plazo, pero lo reconozcan o no, solo si mantenemos la gobernanza de la tecnología tendremos alguna posibilidad de prevenir el cumplimiento de las distopías anunciadas. Por tanto, esto es algo que debe incorporarse en la agenda de todos los partidos políticos. En la regulación correcta de la IA nos jugamos el futuro.

Referencias

Abdalla, M. y Abdalla, M. (2020). "The Grey Hoodie Project: Big tobacco, Big tech, and the threat on academic integrity". URL = <https://arxiv.org/pdf/2009.13676.pdf> (consultado el 1º de marzo de 2021).

Agar, N. (2016). "Don't worry about superintelligence". *Journal of Evolution and Technology, 26*(1): 73-82.

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

Braga, A. y Logan, R. K. (2021). "The singularity hoax: Why computers will never be more intelligent than humans". En Hofkirchner, W. y Kreowski, H.-J. (eds.), *Transhumanism: The Proper Guide to a Posthuman Condition or a Dangerous Idea?* (pp. 133-140). Cham: Springer.

Bryson, J. J. (2019). "La última década y el futuro de la IA en la sociedad". En VV.AA., ¿Hacia una nueva Ilustración? Una década trascendente (pp. 127-159). Madrid: BBVA-Turner Libros.

Chalmers, D. (2010). "The singularity: A philosophical analysis". *Journal of Consciousness Studies, 17*: 7-65.

Coeckelbergh, M. (2020). *AI Ethics*. Cambridge: The MIT Press.

Dafoe, A. (2018). *AI Governance: A Research Agenda*. Oxford: Centre for the Governance of AI, Future of Humanity Institute, University of Oxford. URL = <https://www.fhi.ox.ac.uk/wp-content/uploads/GovAI-Agenda.pdf> (consultado el 17 de marzo de 2021).

Floridi, L. (2014). *The Forth Revolution*. Oxford: Oxford University Press.

Goertzel, B. y Pennachin, C. (eds.) (2007). *Artificial General Intelligence*. Berlin: Springer.

Gottfredson, L. S. (1997). "Mainstream science on intelligence: An editorial with 52 signatories, history, and bibliography". *Intelligence, 24*(1): 13-23.

Grupo de expertos de alto nivel sobre inteligencia artificial (2019a). "Una definición de la inteligencia artificial: Principales capacidades y disciplinas científicas". Bruselas: Comisión Europea. URL = <https://ec.europa.eu/digital-single-market/en/news/ethics-guidelines-trustworthy-ai> (consultado el 18 de marzo de 2021).

Grupo de expertos de alto nivel sobre inteligencia artificial (2019b). "Directrices éticas para una IA fiable". Bruselas: Comisión Europea. URL = <https://ec.europa.eu/digital-single-market/en/news/ethics-guidelines-trustworthy-ai> (consultado el 18 de marzo de 2021).

Heaven, D. (2020). "Historia, mitos, retos y amenazas de la inteligencia artificial general". *MIT Technology Review*. URL = <https://www.technologyreview.es/s/12728/historia-mitos-retos-y-amenazas-de-la-inteligencia-artificial-general> (consultado el 18 de febrero de 2021).

Hutson, M. (2021). "Who should stop unethical A.I.?" *The New Yorker*, February 15. URL = <https://www.newyorker.com/tech/annals-of-technology/who-should-stop-unethical-ai> (consultado el 3 de marzo de 2021).

Julia, L. (2019). *L'intelligence artificialle n'existe pas*. Paris: Éditions First.

Kitcher, P. (2001). *Science, Truth and Democracy*. Oxford: Oxford University Press.

Kitcher, P. (2011). *Science in a Democratic Society*. Amherst: Prometheus Books.

Larson, E. J. (2021). *The Myth of Artificial Intelligence: Why Computers Can't Think the Way We Do*. Cambridge: The Belknap Press.

Marcos, A. (2021). "Sistemas de control delegado (CoDe)". (Ponencia) Seminario *Huella digital: ¿Servidumbre o servicio?* Madrid: Fundación Pablo VI. URL = <http://www.fyl.uva.es/~wfilosof/webMarcos/textos/textos2021/IA_CoDe.pdf> (consultado el 21 de febrero de 2021).

Marcus, G. y Davis, E. (2019). *Rebooting AI. Building Artificial Intelligence We Can Trust*. New York: Pantheon Books.

O'Neil, C. (2016). *Weapons of Math Destruction*. London: Random House.

OECD (2020). *Recommendation of the Council on Artificial Intelligence*. OECD/LEGAL/0449. URL = <https://legalinstruments.oecd.org/en/instruments/OECD-LEGAL-0449> (consultado el 18 de marzo de 2021).

Pettit, P. (2008). "Three conceptions of democratic control". *Constellations, 15*(1): 46-55.

Russell, S. (2019). *Human Compatible*. London: Allen Line.

Sánchez-Migallón Jiménez, S. (2021). *Redes neuronales profundas. Consecuencias filosóficas*. (Tesis de maestría). Universidad de Granada. URL = <https://www.researchgate.net/publication/348734765_REDES_NEURONALES_PROFUNDAS_CONSECUENCIAS_FILOSOFICAS> (consultado el 17 de marzo de 2021).

Schwab, K. (2016). *La cuarta revolución industrial*. Barcelona: Debate.

Signorelli, C. M. (2019). "Can computers become conscious and overcome humans?" En Chella, A. Cangelosi, A., Metta, G. y Bringsjord, S. (eds.), *Consciousness in Humanoid Robots*. (pp. 197-216). Lausanne: Frontiers Media.

Tegmark, M. (2017). *Life 3.0. Being Human in the Age of Artificial Intelligence*. London: Allen Line.

The Public Voice (2018). *Universal Guidelines for Artificial Intelligence*. Brussels: Electronic Privacy Information Center. URL = <https://thepublicvoice.org/AI-universal-guidelines/> (consultado el 17 de marzo de 2021).

Tsamados, A., Aggarwal, N., Cowls, J., Morley, J., Roberts, H., Taddeo, M. y Floridi, L. (2021). "The ethics of algorithms: Key problems and Solutions", *AI & Society*, online first. DOI: 10.1007/s00146-021-01154-8 (consultado el 3 de marzo de 2021).

Véliz, C. (2019). "Three things digital ethics can learn from medical ethics". *Nature Electronics, 2*(8): 316-318.

Véliz, C. (2020). *Privacy is Power*. London: Bantam Press.

Voss, P. (2007). "Essentials of general intelligence: The direct path to artificial general intelligence". En Goertzel, B. y Pennachin, C. (eds.), *Artificial General Intelligence. Cognitive Technologies* (pp. 131-157). Cham: Springer.

Wang, P. (2019). "On defining artificial intelligence". *Journal of Artificial General Intelligence, 310*(2): 1-37.

Wiener, N. (1964). *God and Golem, Inc.* Cambridge: The MIT Press.