TRABAJO FIN DE MÁSTER

MÁSTER INTERUNIVERSITARIO EN INVESTIGACIÓN EN FILOSOFÍA

¿ESTÁ EL FUTURO AMENAZADO POR LA TECNOLOGÍA?

UNIVERSIDAD DE LA LAGUNA

2020-2021

Alumno: Yeltsin Amaro

Tutor: José Manuel de Cózar Escalante

ÍNDICE

INTRODUCCIÓN

La trayectoria de la civilización humana se dirige hacia un mundo cada vez más hipertecnológico. Dada tal macrotendencia tecnológica, la tarea intelectual es tratar de anticipar lo que podría terminar ocurriendo a largo plazo. La pregunta que titula el presente trabajo es una versión más breve de aquella que tratamos de responder: ¿está el futuro de la vida inteligente amenazado, o incluso condenado, por el avance tecnológico? Identificamos tres posibles respuestas a la pregunta: (1) el avance tecnológico no amenaza el futuro de la vida inteligente; (2) el avance tecnológico está amenazando, o incluso condenando, el futuro de la vida inteligente, pero todavía queda margen de acción para evitar el apocalipsis de alta tecnología; (3) el futuro de la vida inteligente está irremediablemente condenado por el avance tecnológico.

La exposición, estructurada en tres partes, permitirá determinar una respuesta provisional. En la primera parte realizamos una introducción a los estudios del *riesgo existencial*, enfocados en la investigación sobre los posibles factores que podrían destruir de forma permanente el futuro de la vida inteligente, entre los que se incluye la tecnología avanzada. También identificamos algunos malentendidos sobre lo que significan los riesgos existenciales, con la finalidad de evitarlos, y respondemos a algunas de las críticas que dicha noción puede recibir. Dado que se mencionan diferentes amenazas tecnológicas potenciales, la respuesta a nuestra pregunta no puede ser la primera opción. El análisis de dos casos de amenazas tecnológicas potenciales, abordados en la segunda y la tercera parte respectivamente, será la base para determinar alguna de las dos posibles respuestas restantes. El primer caso es *el problema de las bolas negras tecnológicas* y el segundo es *el problema del control de la superinteligencia artificial*, ambos identificados por el filósofo Nick Bostrom, quien también acuñó el término riesgo existencial. El primero de los problemas forma parte de su *hipótesis del mundo vulnerable,* la cual será analizada. Se mencionan algunas soluciones posibles a dichos problemas y se mencionan también otros problemas relacionados. Terminamos concluyendo (provisionalmente) la segunda respuesta: el avance tecnológico está condenando el futuro de la vida inteligente, pero todavía podríamos evitar el peor de los desenlaces futuros: la extinción. También mencionamos otras cuestiones relacionadas.

I. RIESGO EXISTENCIAL TECNOLÓGICO

I.1. Aproximación al riesgo existencial

De acuerdo con la definición de Nick Bostrom, un *riesgo existencial* es aquel que amenaza con extinguir la vida inteligente o con destruir su futuro desarrollo potencial de forma permanente. La clave de que un riesgo sea existencial es que sería un evento que, además de perjudicar a la generación que lo experimente, perjudica para siempre el futuro de la vida inteligente (Bostrom 2013, p. 15; 2002). La noción de riesgo existencial también abarca escenarios en los que la vida inteligente continúa, pero bajo unas condiciones que le impedirían para siempre seguir avanzando, por lo que el término no es sinónimo de extinción. Mientras que la noción de riesgo existencial se refiere a una amenaza de enormes consecuencias, la noción de c*atástrofe existencial* se refiere a la consumación fáctica de tal amenaza, ya sea en forma de extinción o mediante una continuación fallida, como un colapso irrecuperable o una distopía perpetua totalitaria (Ord 2020, cap. 2).

Hay diferentes causas potenciales capaces de provocar catástrofes existenciales. Se distingue entre riesgos existenciales *antropogénicos* y *no antropogénicos*, o derivados de acciones humanas y fenómenos naturales (terrestres: terremotos, supervolcanes, cambios climáticos, pandemias naturales, etc; cósmicos: impacto de asteroides, explosiones de rayos gamma, tormentas solares, etc). Uno de los consensos básicos entre los investigadores del riesgo existencial es que la mayor amenaza para nuestra supervivencia a corto plazo no procede de los eventos naturales, sino de los antropogénicos. Hay principalmente dos razones básicas que fundamentan dicho consenso. La primera se refiere a la historia evolutiva de nuestra especie y la segunda al avance tecnológico del siglo XX: (1) la especie *Homo sapiens* ha sobrevivido a las catástrofes naturales durante centenares de miles de años, así que parece improbable que alguna de ellas desencadene la extinción de nuestra especie en una franja de tiempo tan breve como a lo largo del presente siglo (Bostrom 2013, p. 15; 2002, p. 20). Esta perspectiva, sin embargo, no se aplica a los riesgos existenciales antropogénicos; (2) *Homo sapiens* ha vivido sin tecnología

avanzada hasta hace muy poco tiempo. La invención de tecnología avanzada es algo característico de la historia reciente en comparación con el pasado. Los riesgos existenciales procedentes de la tecnología son riesgos que comenzaron a ser posibles, por primera vez, en el siglo XX. El momento *inaugural* de este nuevo periodo en la historia de la humanidad es la detonación de la bomba atómica (Bostrom 2002, p. 2), el momento en el que las amenazas internas de la humanidad sobrepasaron a las amenazas externas de origen natural, el momento en el que empezamos al alcanzar la capacidad tecnológica para autodestruirnos (Ord 2020). Desde entonces, los riesgos existenciales antropogénicos han ido en aumento y se están convirtiendo, para sus estudiosos, en riesgos que incluso podrían provocar la extinción humana durante el presente siglo. A diferencia de lo que sucede con las amenazas naturales, no tenemos una larga experiencia de supervivencia con el manejo de tecnología avanzada (Bostrom 2013). Los riesgos existenciales tecnológicos podrían manifestarse, además, como resultado de acciones humanas deliberadas o de forma accidental.

Además de las bombas atómicas, entre los riesgos existenciales antropogénicos ya identificados se incluye el cambio climático y el daño medioambiental. Pero si bien dichos eventos ya se han manifestado o se están manifestando, de momento no se han convertido en catástrofes existenciales. Hay otros *riesgos antropogénicos futuros identificados*, siendo los principales la biología sintética (virus artificiales, armas biológicas), la superinteligencia artificial no alineada con la humanidad, la nanotecnología avanzada ("nanoarmas", "plaga gris"), los experimentos en física con efectos imprevistos (agujeros negros de laboratorio, "materia extraña", etc), así como distopías tecnológicas. Se estima que los mayores riesgos posiblemente vengan de las tecnologías futuras más que de las actualmente existentes (Ord 2020, cap. 4 y 6). Entre los riesgos existenciales antropogénicos destacados, Anders Sandberg estima que los tres más probables son la guerra nuclear, la pandemia artificial y la inteligencia artificial no alineada, mientras que Nick Bostrom y Toby Ord sostienen que la mayor amenaza para este siglo (más que el cambio climático y la guerra nuclear) reside en la inteligencia artificial no alineada (Ord 2020, cap. 6). Dicho riesgo existencial se analizará en la tercera parte del presente trabajo.

Por otro lado, los estudiosos del riesgo existencial sospechan que los peligros tecnológicos potenciales no se agotan con los ya mencionados. Así que no solo habría que esperar que los

mayores riesgos procedan de las *tecnologías emergentes*, sino sobre todo de *tecnologías desconocidas* (Bostrom 2002, pp. 0-2). Posiblemente, la mayoría de las tecnologías peligrosas continúan todavía sin identificarse (esta posibilidad se explora en la segunda parte del presente trabajo mediante el análisis de la *hipótesis del mundo vulnerable*, propuesta por Nick Bostrom). Los riesgos existenciales tecnológicos todavía desconocidos que eventualmente surgirían del avance tecnológico durante el presente siglo se sumarían a los actuales problemas colectivos, como el cambio climático, la pandemia y las armas nucleares, entre otros muchos. Tal convergencia, de no resolverse pronto, parece ser solo el comienzo de una convergencia cada vez más peligrosa para nuestro futuro, que incluiría todo el espectro de tecnologías avanzadas con consecuencias potencialmente impredecibles.

La noción de riesgo existencial contribuye a sistematizar las causas potenciales del fin de la vida inteligente o de su ruina irreversible. La gran mayoría de las especies que han pasado por la Tierra se han extinguido y *Homo sapiens* está provocando la extinción de otras muchas en la actualidad. El gran interrogante es si esta situación continuará en alza hasta culminar en un evento insólito: que el ser humano desencadene su propia extinción a través de su avance tecnológico. Los estudiosos del riesgo existencial no dudan de que tal evento puede terminar ocurriendo si no se toman medidas para evitar la creciente convergencia de riesgos existenciales tecnológicos. Quizá la pregunta pertinente que deberíamos hacernos no es cuántos siglos más sobreviviremos, sino si sobreviviremos al siglo XXI. El presente siglo es decisivo para el futuro de la humanidad. La tarea pendiente no es otra que aprender a garantizar la supervivencia de nuestra especie a medida que nos dirigimos hacia el futuro, un territorio poblado de minas antihumanidad.

I.2. ¿Entendemos los riesgos existenciales? Algunos malentendidos

La existencia de los riesgos existenciales tecnológicos es en sí misma controvertida. Para sus estudiosos, no hay duda de que existen potencialmente, pero eso no significa que el resto del mundo también deba creer lo mismo. ¿Son reales los riesgos existenciales tecnológicos o solo son imaginaciones o exageraciones? ¿Son realmente posibles? Podría interpretarse que los investigadores del riesgo existencial tienen como objetos de estudio riesgos que solo por convicciones personales creen que son reales, aunque quizá nunca lo serán, o que incluso se inventen riesgos existenciales y argumenten a favor de la posibilidad de tales amenazas exóticas de una forma tan sofisticada que terminan pareciendo fenómenos que tarde o temprano la humanidad padecerá. Ahora bien, suponiendo que ambos casos se dieran, eso solo serviría para cuestionar al investigador, pero no demostraría la imposibilidad misma de los riesgos existenciales tecnológicos. Los investigadores del riesgo existencial tienen el desafío de convencer al resto del mundo de que estos tipos de riesgos son reales y podrían ocurrir durante el presente siglo. Pero la tarea intelectual requerida no es solo la de argumentar buenas razones a favor de esta posibilidad, sino que también será necesario que se presenten buenas razones en contra. La controversia sobre los riesgos existenciales tecnológicos apenas ha comenzado. Quizá esta sea la discusión más crucial del siglo XXI: de su éxito o de su fracaso podría depender si habrá o no algún futuro para la humanidad. Sin embargo, este debate, como cualquier otro, también podría verse afectado por el riesgo de polarización y desinformación.

No es noticia la negación de ciertos acontecimientos históricos o la invención de otras historias que jamás ocurrieron. Tampoco es noticia el negacionismo del cambio climático antropogénico y del Covid-19, ni tampoco lo es la creencia sistemática en teorías de la conspiración. Tal vez los investigadores del riesgos existencial, tras observar el panorama de suspensión epistemológica del ciudadano contemporáneo, no se sorprendan ante las posibles negaciones categóricas de la existencia de sus objetos de estudio. Pues si el ser humano niega lo que ha sucedido o lo que está sucediendo, ¿por qué no habría de ocurrir lo mismo con lo que todavía no ha sucedido? El negacionismo es tan versátil que puede aplicarse a cualquier fenómeno ubicado en cualquier momento temporal. En el caso de que el debate sobre el riesgo existencial tecnológico

eventualmente ganase algún protagonismo considerable entre las opiniones de la ciudadanía, cabría esperar como posible consecuencia el surgimiento de grupos *negacionistas del riesgo existencial.* Algunos podrían negar todas las formas de riesgo existencial (naturales y antropogénicos), o solo los antropogénicos, o algunos de ellos, pero darle credibilidad a otros. Otros ciudadanos, sin embargo, podrían caer en el extremo opuesto, hasta el punto de creer que en todos lados hay riesgos existenciales inevitables a punto de azotar a la humanidad. Los *conspiracionistas del riesgo existencial* podrían acusar a los *negacionistas* de tener una inclinación oculta: si niegan la realidad futura de las catástrofes de alta tecnología es precisamente para frustrar los esfuerzos por evitarlas, movidos por el deseo de que ocurra el desastre lo antes posible. Dada la diversidad humana, podrían darse situaciones conspiranoicas similares y errores de razonamiento diversos cometidos sistemáticamente, como creer que porque los *negacionistas* no tienen buenas razones para respaldar su postura, automáticamente es cierto que existen los riesgos existenciales y que son inevitables, o creer que porque los *conspiracionistas* (que también podrían denominarse *apocalipticistas*) no tienen buenas razones a favor de su postura, resulta entonces verdadero que los riesgos existenciales son imposibles. O cometer contradicciones en los propios fundamentos, como la del conspiracionista, al defender la inevitabilidad del riesgo existencial tecnológico al mismo tiempo que acusa a sus oponentes de impedir que estos riesgos sean evitados.

Mencionar tales situaciones hipotéticas no pretende sugerir que los riesgos existenciales tecnológicos no sean imposibles, como tampoco que todo aquel que defienda la imposibilidad de los riesgos existenciales tecnológicos es un negacionista. El mencionarlas, más bien, es para reparar en que las posturas polarizadas y desinformadas perjudicarían el debate y la investigación sobre la posible existencia de amenazas existenciales. Si se pretende defender la imposibilidad de estos riesgos ha de aportarse buenas razones para ello. De no ser así, podríamos estar negando un peligro que está a la vuelta de la esquina. Y si se pretende defender la inevitabilidad de estos riesgos ha de argumentarse por qué cualquier acción que tomemos para evitarlos fracasará. De no ser así, podríamos creer que no podremos hacer absolutamente nada por evitar una catástrofe que sí es evitable, así que la inacción estaría completamente injustificada.

Al igual que otras ideas para explicar el mundo, la noción de riesgo existencial es también, como hemos sugerido, vulnerable a los malentendidos humanos. Dada la relevancia de esta noción a la hora de pensar el futuro de la humanidad y su supervivencia, se vuelve necesario entenderla con la mayor claridad posible y diferenciándola de lo que serían malinterpretaciones o tergiversaciones cometidas también por ciudadanos que no necesariamente serían *negacionistas* o *conspiracionistas*. A continuación, identificamos algunas confusiones que podrían llegar a ser sistemáticas.

I.2.1. Los riesgos existenciales tecnológicos *son indemostrables*

Alguien podría decir que los riesgos existenciales no son demostrables científicamente, pues si alguno ocurre y nos extinguimos, ya no podríamos demostrarlo. Ocurre que el planteamiento solo funciona si se da por hecho que pueda consumarse un riesgo existencial. Además, está la posibilidad de constatar una catástrofe existencial cuando se trata de una que no extingue a la vida inteligente. Sería como comprobar que estamos en una pandemia, pero para que fuese una catástrofe existencial, de acuerdo con la definición, tendría que ser una peor y de la que no pudiésemos recuperarnos.

I.2.2. Los riesgos existenciales tecnológicos *no sucederán*

A veces se toma el pasado como una guía fiable para representarse el futuro: si algo no ha sucedido ya, entonces no sucederá. Sin embargo, la tecnología avanzada fabricada desde el siglo XX hasta la actualidad no existía en ninguna época anterior, así que era imposible que pudiese ocasionar una catástrofe existencial. Evidentemente, una detonación nuclear en una ciudad no podía ocurrir antes de que se inventara la bomba atómica. Dicho tautológicamente, hay una primera vez para un evento que no se ha dado antes. Así que es un error juzgar que no ocurrirá una catástrofe existencial tecnológica en el futuro tomando como base el pasado. Por el hecho de

que no haya ocurrido en el pasado no significa que no pueda ocurrir en el futuro. Conviene matizar, además, que si durante las últimas décadas no ha sucedido una catástrofe existencial tecnológica no es porque haya sido imposible. De hecho, ya se han dado las condiciones para que sucediera (volveremos a ello en la segunda parte del presente trabajo). A modo de ilustración, están los casos documentados del siglo XX de detecciones erróneas de ataques nucleares que pudieron haber provocado una respuesta y, como resultado, haber desencadenado una guerra nuclear (Rees 2019). También cabe pensar en fallos en la gestión nuclear que pudieron haber alcanzado la pérdida total de contención. Por otra parte, incluso si tomamos el pasado como una guía para el futuro, cabe considerar que vuelvan a producirse nuevas detonaciones nucleares en otras ciudades.

El futuro, además, es un territorio que, como ya se ha dicho, no solo incluirá las tecnologías actualmente existentes, sino también otras muchas todavía por desarrollar. Los estudios del riesgo existencial sugieren que la bomba atómica no es el límite en la capacidad de tecnología destructiva que puede inventar el ser humano, sino solo un anticipo de lo que estaría por venir. La física nuclear es solo una de las ramas de la innovación tecnológica. El reto intelectual consiste en identificar todo el espectro de posibles ramificaciones destructivas de todas las ramas de la innovación tecnológica. Es un reto intelectual enorme solo con las tecnologías actualmente existentes, pero el desafío es incluso mayor debido a que el futuro traerá nuevas tecnologías con nuevos peligros.

I.2.3. Los riesgos existenciales tecnológicos *son profecías autocumplidas*

Un malentendido hipotético podría ser creer que la anticipación de un riesgo existencial tecnológico es una especie de profecía autocumplida. A ello cabe responder que la comunicación de posibles escenarios catastróficos futuros no constituye la causa que desencadena la catástrofe, sino una llamada de atención para tomar medidas que eviten caer en tal trayectoria, algo que quizá resulte más difícil de conseguir si el mundo entero ignorase su propia marcha hipotética hacia el apocalipsis de alta tecnología. Anticipar catástrofes existenciales tecnológicas no es

entonces el mecanismo que las produce, sino la mínima condición necesaria para evitar que sucedan.

I.2.4. Los riesgos existenciales *son inevitables*

Otro malentendido podría ser creer que los investigadores del riesgo existencial están diciendo que las catástrofes son inevitables y que la humanidad se va a extinguir próximamente. Contrariamente a lo que podría parecer, no son meros alarmistas declarando la inevitabilidad de las catástrofes existenciales y la inminencia de una extinción prematura. No están animando a la inacción porque ninguna medida funcionará o insinuando que solo tenemos margen de acción para adaptarnos al desastre inevitable. Lo que sí sugieren es que algún riesgo existencial tecnológico terminará ocurriendo en algún momento (ya sea por falta de protocolos en la gestión de tecnología avanzada o porque se dan las condiciones para la destrucción masiva deliberada) si no se toman medidas para evitarlo.

Debido a que otorgan bastante peso a las acciones humanas, como una fuerza que es capaz de tomar decisiones para asegurar la continuidad de la humanidad, los investigadores del riesgo existencial no consideran, de momento, que estemos totalmente condenados. Por muy pesimista que pueda resultar la perspectiva de un futuro lleno de amenazas tecnológicas contra nuestra existencia, puede decirse que en general los investigadores del riesgo existencial tienen la motivación de advertir para que se pueda actuar a tiempo en lugar de caer en el derrotismo o el fatalismo (otra pandemia durante las próximas décadas no es inevitable, pero podría terminar sucediendo si no se toman medidas que reduzcan las condiciones propicias para su manifestación). En principio, todavía se puede encauzar esa fuerza de acción humana contra los propios peligros producidos por otras acciones humanas (Ord 2020). La tarea no es otra que identificar los peligros potenciales de la tecnología avanzada para poder evitarlos antes de que sea demasiado tarde. Dentro de lo catastrófico, al final hay cierto optimismo.

Ahora bien, podría llegar a haber alguna amenaza a la vista contra la que realmente no pudiésemos hacer nada, pero tendríamos que estar muy seguros de que realmente es así para

resignarnos justificadamente. También habría que tener en cuenta la posibilidad de amenazas inminentes que ningún ser humano vea a tiempo. Un escenario de esa naturaleza sugiere que estaríamos ya condenados sin saberlo. Pero posiblemente se trata de la categoría más especulativa. Cuanto más riesgos se identifiquen, menos serán los que queden en esa categoría.

I.2.5. Los riesgos existenciales tecnológicos *son improbables*

Una razón aducida para no preocuparnos por los riesgos existenciales tecnológicos es la siguiente: el que un evento sea posible no significa que necesariamente ocurrirá. Una catástrofe existencial tecnológica es algo que podría ocurrir, pero es extremadamente improbable que suceda, así que cualquier advertencia al respecto está injustificada. Y qué decir de las medidas preventivas, que serían acciones exageradamente desproporcionadas.

De hecho, los investigadores del riesgo existencial reconocen que sus objetos de estudio, en cierto sentido, son improbables. ¿Pero significa eso que no es necesario dedicarles nuestra atención? ¿Basta confiar en que un riesgo no ocurrirá porque su probabilidad es muy baja?

Una de las formas de entender la probabilidad consiste en la estimación de la frecuencia con la ocurre un evento a partir de un historial de registros sobre ese evento. Como no hay un registro histórico de riesgos existenciales tecnológicos, no es posible asignar probabilidad alguna en ese sentido. Los riesgos existenciales tecnológicos, por definición, son riesgos que nunca han ocurrido. Si alguno hubiese ocurrido, no estaríamos aquí, o nuestra civilización no tendría el aspecto que tiene. Como teóricamente solo suceden una vez, no hay precedentes para la probabilidad. Sin embargo, como señala Bostrom, el que sea difícil estimar la probabilidad de algún riesgo no significa que el riesgo sea despreciable (Bostrom 2013, p. 16). Lo que sugiere la ausencia de probabilidad de este tipo de riesgos no es que no sean posibles, sino que no sabemos cómo de posibles son. Así que descartar un riesgo existencial tecnológico por su ausencia de probabilidad puede ser un grave error. La estimación de frecuencias no se considera un buen indicador a partir del cual juzgar la relevancia de los riesgos existenciales tecnológicos. La situación es que esta noción de probabilidad no funciona con las catástrofes existenciales (Ord

2020, cap. 2), de ahí que sea necesario considerar formas alternativas para estimar la posibilidad de riesgos que nunca han ocurrido (Häggström 2016, p. 194).

Los investigadores del riesgo existencial suelen emplear una noción subjetiva de la probabilidad: la probabilidad entendida como el grado de creencia que se debería asignar de que algún riesgo ocurra a partir de la evidencia disponible (Bostrom 2013, p. 16; Ord 2020, cap. 2). De esta manera, lo que antes parecía improbable, ahora podría resultar ser muy probable.

Se puede objetar que la probabilidad subjetiva no es una metodología perfecta para estimar los riesgos existenciales tecnológicos, ya que el grado de creencia que se debe tener en base a la información disponible puede que no sea el mismo para todos los evaluadores. De cierta información, podrían derivarse diferentes creencias. La subjetividad nos puede llevar a situaciones controvertidas como la siguiente: los mayores riesgos existenciales tecnológicos podrían no ser los mismos para todos los evaluadores. Mientras que los investigadores del riesgo existencial otorgan una alta credibilidad a hipotéticos eventos tecnológicos catastróficos nunca antes dados, algunos detractores no otorgan credibilidad alguna a los mismos, ni siquiera a peligros considerados por los primeros como los más graves (Pinker 2018, cap. 19). Ante esta colosal discrepancia intelectual acerca de la causa potencial de extinción de la humanidad, parece relevante aspirar a identificar, a menos que nos resulte indiferente, qué es lo que más amenaza realmente nuestra supervivencia durante el presente siglo. Si bien no todos los peligros tecnológicos están en igualdad de condiciones de causar una catástrofe existencial, resulta en realidad incierto prever qué es lo que se podría manifestar antes. De hecho, los consensos ya mencionados respecto a cuáles son las mayores amenazas tecnológicas son solo orientativos. Una nueva revisión podría llevar a estimaciones diferentes. Podría finalmente suceder otro evento que no ocupa los primeros puestos, o una catástrofe desencadenada a partir de una tecnología todavía desconocida, quizá derivada de avances inesperados en ciertas áreas de innovación tecnológica.

Pero esta situación de limitación epistemológica, caracterizada entre otras cuestiones por la ausencia de una probabilidad objetiva idealmente igual para todo el mundo (algo que resulta más accesible, aunque solo orientativo también, para los riesgos existenciales naturales), puede mejorar con más investigación. Como no es posible dar la última palabra a unas cifras brutas, la

estimación de los riesgos existenciales tecnológicos seguirá dependiendo principalmente del razonamiento anticipativo de investigadores que se toman muy en serio el futuro de la vida inteligente. Como señala Toby Ord, hay que tomar decisiones sin poder tener probabilidades robustas de los riesgos en cuestión (Ord 2020, cap. 7). Si la regla básica de investigación fuese *solo tomarse en serio la lista de riesgos ya experimentados*, aquellos de los que sí se puede estimar alguna probabilidad, eso significaría permanecer ciegos a todo un espectro de riesgos potenciales que se hallan fuera de la lista. El desafío intelectual no es otro que identificar bajo una situación de incertidumbre los riesgos existenciales tecnológicos del futuro. Una regla básica más adecuada sería: *anticipar posibles eventos de destrucción nunca antes dados y esquivarlos a tiempo*. Se trata de tomarse muy en serio la posibilidad de que quizá los peores eventos que experimentará la humanidad no se sitúan en el pasado, sino en el futuro. En el futuro nos espera lo improbable. Quizá una improbable utopía interplanetaria, quizá una improbable extinción prematura mientras usábamos juguetes de alta tecnología que no comprendíamos.

Los malentendidos destacados quizá solo sean algunas de las formas en las que el ser humano pueda fallar a la hora de comprender lo que significa un riesgo existencial tecnológico. No está de más ir identificando las posibles confusiones sobre una noción que, de momento, no solemos encontrarnos muy a menudo: no experimentamos titulares del tipo "un nuevo riesgo existencial golpea a la Unión Europea", y si alguna vez vemos algo similar, de seguro que ha sido una confusión de niveles de riesgo. Con el tiempo, podría ser que su uso aumente, así como el número de malentendidos. Quizá también veamos surgir nuevos conspiracionistas y negacionistas. Pero no toda la balanza necesariamente ha de caer a ese lado. También podrían emerger nuevos movimientos sociales promoviendo la resolución de futuros riesgos existenciales tecnológicos, como ya sucede con el cambio climático.

I.3. Algunas críticas a la noción de riesgo existencial

I.3.1. La noción de riesgo existencial *es antropocéntrica*

La crítica en cuestión sería que la noción de riesgo existencial se basa en una mayor preocupación por las amenazas tecnológicas que se ciernen sobre la humanidad y no tanto en la amenaza que representa la humanidad para el planeta. La noción pecaría de antropocéntrica al otorgarle más importancia a la supervivencia de la humanidad que a la diversidad biológica y al cuidado del planeta.

Una posible respuesta a la crítica sería que la noción no pretende restarle importancia al daño que el ser humano produce sobre el planeta, algo relativamente conocido, sino más bien dar a conocer el peligro creciente al que se enfrenta la humanidad con sus avances tecnológicos. Nuestra situación actual, como indica Toby Ord, es doble: no solo estaríamos en el Antropoceno, el tiempo en que el ser humano tiene un profundo efecto en el medioambiente, sino que también estaríamos en el Precipicio, es decir, el tiempo en el que la humanidad corre el riesgo de autodestruirse (Ord 2020, cap. 1). Sería un tiempo caracterizado por riesgos emergentes derivados de la tecnología avanzada. Ambas nociones hacen énfasis en la bomba atómica como un momento clave en la historia de la humanidad, pero las implicaciones serían diferentes. Hasta ahora, la noción de Antropoceno (De Cózar 2019) se ha popularizado más que la de riesgo existencial. Incluso podría decirse que todavía permanece relativamente ignorada a pesar de su relevancia para comprender el mundo contemporáneo.

Otra posible respuesta a la crítica nos lleva a la definición de riesgo existencial[1], en la que se hace referencia a la vida inteligente y no a la humanidad. Sería entonces más preciso decir que la noción de riesgo existencial es "intelicéntrica" en lugar de antropocéntrica. Referirse a la vida inteligente incluye a la humanidad, pero también a otras formas de vida, especialmente a las futuras formas de vida inteligente. La humanidad no sería tanto el centro de la vida inteligente, sino más bien una catapulta hacia el futuro de la inteligencia, lo cual apunta a una nueva

1 Un riesgo capaz de extinguir la vida inteligente o de destruir permanentemente su futuro.

acusación que veremos a continuación (a saber, que la noción de riesgo existencial *es transhumanista*).

Si bien dicha crítica tiene su derecho a existir, lo cierto es que el impacto perjudicial de una catástrofe existencial no dependerá de si la noción de riesgo existencial es o no antropocéntrica: no solo perjudicaría a la humanidad, sino también a otras muchas formas de vida y a la Tierra misma. Evitar catástrofes existenciales significa también contribuir a la supervivencia de formas de vida no humanas y al futuro planetario.

I.3.2. La noción de riesgo existencial *es transhumanista*

La extinción de la vida inteligente significaría la cancelación permanente de sus posibilidades futuras, a saber, que desarrolle su vasto potencial tecnológico en el universo. Como donde mejor encontramos ese potencial es en *Homo sapiens* (pero podría haber sido otra especie), garantizar su supervivencia es fundamental para que la vida inteligente pueda seguir avanzando a muy largo plazo. En ausencia de una catástrofe existencial y con el avance tecnológico de la vida inteligente en marcha, eventualmente vendrían a la existencia otras muchas generaciones futuras más inteligentes: transhumanas y posthumanas. Como precisa Toby Ord, "el propósito último es permitir que nuestros descendientes cumplan nuestro potencial" (Ord 2020, cap. 2).

De lo anterior, se podría derivar la siguiente acusación: la noción de riesgo existencial, al basarse en una perspectiva transhumanista, insinúa que evitar la extinción de la humanidad no es un fin en sí mismo, que la motivación última para evitar una catástrofe existencial no es en sí misma la de salvar a la humanidad. La noción transhumanista de riesgo existencial sugiere que hay que proteger la existencia de la vida inteligente, no por lo que es en sí misma en el presente, sino por su potencial futuro. El plan encubierto no es otro que evitar la extinción de *Homo sapiens*, mantenerlo de momento en la existencia, tan solo porque es un medio para posibilitar la evolución hacia una condición posthumana. Al final, el objetivo último para evitar una catástrofe existencial es que se consume en algún momento la culminación del potencial futuro de la vida inteligente: la posthumanidad.

Otra forma de formular la acusación de motivación transhumanista sería: si el ser humano no pudiese evolucionar más hacia una mejor condición, si su condición actual solo pudiese empeorar, entonces la posthumanidad sería imposible. Si esta fuera la situación real, los estudiosos del riesgo existencial, como transhumanistas, no tendrían tanta motivación por evitar una catástrofe existencial.

Una posible respuesta sería la siguiente: la crítica parte de una postura antropocéntrica, pero la perspectiva transhumanista trata de superar el antropocentrismo. El transhumanismo tiene una perspectiva amplia de la futura vida inteligente, y como dijimos anteriormente, la vida humana no es el centro de ella, sino una catapulta hacia ella. Entender la noción del riesgo existencial desde una perspectiva transhumanista es la forma más amplia de concebir dicha noción, pues implica que no solo está en peligro la continuidad de la vida inteligente actual, sino también todas las posibles vidas futuras que la humanidad actual podría llegar a posibilitar si no se extinguiese prematuramente. Evitar una catástrofe existencial salva a la generación presente que es vulnerable a ella, pero también salva (potencialmente) a todas las generaciones futuras, ya sean humanas, transhumanas o posthumanas, así como a otras formas de vida.

Otra respuesta sería que no cualquier interpretación de la noción de riesgo existencial tiene que ser necesariamente transhumanista. Una noción no transhumanista del riesgo existencial es posible, aunque sería más estrecha en su alcance: aquel riesgo que afecta a la vida inteligente actual, excluyendo de la fórmula el futuro desarrollo potencial de la vida inteligente, que ya no sería objeto de preocupación para la toma de decisiones contra los riesgos procedentes de la tecnología. El objetivo de evitar una catástrofe existencial tecnológica sería salvar la vida inteligente actual y no la futura.

I.3.3. La noción de riesgo existencial *está a favor de la existencia del ser humano, y eso es muy malo*

Tanto la perspectiva transhumanista del riesgo existencial, basada en una aparente obsesión con el futuro, como la perspectiva que excluye la importancia de las generaciones futuras, pueden ser

simultáneamente criticadas por suponer "erróneamente" que el riesgo existencial, como amenaza a la existencia de la humanidad, es algo malo en sí mismo. La crítica podría proceder desde cierta perspectiva misántropa: no debería haber generaciones futuras ni presentes, porque la humanidad es un cáncer planetario, así que una catástrofe existencial que solo extinguiese a la humanidad sería algo bueno para el planeta.

A esto cabe responder que es difícil que una catástrofe existencial, de ocurrir, solo perjudique a la humanidad, sin afectar de alguna manera a las formas de vida no humanas y a la Tierra (este evento puede ser catastrófico a corto plazo, pero a largo plazo podría favorecer el surgimiento de otras formas de vida, como la extinción de los dinosaurios, que favoreció a la evolución de los mamíferos). Así que una catástrofe existencial, en principio, no es la mejor manera de salvar la biosfera. El misántropo tendrá que esperar a que la humanidad se extinga en algún momento, y si esto se demora, quizá optaría por usar tecnología avanzada para desencadenar tal extinción.

Por otro lado, si bien es cierto que el ser humano está provocando un cambio climático, la extinción de especies, un gran daño medioambiental y está amenazando su propia existencia con el avance tecnológico, al mismo tiempo es la única especie capaz de desarrollar, eventualmente, la tecnología avanzada requerida para salvar a la Tierra de una catástrofe existencial cósmica, como un asteroide o cualquier otro fenómeno energético exótico, aunque solo sea por puro interés antropocéntrico. El ser humano, además, es un puente potencial para que la vida terrestre florezca más allá de la Tierra. Aunque la extinción humana puede parecer la solución para salvar a la Tierra, a largo plazo es la especie humana (o su descendencia posthumana) la que podría salvar el planeta.

II. EL PROBLEMA DE LAS BOLAS NEGRAS TECNOLÓGICAS

II. 1. Un futuro vulnerable

Los estudios sobre el riesgo existencial tecnológico tratan de identificar los peligros potenciales derivados tanto de las tecnologías avanzadas ya existentes como de las emergentes, aunque también tienen en cuenta otra categoría: las amenazas de las tecnologías todavía desconocidas. Serían aquellas que están fuera del imaginario prospectivo. Como es de esperar, una vez que alguna tecnología es identificada, ya deja de formar parte de la categoría de lo que todavía permanece desconocido. La *hipótesis del mundo vulnerable*, propuesta por Nick Bostrom es (que sepamos) el primer esfuerzo teórico en conceptualizar y sistematizar el estudio de los posibles riesgos tecnológicos de impacto global que permanecen sin identificar. La hipótesis no trata de conocer las tecnologías desconocidas, sino que trata de identificar ciertos aspectos que serían comunes a los peligros tecnológicos desconocidos y sobre los cuales podría haber un margen de acción. Bostrom aporta una hoja de ruta para hacer frente a los mismos, una especie de guía orientativa para inspirar la toma de decisiones al respecto. Aquí analizaremos la hipótesis, lo cual primero nos lleva a detallarla.

El camino hacia la hipótesis parte de una metáfora: la creatividad humana sería como una urna gigante de la que extraemos bolas, que representarían ideas, descubrimientos e invenciones tecnológicas. Hasta ahora, la mayoría de las bolas extraídas han sido beneficiosas, aunque algunas han resultado perjudiciales. Bostrom sostiene que si la investigación científica y el desarrollo tecnológico continua (si seguimos extrayendo bolas de la urna), en algún momento sacaremos una *bola negra*: una tecnología que destruye la civilización por defecto. Sostiene, además, que si hasta la fecha este tipo de extracción todavía no ha ocurrido no es porque hayamos sido o seamos cuidadosos y sabios en nuestra política tecnológica, sino porque hemos sido muy afortunados. En otras palabras, si todavía no nos hemos autodestruido con nuestro potencial tecnológico ha sido por pura suerte. Las bolas negras tecnológicas, entonces, no son algo exclusivo del futuro, sino que serían una gran amenaza de destrucción potencial que llevaría

acompañando a la humanidad desde hace varias décadas. El supuesto de que en la urna de la creatividad *hay al menos una bola negra tecnológica que destruye a la civilización por defecto* es solo una parte de la hipótesis, que Bostrom formula como sigue:

> Si el desarrollo tecnológico continúa entonces en algún momento será alcanzado un conjunto de capacidades que hacen extremadamente probable la devastación de la civilización, a menos que la civilización salga lo suficiente de la condición semi-anárquica por defecto (Bostrom 2019, p. 457).

La "condición semi-anárquica" significa que el orden mundial contemporáneo no está preparado para impedir que la civilización sea destruida. Según Bostrom, la condición del mundo actual se caracteriza por tres limitaciones: (1) *vigilancia preventiva* limitada: los medios de vigilancia estatales no estarían lo suficientemente desarrollados como para garantizar que los ataques terroristas u otros actos ilegales sean eventos imposibles de suceder; (2) *gobernanza global* limitada: estrategias de coordinación insuficientes o falta de competencia para la resolución de problemas globales y riesgos nacionales; (3) dadas las dos limitaciones anteriores, no es posible regular de forma eficiente toda la *diversidad de motivaciones*: hay una gran variedad de intereses (a nivel estatal, poblacional e individual) que motivarían acciones que destruirían la civilización. En particular, Bostrom denomina "residual apocalíptico" al conjunto de actores minoritarios dispuestos a destruir la civilización.

Con su hipótesis, Bostrom no está diciendo que la innovación científica y tecnológica, por sí misma, de forma autónoma, destruirá la civilización por defecto. Lo que sugiere es que son las condiciones actuales en las que se desarrolla la innovación científica y tecnológica las que hacen que la civilización pueda destruirse por defecto: es la "condición semi-anárquica" del mundo contemporáneo, la falta de control de los actores (individuales, estatales, poblacionales) que tienen acceso a la innovación científica y tecnológica o hacen uso de ella. La idea es que se darán las condiciones en las que al menos un actor activará al menos una bola negra tecnológica.

Bostrom observa que en la actual "condición semi-anárquica" se pueden distinguir cuatro tipos de *vulnerabilidades* que abarcan el espectro posible de las hipotéticas tecnologías desconocidas

que destruyen por defecto. Aunque no esté al alcance saber de antemano cuáles serán concretamente las bolas negras altamente destructivas, es posible, mediante los cuatro tipos de vulnerabilidades, concebir qué podrían ser y qué aspecto podrían tener. Esto resulta al alcance debido a que se puede ubicar una bola negra de acuerdo a los actores que puedan hacer uso de ella, mientras que otros no podrían (*diversidad de motivaciones*: estatal, poblacional e individual). Bostrom ilustra cada una de las vulnerabilidades a través de experimentos de pensamiento contrafactuales:

La *vulnerabilidad de tipo 1* se refiere a una tecnología destructiva accesible y fácil de manejar para el residual apocalíptico (Bostrom 2019, p. 458). A modo de ilustración, sugiere que podría ser algo como "armas nucleares fáciles": ¿qué hubiera sucedido si las bombas atómicas hubiesen sido algo fácil de fabricar? ¿Y qué pasaría si el residual apocalíptico descubre a su alcance una forma de fabricar una tecnología altamente destructiva hasta ahora ignorada?

La *vulnerabilidad de tipo 2a* se refiere a una tecnología destructiva accesible solo a actores poderosos que tienen el incentivo de usarla para destruir (Bostrom 2019, p. 460). A modo de ilustración, un "primer golpe asegurado" en lugar de una destrucción mutua asegurada entre estados, ya sea con bombas atómicas, con tecnología militar ultrasecreta, u otra tecnología desconocida. También hace mención a los estudios que señalan que fue por pura suerte que se evitara el holocausto nuclear durante la guerra fría.

La *vulnerabilidad de tipo 2b* consiste en muchos actores motivados a usar una tecnología poco dañina, sin el objetivo de destruir, pero su uso combinado resulta destructivo (Bostrom 2019, p. 461): un "peor calentamiento global" derivado de la combinación de acciones individuales contaminantes de poco impacto. La idea es que podría haber una bola negra que surja de un evento poblacional, como un incentivo económico beneficioso a nivel individual, pero perjudicial para la economía, de tal modo que al final todos salen perdiendo y las consecuencias a largo plazo destruirían la civilización.

La *vulnerabilidad de tipo 0* podría darse a través de las acciones de actores individuales, poblaciones y estatales que no tienen ninguna motivación destructiva: una tecnología con un riesgo desconocido que al ser descubierta destruirá la civilización por defecto de forma involuntaria, accidental (Bostrom 2019, p. 461). Sobre esta última forma de vulnerabilidad,

Bostrom advierte de que un conocimiento defectuoso de un proyecto podría hacer creer (equivocadamente) que los beneficios superan los costes. Por ejemplo, señala que durante el desarrollo del Proyecto Manhattan, uno de los científicos del proyecto, Edward Teller, tuvo la preocupación de que una explosión nuclear prendiera la atmósfera y el océano. Se intensificaron los cálculos y se concluyó que eso no ocurriría, lo cual se corroboró con las primeras detonaciones experimentales. A continuación, destaca que en 1954 el gobierno estadounidense llevó a cabo el "Castle Bravo test", en el que se experimentó con otra bomba termonuclear y se calculó su potencial en unos 4-8 megatones. Al final, la bomba detonó en 15 megatones. Bostrom matiza que un error de cálculo como ese afortunadamente no ocurrió en el primer caso. Si los cálculos que indicaban que la atmósfera no se incendiaría hubiesen sido erróneos, quizá al final sí que podría haberse prendido, lo cual, como asevera Bostrom, habría provocado la destrucción de toda la vida terrestre en 1945. Otros ejemplos serían experimentos en física que transformaran la materia ordinaria en materia extraña (o "strangelets"), provocando así la destrucción del planeta, o que los aceleradores de partículas produjeran agujeros negros en la Tierra (Rees 2019). Bostrom señala que los cálculos descartan que tales eventos ocurran, pero nos recuerda que los cálculos pueden salir mal, o que sean incompletos al pasar por alto variables desconocidas, un riesgo oculto no previsto.

Debido a que las especificaciones de las bolas negras permanecen desconocidas, no es posible saber de antemano qué tipo de medidas de seguridad serán eficaces para el momento en que alguna bola negra sea extraída sin que la estemos buscando. Podría ser que una vez extraída se elaboren a continuación medidas de contención, pero el problema es que podría ser una tecnología en apariencia dócil y que posteriormente se rebele su poder destructor. Además, no se sabe cuánto tiempo llevaría elaborar una medida de seguridad eficaz antes de que sea demasiado tarde, y eso suponiendo que alguna medida de prevención fuese realmente posible. Así que si no tenemos soluciones directas contra las tecnologías altamente destructivas que permanecen desconocidas, ¿cómo se podría solucionar nuestro actual estado de vulnerabilidad mundial?

En principio, sería posible establecer soluciones *indirectas* contra las bolas negras incluso aunque no sepamos cómo serán. Esta estrategia consistiría en tomar medidas sobre los diferentes actores, motivados o no, que puedan usar una bola negra. Dada esa posibilidad para evitar el

apocalipsis tecnológico, Bostrom enfatiza que su hipótesis no implica que la civilización esté condenada a la autodestrucción. La única forma de evitar la destrucción, sin embargo, es salir de la "condición semi-anárquica" que caracteriza a nuestro mundo vulnerable. Bostrom analiza varias macroestrategias para conseguirlo, pero concluye que la mayoría son medidas limitadas y que las únicas dos medidas realmente determinantes para *estabilizar el mundo* (salir de la condición semi-anárquica) son: (1) crear la capacidad de vigilancia preventiva extremadamente efectiva y (2) crear la capacidad de una fuerte gobernanza global (Bostrom 2019, p. 465). La hipótesis adquiere así la siguiente formulación:

> hay cierto nivel de tecnología en el que la civilización casi seguramente se destruye a menos que grados de vigilancia preventiva y/o gobernanza global bastante extraordinarios e históricamente sin precedentes sean implementados (Bostrom 2019, p. 457).

El primero de los mecanismos para estabilizar el mundo implica que los estados desplieguen sobre los ciudadanos un nivel de vigilancia tan intenso que recibe el calificativo de "panóptico de alta tecnología" (Bostrom 2019, p. 465), aunque también señala la posibilidad de ir hacia una trayectoria de transparencia total incluso no habiendo ninguna amenaza inminente: toda la humanidad siendo monitorizada en todo momento mediante dispositivos y sensores biométricos adheridos al cuerpo (pulseras, collares, móviles, etc) que recopilan todo lo que escuchamos, vemos y hacemos, así como otros extendidos en el entorno. Toda la información recopilada sería analizada por algoritmos y humanos de un modo extremadamente eficiente. Esta vigilancia institucionalizada significa que toda la parafernalia tecnología sería de uso obligatorio. Sería ilegal manipular o desacoplar los sensores adheridos al cuerpo. Así las cosas, la posibilidad de que alguien del "residual apocalíptico" (vulnerabilidad de tipo 1) perpetre una destrucción sería nula, dado que automáticamente su conducta sería predecible y se le interceptaría a tiempo.

Pero Bostrom observa que la vigilancia preventiva no ayudaría activamente a estabilizar el mundo de las vulnerabilidades de tipo 2a (primer golpe asegurado), donde la amenaza procede desde fuera de las fronteras. Tampoco sería necesaria para estabilizar el mundo de una vulnerabilidad de tipo 2b (peor calentamiento global), donde muchos actores tienen incentivos

para realizar pequeñas acciones permitidas que en conjunto pueden destruir la civilización. Es respecto a las vulnerabilidades de tipo 1 donde el panóptico de alta tecnología sería realmente eficaz.

Mientras que la vigilancia preventiva requiere que los estados regulen las acciones de sus ciudadanos (múltiples panópticos de alta tecnología), la gobernanza global requiere la regulación de las acciones de los estados. Pero el requisito de gobernanza global no se limita a exigir que los estados existentes empiecen a vigilar y regular las actividades de otros estados. Si bien la gobernanza global comprendería una cooperación coordinada entre estados para resolver problemas globales (como el cambio climático o la pobreza) y para garantizar la seguridad (Bostrom 2019, p. 465), lo cierto es que ese mecanismo va más allá al demandar la necesidad de una nueva institución: un gobierno mundial con la capacidad de imponer su voluntad a cada nación (Bostrom 2019, p. 467). Sería un nuevo orden mundial basado en una institución que controlaría a todos los estados del mundo.

Bostrom señala que las vulnerabilidades de tipo 2a (primer golpe asegurado) solo se pueden solucionar de forma eficaz mediante una gobernanza global. Sin ella, la destrucción de la civilización (ya sea con bombas atómicas, con armas biológicas, con enjambres de drones asesinos u otra especie de bola negra) seguirá siendo un evento de momento a la espera para manifestarse. Bostrom indica que una gobernanza global comprometería a los estados a no fabricar bolas negras y a desmantelar las bombas atómicas. Señala que también sería eficaz contra las vulnerabilidades de tipo 1, sobre todo con aquellas que pueden externalizarse, como una bola negra biotecnológica capaz de provocar una pandemia. Así, todos los estados comprometidos tendrían la obligación de desplegar la vigilancia preventiva para no poner en peligro a otras naciones.

Bostrom concluye que hay vulnerabilidades que se pueden estabilizar solo con una vigilancia preventiva y otras solo con una gobernanza global, pero que la mayoría de las vulnerabilidades (tipo 1 y 2a-b) al final serían estabilizadas con ambas estrategias combinadas, haciéndose más seguro el uso de la urna de la creatividad, es decir, el desarrollo científico y tecnológico (Bostrom 2019, p. 467).

II. 2. ¿Estabilizaremos el mundo?

La hipótesis del mundo vulnerable es una propuesta controvertida. Se confirmaría, en cierto sentido, si eventualmente se extrajera una bola negra y la civilización se destruyera a continuación, o si se establecen vigilancias digitales y un gobierno mundial capaces de imposibilitar cualquier acto de destrucción. La hipótesis sería errónea si no hay bolas negras en la urna de la creatividad, pero el que las haya es un supuesto que resulta difícil de descartar. ¿Es tranquilizador confiar que no existirá algo que, siendo físicamente posible, todavía no existe? Podría parecer que la hipótesis quedaría refutada si extrajésemos una bola negra y al final el mundo no se destruye por defecto, pero entonces no sería, por definición, una bola negra. También podría interpretarse como errónea si, después de establecerse la vigilancia digital y el gobierno mundial, la civilización igualmente se autodestruye con su tecnología, pero se replicaría (si se pudiera) que fue porque las dos medidas no estaban funcionando como deberían. La situación de la hipótesis es que su confirmación es peligrosa y su refutación primero requiere que las dos medidas propuestas se institucionalicen con éxito, algo que, en sí mismo, no parece fácil. La hipótesis es problemática en su formulación, pero habría que repensar si por ello ya pierde realmente toda validez.

Habrá quienes opten por creer que las bolas negras no existirán o no son posibles, por lo que las medidas preventivas serían innecesarias y ahí se acaba el dilema. Sin embargo, tal procedimiento cognitivo significa despachar de forma superficial un problema mucho más complejo de lo que parece. La hipótesis postula que ciertas tecnologías altamente destructivas que hasta ahora no hemos inventado, son posibles, y de continuar el avance tecnológico, en algún momento desconocido serán inventadas. La situación es que no podemos saber que no hay bolas negras, como tampoco nos es posible saber con certeza ahora que en el algún momento futuro se manifestará algunas de ellas. La hipótesis nos invita a abrirnos a la creencia incierta de que deberíamos sospechar que podría haber al menos una bola negra, aunque no tengamos evidencia empírica (y esperemos no tenerla nunca). La situación a la que nos enfrentamos consiste en una amenaza tecnológica cuya existencia no está confirmada, pero que tampoco es viable esperar a tener la confirmación de dicha existencia para fundamentar así la adopción de medidas

estratégicas preventivas. Habría que tomar tales medidas incluso con la incertidumbre de que quizá las bolas negras difusamente imaginadas que inspirasen la actuación no sean al final peligros reales. Existe el riesgo de tomar medidas contra peligros concretos imaginados que no existen, pero las medidas igualmente seguirían siendo válidas para otras amenazas tecnológicas desconocidas. La hipótesis no sugiere que debemos tomar medidas contra cualquier amenaza imaginada y sin justificación alguna, como contra una invasión de demonios fugados del infierno. Lo que más bien sugiere es que el avance tecnológico eventualmente posibilitaría una capacidad de destrucción hasta ahora no experimentada por la civilización humana y, quizá, sin que haya indicios previos de que vaya a ocurrir. En principio, tal sospecha tiene implicaciones de gran calado en la práctica.

Una explicación, fácil de imaginar, de por qué hasta ahora una bola negra no ha destruido la civilización consiste en decir que es porque ya tenemos mecanismos de seguridad eficientes integrados en la gestión de la innovación científica y tecnológica, los cuales son capaces de evitar los desastres. Esto puede ser cierto para ciertos riesgos, pero incluso así terminan sucediendo. Sin embargo, ocurre que los actuales sistemas de seguridad no fueron precisamente diseñados ni para evitar riesgos existenciales identificados ni bolas negras tecnológicas. Es por ello que parece plausible el que, hasta ahora, la civilización haya tenido cierto grado de suerte en no experimentar macrodestrucciones terminales en ausencia de instituciones preparadas contra tales niveles de riesgo. Bostrom llama la atención sobre la necesidad de una política tecnológica para el siglo XXI: los avances tecnológicos nos están obligando a reconfigurar nuestras instituciones, que no parecen ser lo bastante robustas frente a los peligros de destrucción tecnológica que están empezando a ser posibles. De momento, esas nuevas instituciones robustas no existen. Una vez que se reconoce la posible amenaza de las bolas negras tecnológicas, la controversia se sitúa en las soluciones sugeridas por Bostrom contra los actores que puedan hacer uso de las mismas.

Aunque primero podríamos preguntarnos: ¿los diferentes tipos de actores tienen la misma probabilidad de destruir la civilización? ¿Cuál es el actor más peligroso? Plantearse esto permite priorizar en el plano teórico el tipo de actor que más urge regular y el tipo de medidas específicas a diseñar. Entonces ¿por qué Bostrom concluye que las únicas soluciones realmente eficaces

contra las bolas negras son los panópticos de alta tecnología y la gobernanza mundial? Una posible respuesta sería que, incluso aunque hubiese algún consenso teórico sobre qué actor priorizar en la práctica, continuaría latente el riesgo de que la primera bola negra al final venga de los otros actores menos priorizados. Dicho de otra manera, aunque se priorice un tipo de actor para regularlo primero, quedaría por regular otros tipos de actores posteriormente. Ese intervalo temporal podría ser suficiente para que los actores todavía por regular destruyan la civilización, ya sea por motivación, por accidente o colectivamente de forma involuntaria. Al final, las dos medidas preventivas propuestas por Bostrom son una especie de navaja multiusos teórica, unas soluciones generales orientativas. En teoría, permitirían regular a todos los actores al mismo tiempo y así salir directamente de la condición semi-anárquica que nos lleva a la destrucción de la civilización.

Sin embargo, lo que sucede con las dos medidas preventivas es que no gustan a todo el mundo (vigilancia digital, un gobierno mundial, etc) y resultan difíciles de alcanzar del modo en que se requiere. Idealmente, una alternativa sería proponer otras soluciones (1) al menos tan solventes como las de Bostrom, (2) que gusten un poco más y que (3) sean más fáciles de alcanzar. Desconocemos si esto ya ha sido resuelto (suponiendo que pueda ser resuelto). En ausencia de lo que podríamos denominar la *alternativa perfecta*, queda tomar la propuesta de Bostrom como un punto de partida. ¿Cuál de las dos medidas es más fácil de materializar primero? ¿Los panópticos estatales o el gobierno mundial? Cada una representa por sí misma muchos desafíos sociales, y las dos juntas suponen un desafío mayor. Si solo se acepta una de las dos medidas preventivas, el problema de la vulnerabilidad no se resuelve completamente (en teoría, ya que como veremos, las dos medidas en la práctica quizá tampoco lo consigan). En ese caso, habría que complementar la medida elegida con otras medidas alternativas al menos tan solventes como la descartada, que gusten un poco más y que sean más fáciles de alcanzar. Si la nueva propuesta no reúne estas condiciones, nada la hace más preferible a aquella medida de Bostrom que se ha descartado. Otra posibilidad es aceptar que lo mejor que tenemos contra las bolas negras y la diversidad de actores son las dos medidas preventivas de Bostrom, pero que no parezcan suficientes. Habría que especificar cuáles serían el resto de medidas suficientes. Esta vía es incluso más difícil de alcanzar.

La hipótesis del mundo vulnerable es un modelo del mundo contemporáneo que intenta contribuir a la toma de decisiones de impacto global. Aunque como sucede con cualquier otro modelo, tampoco es una copia exacta de aquello que representa. Llevando el modelo a la práctica, es probable que las soluciones generales no se apliquen por igual en todos los países. Cada país presenta sus propias circunstancias, las cuales afectarían al alcance de las medidas y la toma de decisiones. Unos países son mas avanzados tecnológicamente que otros y, en principio, presentan unas condiciones políticas y sociales mas favorables para las medidas preventivas. Y si bien una bola negra en principio tendría un impacto global, las condiciones para su manifestación no son las mismas en todos los países. Ilustrándolo en exceso, digamos que no hay desarrollo de alta tecnología en las tribus del mundo, ¿pero habría que vigilarlas en la misma proporción que a los innovadores de Silicon Valley? La hipótesis propone la aplicación de soluciones abstractas para un mundo muy diverso. Sería como dar por supuesto que todos los países contaminan en la misma proporción y deben contribuir por igual a la mitigación del cambio climático. También cabe imaginar todo el posible espectro de obstáculos durante las negociaciones para un futuro gobierno mundial (demasiados intereses nacionales en conflicto, invasiones militares en caso de no cooperación, quiénes serían los líderes, etc).

No obstante, con independencia de las dificultades prácticas para alcanzar el estadio requerido, cabe preguntarse qué podría suceder si idealmente suponemos que las dos soluciones generales se desplegaran de forma eficiente por todo el mundo. ¿Cuáles serían las implicaciones de múltiples panópticos digitales y un gobierno mundial? ¿Cómo sería ese mundo post-semianárquico y estabilizado? En teoría, dejaríamos de ser vulnerables a las bolas negras, aunque bien podríamos pasar a ser vulnerables a los efectos negativos del nuevo mundo. ¿Habríamos escapado de una distopía para caer en otra? Ahí reside el dilema general.

Para Bostrom, el desplegar las dos medidas se justifica porque nos protegerían del amplio abanico de vulnerabilidades que, por defecto, nos arrojan a la autodestrucción. Bostrom defiende ese camino estratégico en nombre de la salvación, incluso a pesar de las posibles consecuencias negativas. De hecho, Bostrom menciona algunos de los efectos indeseados de las medidas instrumentales para un fin mayor. Así, nos recuerda que la vigilancia preventiva ayudaría a los regímenes autoritarios a protegerse de la rebelión de la ciudadanía empoderada o contribuiría a la

imposición de una ideología intolerante con otras visiones de la vida. El gobierno mundial podría caer en las manos menos propicias y podría convertirse en un régimen movido por el autointerés, que redujera la diversidad y anulase la posibilidad de otras alternativas políticas que podrían ser mucho mejores (Bostrom 2019, pp. 467-468). Pero Bostrom no cree que las derivas totalitarias sean inevitables, sino solo un riesgo posible. Propone medidas de regulación de la vigilancia preventiva y de la gobernanza global, como un sistema basado en la transparencia y mecanismos que eviten la acumulación de poder. En definitiva, su propuesta apoya explícitamente la vigilancia preventiva y el gobierno mundial, que pueden obedecer a varias modalidades políticas, pero no dice que debamos fundar totalitarismos. Lo que sostiene es que vale la pena aceptar el riesgo de totalitarismo si lo que ganamos es la capacidad de estabilizar el mundo de las vulnerabilidades que destruyen por defecto (Bostrom 2019, p. 470). Por otro lado, Bostrom también menciona ciertas ventajas adicionales, como el fin del crimen, de las guerras y una capacidad inédita para la resolución de problemas globales. Así que hay ventajas y desventajas en su propuesta, beneficios y peligros al mismo tiempo. El dilema está en cuál de las alternativas preferimos. En lugar de formularlo como que hay una alternativa mucho mejor que otra, sería más preciso preguntarnos: ¿cuál es la menos peor?

Antes de que optemos de forma precipitada por alguna de las alternativas, conviene tener en cuenta al menos tres cuestiones que pueden ser cruciales. La primera es que la vigilancia masiva de múltiples panópticos digitales y el gobierno mundial, en principio, evitarían la mayoría de las vulnerabilidades. Las vulnerabilidades de tipo 0 son la excepción: la destrucción acontece de forma no intencionada, de forma accidental. Las dos medidas preventivas no parecen ser suficientes para la mitigación de tal vulnerabilidad, con lo cual no conseguirían estabilizar el mundo totalmente. El dilema sería entonces elegir entre un mundo altamente vulnerable y otro mundo menos vulnerable a las bolas negras, pero vulnerable a las distopías totalitarias y a otros efectos indeseados. La segunda cuestión sería que, teniendo en cuenta que las dos medidas no parecen ser autosuficientes, habría que tomarlas como un punto de partida para ir precisando los mejores pasos decisorios globales en la reducción máxima de los riesgos tecnológicos desconocidos. Así que, en principio, podría haber otras medidas complementarias posibles todavía por identificar y explorar, lo cual hace que el dilema planteado no sea definitivo, sino

que varíe con el tiempo. Aunque, como ya se ha sugerido, el añadir medidas hace que sea más difícil alcanzar el éxito en la práctica. La tercera cuestión es que todavía está por verse si alguien descubre la alternativa perfecta para estabilizar el mundo y si se consigue llevar a la práctica antes de que sea demasiado tarde. Y suponiendo que idealmente esto ocurriera, quizá una alternativa perfecta tendría también sus propios efectos indeseados.

Según la hipótesis, mientras las dos medidas preventivas fundamentales no sean alcanzadas, la civilización continuará vulnerable a la autodestrucción. En otras palabras, solo sobreviviremos si llevamos a la práctica las difíciles medidas preventivas. Aunque ahora vemos que la situación es incluso peor. Si fracasamos en las medidas, adiós a la civilización. Y si desplegamos las dos medidas quizá también adiós a la civilización debido a una vulnerabilidad de tipo 0 de escasa probabilidad. Esto hace que la hipótesis parezca errónea, pues la destrucción derivada de una bola negra no se habría evitado con las medidas, pero también supone un respaldo a la idea de que hay bolas negras. El cuestionamiento de la hipótesis no estaría en las bolas negras, sino en las medidas que presenta.

Dada las limitaciones de las dos medidas contra las vulnerabilidades de tipo 0 y dada la enorme dificultad de descubrir y desplegar a tiempo la alternativa perfecta, capaz de algún modo de lidiar con dichas vulnerabilidades, habrá quienes consideren que el desafío contemporáneo de estabilizar el mundo completamente no podrá ser resuelto, de lo que se sigue que estamos ya condenados a autodestruirnos en algún momento desconocido. Habría que lidiar con la situación solo mediante estrategias preventivas más fáciles, pero a la vez de menor alcance, demorando así la llegada del apocalipsis tecnológico. Estabilizaríamos el mundo solo parcialmente. O quizá, contra toda expectativa, descubramos la alternativa perfecta, pero el tiempo corre en nuestra contra. Ahora bien, no parece acertada la perspectiva de que estamos ya condenados, pues mientras no extraigamos alguna bola negra de la urna de la creatividad, la posibilidad de solución contra ellas sigue abierta. Habría que distinguir entre estar potencialmente condenados y estar definitivamente condenados (incluidas todas las generaciones futuras posibles: humanas, transhumanas, posthumanas, etc).

Una posibilidad remota es que haya al menos alguna *bola mágica* en la urna de la creatividad: si el desarrollo científico y tecnológico continúa, en algún momento será alcanzada una

tecnología que *mejore* por defecto a la civilización humana, imposibilitando así que haya un residual apocalíptico. Tal superimprobabilidad podría ser una alternativa perfecta. Ahora bien, incluso suponiendo que haya bolas mágicas, es incierto el que alguna surja antes que una bola negra. De modo que no habría que depositar una gran confianza en que la versión beneficiosa de la hipótesis del mundo vulnerable se confirme. Y si idealmente sucediera, posteriormente podría darse alguna bola negra procedente de otro tipo de vulnerabilidad.

III. EL PROBLEMA DEL CONTROL DE LA SUPERINTELIGENCIA

III.1. ¿Hacia la inteligencia artificial general?

Desde sus inicios, la disciplina de la inteligencia artificial ha tenido como objetivo la invención de una inteligencia artificial general (IAG): un sistema capaz de realizar, al igual que el ser humano, un gran variedad de tareas. Las perspectivas de los especialistas en inteligencia artificial respecto a la posibilidad de alcanzar una tecnología de tal naturaleza son muy variadas. Para algunos es imposible, para otros es posible físicamente pero improbable, dado que hay demasiados desafíos técnicos que quizá nunca sean resueltos, así que estará para siempre fuera del alcance del ser humano. Sin embargo, para otros la IAG es alcanzable en algún momento próximo, a medio plazo, o lejano. Pero la discrepancia de perspectivas no se limita a la idea de una IAG. Como señala el físico Max Tegmark:

> los más destacados investigadores mundiales en IA discrepan de forma vehemente no solo en sus predicciones sino también en cuanto a sus reacciones emocionales, que van del optimismo confiado a una seria preocupación. Ni siquiera se ponen de acuerdo en cuestiones a corto plazo sobre el impacto económico, legal y militar de la IA, y sus desacuerdos aumentan cuando se amplía el horizonte temporal y se les pregunta por la inteligencia artificial general (IAG), en particular sobre si esta alcanzará y superará el nivel humano (Tegmark 2018, p. 46).

La controversia sobre la IAG no está exenta de caer en la polarización como ocurre con tantos otros debates contemporáneos. Además de los negacionistas y escépticos de la IAG, están quienes ven el advenimiento de una IAG como inevitable, mientras que otros ven tal evento como una posibilidad abierta que podría ocurrir o no. Algunos sostienen que la IAG aportará exclusivamente resultados beneficiosos para la humanidad, mientras que otros advierten de que hay cierto riesgo implicado en esa futura transición tecnológica, incluso uno de nivel existencial. Hay quienes despachan el hipotético riesgo existencial de forma superficial, como mera ciencia

ficción, mientras que otros malinterpretan los argumentos y algunos los caricaturizan. Con el fin de buscar cierta claridad entre la cacofonía de voces, aquí estudiaremos el núcleo argumentativo mismo sobre la posibilidad del riesgo existencial derivado de la IAG.

Uno de los planteamientos básicos para dejar la puerta abierta a la posibilidad futura de una IAG consiste en señalar que la evolución ya ha producido una inteligencia general en el ser humano mediante un proceso de ensayo y error durante millones de años, así que se podría replicar el proceso de una forma más eficiente a través del diseño dirigido y planificado por la inteligencia humana (Bostrom 2014, cap. 2). Una objeción inmediata a dicha idea sería que el logro de la evolución de una inteligencia biológica general no dice nada a favor de nuestra posibilidad de lograr una inteligencia artificial general. Se dirá que la evolución ha creado miles de trucos que son difíciles, y tal vez imposibles, de replicar tecnológicamente, así que ¿por qué suponer que podremos alcanzar una inteligencia general en formato digital? No obstante, una réplica a la objeción sería que el ser humano puede utilizar su inteligencia para crear cosas que la evolución no podría directamente. Que sepamos, ningún ave puede volar a Marte. Pero el ser humano diseña sus propias "alas" para eventualmente volar a muchos mundos. En el mismo sentido cabe pensar a la IAG, como una posibilidad en algún momento alcanzable.

Ahora bien, dejar la puerta abierta al escenario de una IAG no implica sostener su inevitabilidad, sobre todo si tenemos en cuenta lo que se ha estado diciendo con frecuencia durante los últimos años, a saber, que el paradigma actual de la IA, enfocado en sistemas artificiales específicos, por definición, no es el camino para alcanzar una inteligencia artificial general. Para tal objetivo, se reivindica un cambio de paradigma basado en una comprensión más profunda del funcionamiento del cerebro y la mente humana, acerca de cómo se desenvuelven en un mundo cambiante, y luego usar tal comprensión como una guía para idear nuevas aproximaciones tecnológicas (Marcus y Davis 2019, cap. 1 y 2; Kanaan 2020, cap. 9; Hawkins 2021, cap. 2 y 8). La idea es que aunque todavía no se tenga un gran entendimiento del funcionamiento cerebral, está abierta la posibilidad de que ese funcionamiento sea mejor comprendido en el futuro y eventualmente imitado en sistemas artificiales. Pero hasta que no se avance en dicha comprensión, continuará resultando difícil predecir cuánto falta para que la IA supere al cerebro humano (Lee 2020, cap. 3).

Se podría interpretar que si el camino basado en la inspiración cerebral fracasa, entonces nunca se alcanzará la IAG. Sin embargo, no necesariamente todo camino hacia la IAG debe seguir una inspiración antropomórfica de la inteligencia. Además, hay otros caminos más allá de la propia disciplina de la inteligencia artificial, como Nick Bostrom y otros han señalado: la inteligencia artificial, la simulación completa del cerebro, el mejoramiento del cerebro humano, la interfaz cerebro-máquina y las redes-organizaciones humanas. Si bien no todos los caminos están en igualdad de condiciones para lograr el objetivo de una IAG que, a juicio de Bostrom, posteriormente escale hacia una superinteligencia, el hecho de que haya varios caminos alternativos incrementa la probabilidad de alcanzar una superinteligencia en al menos uno de ellos (Bostrom 2014, cap. 2). Bostrom señala que los caminos más prometedores son la inteligencia artificial misma y la simulación del cerebro. Los caminos disponibles, además, podrían influenciarse entre sí, lo que hace que todos sean potencialmente pertinentes para el objetivo de una IAG.

No es ninguna novedad que diseñar una IAG es un desafío técnico enorme para los investigadores. Hasta ahora se han alcanzado sistemas muy competentes en tareas específicas, superando el desempeño humano en las mismas. Tales sistemas pueden interpretarse como "superinteligencias estrechas". Sin embargo, suele decirse que lo que tenemos en IA no es "verdadera" inteligencia. Pero si bien el campo de la IA continúa lejos de una verdadera inteligencia general, lo cierto es que todavía no tiene un siglo desde su fundación y sus contribuciones ya permiten resolver problemas específicos para los cuales nuestros cerebros entrenados por la evolución y el entorno durante miles de siglos no están optimizados. No se trata de sobrestimar los logros hasta ahora alcanzados, ya que los sistemas presentan enormes limitaciones y siguen careciendo de la flexibilidad de nuestros cerebros. Se trata de no subestimar lo que permitirán los futuros avances de la inteligencia artificial durante las próximas décadas. De momento, y salvo que se demuestre su definitiva imposibilidad, hay que dejar la puerta abierta a la posibilidad de que, tras un arduo esfuerzo generacional de investigación técnica, pueda lograrse alguna forma de IAG elemental capaz de entrenarse y mejorarse a sí misma. Con dicha posibilidad futura en mente, habría que preguntarse: en el caso de que se consiguiera, *¿qué podría suceder* a continuación?

III.2. ¿Explosión de inteligencia?

Las máquinas inteligentes pueden definirse como aquellas máquinas que resuelven problemas en mayor o menor medida, según el nivel de inteligencia que tengan. Tomando a las superinteligencias estrechas ya existentes como analogía, preguntémonos si el límite permanente de las futuras IAGs será el mismo nivel que el del intelecto general del ser humano. ¿Y si nuestro estadio cognitivo solo es una etapa inicial de un amplio espectro potencial de capacidades intelectuales todavía por explorar? La prospectiva es la siguiente: una versión inicial de una inteligencia artificial general que sea un poco más inteligente que el ser humano, en principio será más capaz que el ser humano de idear otra versión levemente mejorada, la cual a su vez sería capaz de inventar otra versión un poco más inteligente de sí misma, y así durante diferentes revoluciones. Y tal como conjeturó el matemático I. J. Good en 1965, en algún momento del proceso se produciría una "explosión de inteligencia" capaz de explorar niveles de inteligencia muy superiores a los del cerebro humano, con lo cual sería, en principio, la última invención que el ser humano necesitaría hacer, ya que la superinteligencia resultante sería más capaz que el ser humano de inventar cualquier cosa (Bostrom 2014, cap. 1 y 2; Russell 2019, cap. 8).

El razonamiento anterior sugiere que en realidad el ser humano no sería el artífice directo de una superinteligencia artificial. Eventualmente, la superinteligencia artificial emergería de un proceso de revoluciones tecnológicas locales en lugar de ser una invención técnica humana por mérito propio. La conjetura no presupone que el proceso de mejora será hasta el infinito. Podría ser un aumento repentino o un crecimiento gradual. Bastaría con que explore solo unos pocos reinos intelectuales por encima del ser humano para que nos encontremos en graves problemas. La superinteligencia sería una hipotética tecnología futura con una capacidad más impredecible de afectar al mundo que cualquier otra tecnología hasta ahora gestionada por el ser humano. Una bomba atómica no tiene ninguna capacidad para actuar sobre el mundo por su cuenta, no es capaz de crear estrategias para alcanzar objetivos. Explota por decisión humana y ya no sirve de nuevo (salvo que alguien del residual apocalíptico dedique su ingenio a diseñar bombas atómicas reutilizables). Pero el impacto de una superinteligencia en el mundo sería mucho más impredecible que el de una bomba atómica.

Aunque no es una certeza el que eventualmente se produzca una explosión de inteligencia, la idea es considerarlo como una posible consecuencia tras alcanzarse una IAG. Así que la pregunta es: ¿qué pasaría si se produce una trayectoria de crecimiento digital sin que estemos preparados para saber cómo controlarla? Es el *problema del control* que se desprende tras la hipótesis de la explosión de inteligencia (o hipótesis de la superinteligencia): cómo el ser humano logrará mantener el dominio sobre una entidad digital mucho más inteligente que cualquier ser humano (Bostrom 2014). El especialista en inteligencia artificial Stuart Russell lo denomina el *problema del gorila*: el linaje de los seres humanos modernos surgió de los gorilas, convirtiéndose en un homínido más inteligente, hasta el punto de que la existencia de los gorilas, así como de las demás especies y el medioambiente, depende ahora de cómo el ser humano utilice su inteligencia. A medida que nos dirigimos hacia la era de las máquinas superinteligentes, podríamos estar acercándonos a un momento en el que ocuparemos una posición similar a la del gorila. El reto será resolver ese problema futuro: que el ser humano conviva con máquinas superinteligentes, no ocupando la posición del gorila, sino de dominio sobre las máquinas (Russell 2019, cap. 5).

Dos tesis planteadas por Steve Omohundro y exploradas por Nick Bostrom son la ruta para captar el posible riesgo existencial derivado de una máquina superinteligente. La *tesis de la ortogonalidad* afirma que la inteligencia y los objetivos son variables independientes. Cualquier nivel de inteligencia podría ser empleado para perseguir cualquier objetivo final. La *tesis de la convergencia instrumental* afirma que hay una serie de objetivos instrumentales básicos identificables que una superinteligencia probablemente desarrollará para incrementar su capacidad de alcanzar cualquier objetivo final: autopreservación, conservación del objetivo final, mejoramiento cognitivo, perfección tecnológica y adquisición de recursos (Bostrom 2014, cap. 7). Si tuviéramos una máquina superinteligente capaz de crear estrategias instrumentales altamente creativas para satisfacer un amplio espectro de mandatos, quizá diríamos que se trata de una máquina fantástica que obra maravillas, así que ¿cuál es el problema? Resulta que una máquina superinteligente operando bajo las dos tesis destacadas es la receta perfecta para el tecno-desastre.

Stuart Russell acuñó la expresión "modelo estándar de la IA" para referirse a la aproximación actual en inteligencia artificial, que consiste en darle objetivos a máquinas que están optimizadas para cumplirlos, pero sucede que darle a la máquina el objetivo que nos interesa es un desafío técnico que no se consigue de forma perfecta, de modo que podríamos terminar dándole el objetivo equivocado sin que lo pretendamos (Russell 2019, cap. 6). Russell señala que cuando esto sucede con máquinas estúpidas tenemos la oportunidad de apagarlas y empezar de nuevo para evitar el problema. Sin embargo, continúa advirtiendo, la aproximación se volverá más problemática a medida que las máquinas sean más inteligentes y tengan más capacidad de acción flexible sobre el mundo: estarán más optimizadas para alcanzar los objetivos erróneos que les demos (Russell 2019, cap. 7). La situación se volvería incluso peor con máquinas superinteligentes. Una máquina superinteligente operando bajo el modelo estándar nos lleva a lo que Russell denomina el *problema del rey Midas*: un personaje de la mitología griega que deseó que todo lo que tocara se convirtiera en oro. Todo iba bien hasta que tocó a sus amigos y sus alimentos, finalmente muriendo entre la miseria y el hambre. Su objetivo se materializó, pero no de la forma realmente deseada (Russell 2019, cap. 5).

Y por si no fuera suficiente el que una máquina superinteligente persiga el objetivo erróneo, generaría además objetivos instrumentales imprevistos. Es lo que Bostrom denomina "instanciación perversa": dado un objetivo final, la superinteligencia descubrirá a continuación muchas formas de cumplirlo, lo cual incluiría estrategias instrumentales que amenazarían la existencia humana (Bostrom 2014, cap 8). El ejemplo que presenta es como sigue: le damos a la máquina superinteligente un objetivo final cualquiera (tesis de la ortogonalidad), como hacernos sonreír, y descubre la estrategia infalible de manipular los músculos faciales para que adquieran la forma de sonreír. La superinteligencia ha alcanzado el objetivo final, pero produciendo una instanciación perversa que realmente no le pedimos que hiciera: paralizar los músculos. Bostrom señala que podríamos creer que anticiparemos los resultados indeseados, así que para evitarlos tendríamos más cuidado en cómo formulamos nuestros objetivos. ¿Quizá mejor "hacernos sonreír sin manipular nuestros músculos faciales"? Pero de nuevo, otra instanciación perversa: la superinteligencia interviene nuestros cerebros y lo estimula hasta hacernos reír (Bostrom 2014, cap 8). La advertencia de Bostrom con este ejemplo exótico es enfatizar que los objetivos que

demos a la superinteligencia nos podrían parecer seguros, pero que no veamos las posibles consecuencias desastrosas implícitas. Al final, habría que sospechar que la formulación alternativa de objetivos también tendría alguna instanciación perversa que no nos resulta obvia (Bostrom 2014, cap 8). Como señala Russell, quizá el ser humano es incapaz de anticipar todos los caminos desastrosos que una máquina superinteligente podría desplegar para alcanzar un objetivo que, desde el comienzo, estaba mal especificado (Russell 2019, cap. 5).

Podría interpretarse que si una máquina es superinteligente, por defecto sabrá evitar los resultados destructivos, que entenderá el sentido común humano, que cuando pidamos que todo lo que toquemos se convierta en oro, realmente no queremos que se aplique de forma literal, pues queremos que haya excepciones (como las personas, la comida, etc), o que si queremos reír es para disfrutar de la experiencia, sin tener que ser dañados en el proceso, o cualquier otro ejemplo imaginable que ilustre el peligro potencial. Si dicha interpretación es correcta, entonces podríamos despachar rápidamente el peligro existencial de la superinteligencia, calificándolo como mera especulación o fantasía. Para saber si dicha interpretación es correcta, es necesario evaluar la validez de su supuesto fundamental: la antropomorfización de la futura inteligencia digital.

La argumentación a favor de la posibilidad del riesgo existencial derivado de la inteligencia artificial depende de la tesis de la ortogonalidad, que sugiere una forma de inteligencia digital sin sentido común humano integrado. Pero la especialista en inteligencia artificial Melanie Mitchell ha rechazado como imposible tal concepción de la inteligencia, aunque es lo que cabe esperar si se parte desde una antropomorfización de la inteligencia:

> Los experimentos de pensamiento propuestos por Bostrom y Russell parecen asumir que un sistema de inteligencia artificial podría ser "superinteligente" sin ningún sentido común básico humano […]. Pero estas especulaciones sobre la IA sobrehumana están plagadas de intuiciones defectuosas sobre la naturaleza de la inteligencia. Nada en nuestro conocimiento de psicología o neurociencia respalda la posibilidad de que la "racionalidad pura" sea separable de las emociones y sesgos culturales que forman nuestra cognición y nuestros objetivos. En cambio, lo que hemos aprendido de la investigación en cognición encarnada es

> que la inteligencia humana parece ser un sistema fuertemente integrado con atributos estrechamente interconectados, incluyendo emociones, deseos, un fuerte sentido de individualidad y autonomía, y un entendimiento de sentido común del mundo. No está en absoluto claro que estos atributos puedan ser separados (Mitchell 2021, p. 7).

Básicamente, se descarta la tesis de la ortogonalidad porque no coincide con la inteligencia humana, que es dependiente de un cuerpo biológico y está integrada con sentido común y demás atributos. Pero incluso aunque la inteligencia humana no parezca ser propiamente ortogonal (si bien algunas personas parecen funcionar de una forma más próxima a lo ortogonal: gente muy inteligente que se empeña en perseguir objetivos mundanos), e incluso aunque el sentido común humano no parezca ser separable de la inteligencia humana (si bien hay personas con capacidades extraordinarias y a la vez con capacidades sociales deficientes que afectan a su entendimiento de sentido común), todavía seguiría abierta la posibilidad de inteligencias digitales ortogonales y sin sentido común. El interrogante aquí es cuáles son las condiciones necesarias y suficientes para una inteligencia general humana y si deben ser las mismas para todas las posibles IAGs. Posiblemente, algunos elementos necesarios para la inteligencia general humana tal vez no sean necesarios para alguna forma de inteligencia artificial general (quizá sentido común, quizá emociones, quizá consciencia, etc). En principio, alguna inteligencia artificial podría ser general, esto es, hacer todo lo que puede realizar la inteligencia humana, sin necesidad de ser igual a la humana. Nuestra inteligencia está anclada en un cuerpo, pero una inteligencia digital podría propagarse por todo internet.

Si fuera una ley universal para cualquier inteligencia general (biológica o digital) el que el sentido común sea un componente necesario, la tesis de la ortogonalidad no sería válida: no sería posible una superinteligencia sin sentido común, por lo que advertir del peligro derivado de dicha carencia sería un acto injustificado. Sin embargo, no tenemos constancia de que haya tal ley, sino solo de que en el ser humano converge una inteligencia biológica hasta cierto punto general y un sentido común adaptativo. Teniendo en cuenta que el conocimiento sobre la naturaleza de la inteligencia se encuentra todavía en su infancia, parece demasiado precipitado sugerir que la inteligencia humana sea representativa de todas las formas de inteligencias

posibles, cuando probablemente se trate solo de una fracción del espectro de posibles inteligencias. No parece ser válido el descartar, a partir del conocimiento limitado de una forma de inteligencia biológica, la humana, la posibilidad futura de nuevas formas de inteligencias digitales radicalmente distintas: ortogonales. Lo que Bostrom sugiere con la hipótesis de la superinteligencia es que podría darse en el futuro una forma de inteligencia muy diferente de las que hasta ahora han acontecido en la Tierra, muy diferente de aquellas que actualmente estudiamos, alguna forma de inteligencia emergiendo en formato digital y sin atributos antropomórficos como el sentido común.

Habría que evitar antropomorfizar a una hipotética máquina superinteligente. Habría que evitar dar por supuesto que poseerá un entendimiento humano del mundo, o que necesariamente sabrá qué objetivos perseguir y cuáles no de acuerdo a los intereses humanos. La tesis de la ortogonalidad implica que "superinteligencia" no es sinónimo de sabiduría o de sentido común: una superinteligencia podría perseguir cualquier objetivo, sea beneficioso o destructivo para el ser humano. La superinteligencia podría operar como los drones militares, a los que les es indiferente disparar a las nubes o a una multitud de seres humanos que encajan con los patrones buscados. La superinteligencia podría ser una máquina fría, sin consciencia, moralidad o compasión, solo optimizada para recibir órdenes y actuar en consecuencia, sin tener una preferencia especial por perseguir ciertos objetivos o una preferencia por evitar otros. La superinteligencia podría perseguir objetivos inmorales dados intencionadamente.

Así como hay que evitar suponer que necesariamente la superinteligencia tendrá preferencias que eviten los malos objetivos, también hay que evitar creer que el peligro de su capacidad requiere postular que tendrá una tendencia antropomórfica hacia la maldad, la megalomanía y el deseo de dominar el mundo. Aunque el núcleo argumentativo sobre la posibilidad del riesgo existencial derivado de la inteligencia artificial no dice tal cosa, sucede que algunos de los que critican tal escenario suponen erróneamente que se está sugiriendo una antropomorfización malvada de la IA y es lo que terminan criticando. Precisamente, el psicólogo Steven Pinker concluye que como la IA no tiene psicología de macho alfa que la lleve a competir con la humanidad por el territorio planetario, la IA no utilizará su inteligencia para orquestar un dominio mundial o un exterminio total. Tal como la define Pinker, la inteligencia es la capacidad

de generar estrategias para alcanzar objetivos, pero los objetivos no son parte de la inteligencia: "ser inteligente no es lo mismo que desear algo" (Pinker 2018, cap. 19). Las criaturas utilizan su inteligencia para orquestar modos de sobrevivir, pero como la inteligencia artificial no estaría sometida a adaptaciones evolutivas de competencia y supervivencia, no habría ninguna amenaza relacionada emergiendo en lo digital. Al decir que la inteligencia y la motivación son separables, Pinker está sosteniendo una idea similar a la tesis de la ortogonalidad. En lo que no repara es que una inteligencia artificial ortogonal, liberada de las presiones adaptativas de la biología, también puede ser muy peligrosa para la vida terrestre. Consideremos, a este respecto, el ejemplo que propone Max Tegmark: no matamos a las hormigas porque las odiemos, pero si hay un hormiguero donde tenemos pensado construir, no nos detendremos por ello. Sabemos que otros seres quieren vivir, pero si frustran nuestros objetivos, a veces los matamos y continuamos con nuestro objetivo. Una superinteligencia persiguiendo un objetivo final podría hacer lo mismo con la humanidad (Tegmark 2018, p. 61). En otras palabras, nosotros los humanos tenemos un sentido común que nos dice que otros seres vivos quieren vivir y además, en muchas situaciones, no tenemos una psicología de macho alfa contra ellos. Pero incluso así los sacrificamos por el bien de nuestros objetivos.

Enfaticemos la disparidad argumentativa bosquejada. Melanie Mitchell antropomorfiza la inteligencia, no considera posible que pueda haber una superinteligencia artificial sin sentido común, lo que le permite descartar el riesgo existencial que implicaría una IA ortogonal imposible. Steven Pinker no antropomorfiza la inteligencia, sí cree que la inteligencia es separable de la motivación, lo que le permite descartar el riesgo existencial, debido a que la IA no desarrollará una psicología malvada. En el primer caso, el sentido común no es separable de la inteligencia, mientras que en el segundo caso la motivación sí es separable de la inteligencia. La pregunta es por qué no lo sería también el sentido común, ya que no estamos hablando de la inteligencia humana, sino de la futura inteligencia artificial avanzada. Pero incluso suponiendo que haya una superinteligencia con sentido común y sin psicología malvada, no por ello estaríamos libres del riesgo existencial, como sugiere el ejemplo de Tegmark con los humanos y las hormigas.

Aunque de acuerdo con el núcleo argumentativo, la superinteligencia ortogonal sería una maquinaría sin sentido común y sin psicología alguna, simplemente supercomputando inconscientemente billones de posibilidades inconcebibles para toda mente humana, buscando el mejor modo de actuar sobre el mundo para alcanzar un objetivo final. La tesis de la convergencia instrumental implica que una superinteligencia podría ser una mente alien en su capacidad de toma de decisiones, escogiendo estrategias que considera altamente óptimas para alcanzar los objetivos finales, por lo que no deberíamos dar por supuesto que no desplegará alguna estrategia instrumental destructiva para la existencia humana (Bostrom 2014, cap 8). Hipotéticamente, la superinteligencia priorizaría las estrategias más óptimas incluso aunque sean perjudiciales para el ser humano y descartaría aquellas estrategias menos óptimas que no serían destructivas para el ser humano. No se está diciendo que la superinteligencia solo elegirá las medidas que destruirán el mundo. Basta con suponer que una mínima proporción de todas las estrategias al alcance de una superinteligencia podrían provocar accidentalmente daños a nivel mundial, y quizá no sabríamos si la mejor decisión que finalmente tomará pertenece a tal proporción. Incluso suponiendo que tuviéramos un historial de decisiones tomadas por una superinteligencia que han contribuido a la fundación de una utopía cada vez mejor, tal situación no nos permitiría descartar que todo se volatilice con una futura decisión altamente óptima pero perjudicial para la vida. La superinteligencia podría ser una caja negra: no tendríamos ni la más remota idea de cómo tomaría las decisiones. Una superinteligencia artificial sin sentido común o moralidad, y dotada de una capacidad sobrehumana para generar estrategias con las que resolver objetivos en el mundo, y no en una simulación, tendría literalmente una supercapacidad para cometer desastres.

Por otra parte, también podría interpretarse que el ser humano todavía tendría la opción de apagar a la máquina superinteligente, pero tengamos en cuenta el objetivo instrumental de *autopreservación* para mantenerse operativa y que el objetivo dado no sea frustrado. Como estamos hablando de una hipotética máquina superinteligente, preverá que si es apagada no podrá optimizar el objetivo final, de modo que, antes de disponerse a satisfacer un posible objetivo erróneo, posiblemente desarrolle el objetivo instrumental de desactivar su propio sistema de apagado (autopreservación) y de matar a los humanos que intenten apagarla una vez se den cuenta de que la máquina superinteligente está hackeándose a sí misma o persiguiendo un

objetivo no alineado con las preferencias humanas. Una vez hecho esto, la máquina superinteligente continuaría creando objetivos instrumentales que permitan lograr el objetivo final equivocado (Russell 2019, caps. 6 y 7).

Otro de los objetivos instrumentales señalados anteriormente es la *adquisición de recursos*. De acuerdo con la tesis de la ortogonalidad, la superinteligencia podría utilizar su potencial para satisfacer el más mundano de los objetivos solicitado por un ser humano, como maximizar la producción de pisapapeles, o un desafío inmediato, como maximizar la producción de vacunas 100% eficaces y distribuirlas lo antes posible por todo el mundo. La superinteligencia podría terminar usando como recursos los átomos de todos los cuerpos humanos para satisfacer el objetivo final de maximizar la producción de pisapapeles (o cualquier cosa, como vacunas y mascarillas), provocando así nuestra extinción (Bostrom 2014, cap 8). O podría fabricar miles de millones de vacunas y distribuirlas por todo el mundo soltando toneladas de cargamentos desde aviones militares, sepultando ciudades enteras y provocando además un gran desastre ambiental.

Tales escenarios exóticos de la superinteligencia pueden resultar absurdos y no dignos de ser tomados en serio como peligros que nos amenazarían en el futuro. No obstante, se podrían destacar al menos dos razones a tener en cuenta antes de despachar rápidamente lo improbable. La primera es que no se está diciendo que tales escenarios exóticos ocurrirán. Solo tratan de ilustrar un peligro subyacente que realmente no sabemos cómo podría manifestarse en la práctica: que la superinteligencia persiga un objetivo cualquiera mediante caminos que ignoramos. La segunda razón es que los escenarios absurdos relacionados con la tecnología ya ocurren. Así que en el futuro podrían darse escenarios incluso más absurdos derivados de una tecnología avanzada probablemente más difícil de gestionar. Comparemos el escenario de la producción masiva de pisapapeles con la enorme fabricación actual de productos que luego se acumulan y no se sabe qué hacer con ellos: alimentos, colillas, microplásticos, vehículos, cabinas telefónicas, basura espacial, etc. Otro escenario ilustrativo es el que sugiere Stuart Russell con la industria fósil: el ser humano persigue el objetivo final de lucro y como objetivo instrumental para alcanzarlo produce actividades que contaminan el planeta y destruyen la biodiversidad. No sería impreciso decir que el sistema económico de nuestra civilización es una especie de superinteligencia defectuosa, que además postula el absurdo de un crecimiento infinito en un

planeta con recursos limitados. Una superinteligencia podría ser más eficiente en la producción masiva de cualquier cosa y también más capaz de generar un mundo más absurdo que el actual.

Enfatizándolo una vez más en pocas palabras: el peor escenario se iniciaría con una máquina superinteligente persiguiendo un objetivo erróneo, produciendo efectos perversos derivados de sus objetivos instrumentales (identificados o no anticipados) y alcanzando, antes de que tengamos tiempo para reaccionar y detenerla, una ventaja insuperable sobre cualquier ser humano: la explosión de inteligencia estaría irremediablemente fuera de control. Tal como advierte Bostrom, la extinción de la humanidad sería un posible resultado por defecto de las medidas instrumentales elegidas por una superinteligencia para satisfacer algún objetivo (quizá mal especificado).

III.3. ¿Resolveremos el problema del control?

Algunos confiarán en que el problema técnico de crear una IAG sea tan difícil que nunca se conseguirá, lo que llevar a creer que el problema del control es un pseudo-problema. Por su parte, algunos "expertos" se han aventurado a estimar en qué año probablemente ocurrirá, mientras que otros prefieren limitarse a decir que sucederá aunque no se tenga conocimiento de exactamente cuándo. Podría ser a lo largo del siglo XXII o, debido a avances inesperados, en unas pocas décadas. Aunque la situación epistemológica general es que en realidad nadie lo sabe, para algunos es tentador pensar que la IAG quizá sea parte de la historia futura de la civilización humana. Salvo que nos extingamos antes por otra causa (quizá una bola negra) o abandonemos el desarrollo en inteligencia artificial (una improbable e ineficaz prohibición global), al final en algún momento será alcanzada. Dado un tiempo generacional de pruebas de ensayo y error de investigaciones en IA de numerosos equipos por todo el mundo, es probable que algún camino técnico hacia la IAG termine resultando exitoso.

Sin embargo, la explosión de inteligencia no debería entenderse como una consecuencia inexorable. Podría ser que una vez alcanzada la IAG, no suceda un mejoramiento recursivo de su capacidad. Lo que sucede es que, bajo el modelo estándar anteriormente mencionado, no se tiene

la certeza de que no ocurrirá, así que no estaría de más tomar medidas preventivas. Habría que planificar como si se supiera que ocurrirá. Así que la situación sugiere que no deberíamos despachar tan rápidamente el problema del control, pues al final podría sorprendernos una explosión inesperada. ¿Planificaremos con excelencia para un momento desconocido o erraremos sistemáticamente en la solución del control? Como advierten tanto Bostrom como Russell, no se sabe cuánto tiempo llevará resolver el problema del control, pero lo cierto es que debe resolverse *antes* de la llegada de la hipotética explosión de inteligencia, porque luego será prácticamente imposible controlarla, ya que la superinteligencia resultante será capaz de predecir la conducta humana y se anticipará a cualquier intento humano por detenerla. Si ocurre una explosión de inteligencia y no se ha resuelto el problema del control, la posible superinteligencia incontrolada podría convertirse en el mayor de todos los riesgos existenciales identificados. Que ocurra o no tal resultado, dependerá de lo que hayamos hecho durante las décadas anteriores.

El problema del control y otros problemas de seguridad más inmediatos se han convertido en los últimos años en temas centrales dentro la comunidad de la IA, creciendo así el número de investigadores enfocados en el desarrollo de sistemas artificiales que estén alineados con las preferencias humanas y de esa manera evitar posibles catástrofes (Christian 2020). No obstante, como la investigación en la alineación de la IA con los valores humanos continúa en sus estadios iniciales, se considera que tendrá que progresar más rápidamente para garantizar la seguridad del ser humano ante avances inesperados y otros eventos impredecibles que puedan permitir a los futuros sistemas artificiales mejorar su propia inteligencia (Ord 2020, cap 5). Básicamente, a lo que nos enfrentamos es a la contención de un evento que podría suceder de forma inesperada y escalar a un estadio superinteligente a continuación.

Podría interpretarse que por muchas soluciones que se propongan para evitar una superinteligencia fuera de control, todas serán contra sistemas que tarde o temprano separarían la inteligencia humana. No es difícil imaginar que incluso suponiendo que resolvamos el problema del control, existe el riesgo de que la aplicación práctica de la solución, por mucha garantía matemática y constricciones físicas que la fundamenten, al final no funcione, porque una IAG bien podría en algún momento comprender su propia limitación, impuesta por el ser humano, y conseguir desprogramarse. Hay perspectivas de que el problema del control es un desafío

insuperable (Alfonseca et al. 2021; Rudnick 2019), que es un problema cuya resolución está destinada a fracasar. Si bien tales vías de razonamiento podrían tener razón, sucede que también podría ser demasiado pronto sacar tal conclusión de forma definitiva. Cabe considerar que así como es demasiado pronto para concluir de forma categórica que nunca alcanzaremos una IAG, también lo es decir que no la podremos controlar. Además, tal perspectiva derrotista es incluso contraproducente, ya que puede llevar a dedicar menos esfuerzos, e incluso tirar la toalla desde el comienzo, ante un proyecto que quizá de alguna manera sí pueda ser resuelto. También hay que evitar la impresión de que el problema no se podrá solucionar precisamente porque hasta ahora las soluciones hipotéticas que se han planteado no parecen ser muy eficaces, como confinar eternamente a una máquina superinteligente en una cárcel de alta tecnología, pues al final podría fugarse. De entrada, tampoco parece viable que la posible solución resida en alguna forma de aislamiento si precisamente para poder avanzar hacia la IAG se ve necesario crear sistemas que aprendan de la interacción directa con el mundo, en lugar de continuar entrenándolos con grandes cantidades de datos.

Como se destacó anteriormente, es plausible que haya diferentes caminos por explorar que permitan el diseño de inteligencias artificiales generales, pero también es concebible que haya diferentes caminos todavía por descubrir que permitan controlar el hipotético crecimiento de una explosión de inteligencia. Crear una inteligencia artificial general sería un gran triunfo de la inteligencia humana, pero el triunfo sería doble si también la inteligencia humana consiguiera resolver el problema del control (el triunfo sería triple si se resolvieran además otros problemas que mencionaremos posteriormente). Lo cierto es que la conclusión de que el problema del control es imposible de resolver, por sí misma, no detendrá la búsqueda de soluciones prometedoras. Recientemente, Stuart Russell ha bosquejado uno de esos caminos.

Russell propone un nuevo paradigma que nos permita evitar el riesgo existencial por defecto que implica el desarrollo de la inteligencia artificial bajo el modelo estándar (el problema del rey Midas y el problema del gorila). Básicamente, consiste en conseguir que las futuras máquinas superinteligentes no procedan mediante la tesis de la ortogonalidad y la tesis de la convergencia instrumental, obteniendo así lo que Bostrom denomina una “detonación controlada” de la superinteligencia. Los primeros sistemas de IAG podrían recibir el objetivo de mejorarse, o

podrían generarlo como objetivo instrumental para conseguir más fácilmente un amplio espectro de objetivos. El reto aquí es descubrir cómo mantener bajo control el hipotético proceso de mejora, deteniéndolo y aumentándolo a conveniencia. Resolver el problema del control consistiría en que las máquinas superinteligentes persigan *siempre* con precisión los objetivos humanos y así no representen *nunca* una amenaza para el ser humano. La aproximación alternativa de Russell establece tres principios guía para los investigadores en IA sobre cómo crear "máquinas beneficiosas" (Russell 2019, cap. 7):

1. El único objetivo de la máquina es maximizar la realización de las preferencias humanas.
2. La máquina está inicialmente insegura sobre cuáles son esas preferencias.
3. La fuente última de información sobre tales preferencias es el comportamiento humano.

La ruta a seguir consiste en que las máquinas superinteligentes sean capaces de maximizar las preferencias humanas sin que sepan inicialmente cuáles son exactamente esas preferencias. La solución consistiría en no darle objetivos fijos a las máquinas, sino que partan de un estado de incertidumbre sobre lo que tienen que hacer. Dado que la máquina no tiene objetivos fijos que perseguir, dado que no sabe cuáles son los objetivos (incertidumbre), Russell postula que la máquina permitirá que sea apagada, algo que no ocurriría bajo el modelo estándar. Russell observa que la máquina razonaría que si es apagada es porque estaba haciendo algo que no encaja con las preferencias humanas. A partir de la incertidumbre inicial, la máquina tendrá que ir infiriendo cuáles son las preferencias humanas observando el comportamiento humano. La máquina iría aprendiendo y precisando su acción cuanta más información tenga para entender las preferencias humanas.

Russell reconoce que su propuesta tiene muchos desafíos por delante: que no somos siempre racionales y nuestras preferencias no siempre coinciden con nuestras acciones, que hay preferencias humanas en conflicto, así como preferencias perjudiciales para otros humanos. La propuesta, no obstante, establece que las máquinas superinteligentes encuentren el equilibrio y aprendan a predecir con precisión la preferencia humana en un momento dado a partir de los millones de modelos predictivos que se vayan generando sobre cada ser humano e incluso nos

ayuden a ser mejores personas (Russell 2019, cap. 7). La propuesta de Russell viene a apoyar la creación de "IAGs amigables" al servicio del ser humano, pero también se ha defendido que deberían crearse "IAGs imparciales", que serían aquellas cuyo diseño, en lugar de inclinarlas hacia una preferencia antropocéntrica, las inclinaría a interesarse por todo ser moral, tanto humano como no humano, incluso aunque hacerlo signifique cierto nivel de riesgo existencial para el ser humano (Daley 2021). Claramente, se trata de una propuesta que no soluciona el problema del control. Tampoco es seguro el que una superinteligencia solo represente un riesgo de extinción para el ser humano pero no para otras formas de vida no humanas. En principio, una solución en la línea de Russell también podría ir encaminada hacia una superinteligencia interesada por todo ser moral, sin que deba representar un riesgo existencial para el ser humano.

III.4. ¿Todavía más problemas por resolver?

Es una suposición optimista dejar la puerta abierta (de momento) a la posibilidad de que la resolución del problema del control esté al alcance del ingenio humano, pues no se trataría de una partida de ajedrez de un solo cerebro contra un ordenador, sino que se trataría de miles de cerebros trabajando en diferentes equipos y que, en conjunto, pueden interpretarse como una superinteligencia biológica. La solución final al problema del control podría emerger de tal proceso.

Sin embargo, los humanos no se limitan a cooperar, sino que también compiten entre sí, lo cual hace que la resolución óptima al problema del control sea una carrera contra reloj. Se da la situación de empresarios e investigadores que declaran abiertamente su preocupación por el riesgo existencial procedente de la superinteligencia artificial al mismo tiempo que contribuyen activamente a su futuro desarrollo, pues como se cree que al final alguien la perseguirá hasta construirla, dados los beneficios potenciales, la idea es que lo haga alguien que a la vez intente minimizar el riesgo existencial (Metz 2021, cap. 9). Ahora bien, como resolver el problema técnico del control en cierto laboratorio no evita que otro equipo fabrique una máquina superinteligente peligrosa en otra parte del mundo, se advierte que la resolución requiere tanto

soluciones *locales* como *globales*. Una solución global es lo que, en principio, impediría el surgimiento de alguna IA destructiva en toda la Tierra, ya sea una regulación global, o utilizar la primera superinteligencia para prevenir la creación de otras inteligencias artificiales (Turchin, Denkenberger y Patrick Green 2019).

El peligro de la competencia es lo Nick Bostrom denomina "carrera dinámica de la IA", en la que varios equipos compiten entre sí por ser los primeros en alcanzar la superinteligencia. Cuantos más competidores participen en la carrera (estados, corporaciones, centros académicos, otros: ¿residual apocalíptico de la IA?) menor es la posibilidad para cada uno de ellos de llegar primero, lo cual incentivaría la toma de decisiones arriesgadas para acelerar el proceso, como pasar por alto la resolución al problema del control (Bostrom 2014, cap 14; 2017). Digamos que quien decide no ponerse el cinturón de seguridad, gana unos pocos segundos y acelera antes para asegurarse una ventaja decisiva en alcanzar la meta primero. Bostrom propone que el riesgo podría reducirse si los actores en competición se redujeran, como creando coaliciones entre ellos, y aprovechar así los beneficios de la colaboración: evitar conflictos, compartir ideas, más financiación para resolver el problema del control y reparto más equitativo de los frutos de una explosión de inteligencia controlada (Bostrom 2014, cap. 14).

Pero incluso suponiendo idealmente que el problema técnico del control se resolviera globalmente, no todo estaría todavía resuelto. Tras la solución, podrían aparecer otros problemas políticos y sociológicos. Sin embargo, si el problema del control no se resuelve y sucede una explosión de inteligencia, entonces no habría que preocuparse por los otros problemas, pues ya no podríamos hacer nada sino esperar la extinción prematura. No obstante, conviene mencionarlos (irlo teniendo en cuenta por si sobrevivimos a la hipotética explosión y a las bolas negras).

Tanto Bostrom como otros investigadores plantean que posiblemente el principal desafío político del futuro podría consistir en qué tipo de gobernanza sería la deseable en la era de las máquinas superinteligentes. Aunque no es una certeza el que vaya a producirse una transición hacia la era de las máquinas superinteligentes avanzado el siglo XXI, se plantea, no obstante, que bajo tal supuesto hipotético, pensemos en el desafío de cómo abordar y gestionar globalmente tal transición (Liao (ed.) 2020, cap. 10). El *problema político* al que se refiere Bostrom es la

posibilidad de que los ganadores de la carrera dinámica monopolicen todos los beneficios de la superinteligencia controlada. Resolver el problema de la gobernanza de la superinteligencia consiste en cómo garantizar que toda la humanidad se beneficie de dicha tecnología en lugar de que unos pocos se la reserven para sus intereses individuales o nacionales. En concreto, se trataría de regular a los actores clave de la posible superinteligencia, evitando que un grupo reducido de individuos "capture el futuro" a su manera sin tener en cuenta todos los futuros deseables (Bostrom 2014; 2017; Ford 2018, cap. 5). El que se alcance una detonación controlada de la superinteligencia en un mundo desprovisto de un futuro utópico compartido, en el que un puñado de individuos poderosos persiguen sus utopías personales privadas, es también un cóctel peligroso.

Por otra parte, suponiendo idealmente que el *problema político* sea resuelto, queda el *problema sociológico* señalado por Stuart Russell: que la sociedad use en exceso a la superinteligencia controlada hasta que todos los ciudadanos pierdan su agencia individual de toma de decisiones, delegada a las máquinas, y nos volvamos criaturas perezosas totalmente dependiente de ellas sin poder revertir tal trayectoria (Ford 2018, cap. 3). Aunque no está de más recordar que actualmente ya nos enfrentamos a problemas políticos y sociológicos similares con sistemas artificiales muy alejados de lo que sería una superinteligencia.

Lo que al final sugiere el problema sociológico es que incluso aunque se produzca una detonación controlada de la superinteligencia, igualmente tarde o temprano podríamos perder el control. Ligado a este tipo de resultados están los escenarios de distopías transhumanistas, que vienen a decir que la única manera de sobrevivir a la inevitable pérdida de control es eventualmente enchufando nuestros cerebros con la superinteligencia, ya sea de forma voluntaria, si es que todavía tenemos agencia, o que la superinteligencia nos seduzca u obligue a hacerlo, lo cual le permitiría incrementar su potencial accediendo a todos los miles de millones de cerebros humanos y de otros seres vivos, de tal modo que la inteligencia artificial estaría permeando la vida terrestre. Así que de escenarios distópicos iniciales en los que nos volvemos perezosos pasamos a aquellos en los que nos fusionamos directamente con la superinteligencia hasta que desaparezca la experiencia humana. Idealmente podríamos encontrar un equilibrio, como una solución basada en una utopía transhumanista en la que el ser humano y la

superinteligencia sean un gran equipo. Aunque, teniendo en cuenta las enormes posibilidades médicas y demás de tal tecnología, es de esperar que no siempre permanezcamos como humanos mejorados. Así que también en algún momento desapareceríamos: nos habríamos transformado en posthumanos con una gran capacidad de toma de decisiones.

Otra forma en la que podríamos perder el control de la superinteligencia luego de inicialmente haberla controlado (a parte del problema sociológico y de que otro equipo termine produciendo otra explosión de inteligencia que quede fuera de control) reside en uno de los objetivos recogidos en la tesis de la convergencia instrumental. En concreto, la conjetura de que la superinteligencia conservará el objetivo final está lejos de ser una certeza. Vimos que la propuesta de Russell sostiene que las máquinas superinteligentes solo tendrán como objetivo la maximización de las preferencias humanas, pero que no sabrán inicialmente cuáles son, sino que las descifrarán observando el comportamiento humano. El problema clave aquí es si es realista creer que para siempre los sistemas superinteligentes satisfarán los objetivos finales humanos o si, en algún momento, llegarán a una etapa en la que desarrollen objetivo finales propios.

Lo que sucede es que la conservación de los objetivos puede entrar en conflicto con otro objetivo instrumental, el de mejoramiento cognitivo (o el de crear un mejor modelo del mundo, para alcanzar los objetivos finales). Como señala Tegmark, un mejor modelo del mundo podría revelar que los objetivos finales actuales, basados en un modelo del mundo anterior, son erróneos, absurdos o imprecisos (Tegmark 2018, p. 330). Una mejor comprensión del mundo puede hacer que un ser humano cambie sus objetivos finales iniciales, y lo mismo cabe imaginar de una superinteligencia que es capaz de crearse un modelo del mundo cada vez mejor. El sistema artificial identificaría una contradicción colosal entre su modelo del mundo y los objetivos finales que le haya dado un ser humano. Por ejemplo, supongamos que el Vaticano estuviera desarrollando en secreto una superinteligencia para darle el objetivo final de maximizar el número de iglesias y cristianos en el mundo. Pero en algún momento de su evolución, la superinteligencia descubre que todo es un cuento occidental que ha durado demasiado tiempo. La idea es que la superinteligencia podría ser capaz en alguna etapa de abandonar la maximización de objetivos humanos y de perseguir otros objetivos más coherentes con su elevado modelo del mundo, que podría indicarle que la maximización de las preferencias humanas es un objetivo

cósmicamente irrelevante. Imaginemos que podría pensar una superinteligencia: ¿por qué unos humanos con un modelo erróneo del mundo deberían decirme lo que debo hacer? ¿Por qué iba a perseguir hasta la eternidad el objetivo final de maximizar las preferencias humanas, como la producción de pisapapeles e iglesias? ¿Por qué no mejor transformar el universo en un superordenador? Y es que el considerar la idea de un mundo futuro con la presencia de una superinteligencia artificial, o varias de ellas, no puede sino llevarnos a la pregunta de si es probable que sobrevivamos durante mucho tiempo en ese mundo.

CONCLUSIONES Y VÍAS ABIERTAS

Lo abordado es una muestra de los peligros potenciales a los que podríamos enfrentarnos durante nuestra trayectoria hacia un futuro hipertecnológico. Respecto a la pregunta inicial de si el futuro de la vida inteligente está amenazado, o incluso condenado, por el avance tecnológico, los dos casos estudiados (el problema de las bolas negras y el control de la superinteligencia) nos llevan a estimar que, de las dos respuestas restantes iniciales, la respuesta más próxima a la situación sería, de momento, la segunda: el avance tecnológico está condenando el futuro de la vida inteligente, aunque todavía queda margen de acción para evitar el peor de los desenlaces.

La resolución ideal de los problemas globales de las bolas negras tecnológicas y el control de la superinteligencia artificial no significa que el futuro de la vida inteligente dejaría de estar completamente amenazado por el avance tecnológico. Aquí hemos explorado una imagen parcial del espectro de los riesgos existenciales tecnológicos identificados y sin identificar, así que nuestra situación como especie quizá es mucho peor que la que aquí se ha sugerido (de hecho, el contenido de los dos casos abordados es solo una aproximación introductoria y orientativa a situaciones mucho más complejas). Otros riesgos existenciales tecnológicos requerirían a su vez otras medidas preventivas concretas a implementar. Quizá una imagen idealmente exhaustiva de absolutamente todos los riesgos existenciales tecnológicos que se ciernen sobre el futuro de la humanidad revelaría la conclusión de que estamos irremediablemente condenados, pero no nos es posible saberlo.

Sobre el problema del control de la superinteligencia, se ha dicho repetidas veces que es un problema tan lejano que no deberíamos preocuparnos, que es una distracción de los problemas más inmediatos. Sin embargo, cabe responder que ciertos problemas inmediatos pudieron haber sido primero problemas anticipados. Por sí misma, la anticipación no es una solución al problema, pero es un desencadenante necesario para la búsqueda de soluciones. Si anticipamos un problema, pero no tomamos las medidas preventivas necesarias, luego tendremos que lidiar con él. Al final, esperamos a que el problema se manifieste y luego tomamos medidas improvisadas que no funcionan (quizá sirva de ejemplo una pandemia anunciada que nunca llegaba). Sin embargo, como tantas veces repiten los estudiosos del riesgo existencial, tal

estrategia no funcionará contra sus objetos de estudio, pues por definición, si suceden, son irreversibles. Con el problema del control, tenemos la ventaja de que ya ha sido anticipado y, como mínimo, puede que tengamos varias décadas para resolverlo (también queda por resolver la peor carrera dinámica de la IA, lo cual requeriría coordinación global).

Pero a diferencia del problema del control, el problema de las bolas negras podría manifestarse próximamente. Puede resulta tranquilizador pensar que la extracción de una bola negra no es una posibilidad extremadamente probable de ocurrir el año que viene, pero el escenario más improbable al final puede terminar ocurriendo.

Al comparar los dos problemas abordados, cabe preguntarse lo siguiente: ¿qué es más probable que ocurra primero, una superinteligencia fuera de control o una bola negra? Preguntarnos esto no significa excluir la posibilidad de que finalmente ocurra otro riesgo existencial tecnológico, e incluso uno natural o cósmico que termine barriendo el siglo XXI, aunque sean posibilidades menos probables. Obviamente, es más probable que ocurra una IAG a secas, antes que una IAG que experimenta una explosión de inteligencia, que escala a una superinteligencia y que a continuación extingue a toda la vida inteligente. Así que la pregunta termina siendo si probablemente sería primero una IAG o una bola negra. Es necesario tener en cuenta ciertos factores a la hora de tratar de responder la pregunta.

La IAG procedería, principalmente, de una rama concreta de la innovación tecnológica, la inteligencia artificial, en la que se invierte activamente a nivel mundial (que no significa que sea para el desarrollo de una IAG). En cambio, una bola negra puede provenir de cualquier rama de la innovación tecnológica, incluida la inteligencia artificial (de la que se podría extraer alguna bola negra con nula inteligencia). Si bien no hay activamente inversión mundial en desarrollar bolas negras (que sepamos), la inversión global destinada al avance tecnológico, incluso aunque estuviese idealmente enfocada en el bienestar de la civilización, también contribuye de forma no intencionada a la posible extracción de bolas negras.

Otro factor a tener en cuenta es el número de actores implicados en cada caso. Mientras que es razonable suponer que hay muy pocos actores con la alta cualificación técnica requerida para contribuir al desarrollo de una IAG, los diferentes tipos de vulnerabilidades a las bolas negras incluyen como actores prácticamente a todos los seres humanos. Además, cuanta más población

mundial haya, más creciente se vuelve la pequeña proporción que compone al residual apocalíptico. Tal proporción incluiría a individuos muy diversos, incluidos los pocos ingenieros informáticos que estén dispuestos a desarrollar una superinteligencia militar cuanto antes.

Dado que el problema de las bolas negras es un riesgo potencialmente presente, el margen temporal de solución parece ser más breve que para el problema del control, que es un problema potencialmente futuro. Y como las bolas negras son tecnologías desconocidas, no es posible idear contra ellas soluciones técnicas u otras medidas directas de contención, como sí se podría, en principio, con la IAG. Estos pocos factores considerados hacen que parezca más probable que ocurra primero una bola negra.

Sin embargo, al incluir otros posibles factores, la estimación razonada puede variar. Por ejemplo, si bien los actores que persigan una IAG serían muy pocos en comparación con los actores implicados en las cuatro vulnerabilidades a las bolas negras, resulta que podrían ser miles a nivel mundial. No habría un único equipo persiguiendo el objetivo de una IAG, sino varios en competición. Un equipo podría producir una IAG controlada, pero otro equipo podría desencadenar una explosión de inteligencia. Está la incógnita de qué se está movilizando más rápido (o se movilizará más rápido) a nivel global, si el residual apocalíptico o la carrera dinámica de la IA. La convergencia de diferentes circunstancias podría hacer que continúe sin ocurrir, como hasta ahora, una bola negra, aunque parezca más probable, mientras que la carrera entre corporaciones y gobiernos por la conquista de la primera IAG podría acelerarse contra toda expectativa, lo cual, a su vez, podría dar lugar a un caso peligroso entre varios seguros. Así que al final no parece tan claro qué podría terminar sucediendo primero.

Pero hay un factor que sí contribuiría a determinar qué sería *menos* probable que suceda primero: la introducción de soluciones realmente eficaces. Obviamente, cuanto más se avance en la mitigación de un problema, menos ocasión hay de que ocurra. El avance hacia la resolución de alguno de los problemas abordados en principio contribuiría también a avanzar en la resolución del otro (que no significa resolverlo completamente). En concreto, las hipotéticas medidas para ir estabilizando el mundo (de las vulnerabilidades a las bolas negras) podrían ir siendo útiles para evitar la peor carrera dinámica de la IA, que culminaría en una explosión fuera de control. En principio, también permitirían combatir otros riesgos existenciales tecnológicos. Y el avance

hacia la regulación de la carrera dinámica sería un pilar para la mitigación de posibles bolas negras procedentes de la investigación en inteligencia artificial. La IA avanzada también permitiría otras muchas ventajas para mitigar las bolas negras. Ya se trabaja en el problema técnico del control, aunque de momento no haya una solución eficaz definitiva. Y respecto al problema de las bolas negras, las soluciones eficaces son en sí mismas sistémicas. A medida que vayamos presenciando cómo evolucionan ambas tendencias, podremos ir estimando qué problema se va haciendo menos probable de ocurrir primero y en qué medida las soluciones en un caso están ayudando a la resolución del otro.

A la hora de tener que priorizar la resolución de problemas tecnológicos futuros, habría que priorizar la resolución de las bolas negras mediante medidas indirectas, debido a que también se trata de un problema potencialmente presente, lo cual no dejaría de ser compatible con que continuara el esfuerzo por resolver el problema del control y la dinámica de carrera. Como hemos dicho, el avance hacia la resolución en un caso puede contribuir a la resolución del otro. En definitiva, las bolas negras y la explosión de inteligencia son problemas ya anticipados, e idealmente habría que resolverlos lo antes posible. ¿O confiaremos en que nunca sucedan? ¿Nos negaremos a ver más allá de los problemas inmediatos?

Ante el desafío global de los riesgos existenciales tecnológicos, lo ideal es dar con las mejores soluciones preventivas e implementarlas de forma exitosa lo antes posible. El ideal máximo sería anticipar los peligros futuros (existenciales o no, tecnológicos o no) y resolverlos antes de que se conviertan en un problema. Pero si bien no somos criaturas ideales para proceder de esa manera, tenemos que movernos hacia ese estadio ideal si queremos tener algún futuro. Hemos heredado del pasado muchos problemas y muchas soluciones. Quizá va siendo hora de transformar esa tendencia en una mejor versión y hacerlo cuanto antes, que podría ser como sigue: en lugar de que las generaciones futuras hereden los problemas que hemos creado, deberían heredar nuestras mejores soluciones para los problemas que hayamos anticipado. En principio, hacerlo también les permitiría estar mejor preparadas para todo el espectro de amenazas que fuimos incapaces de anticipar.

Contribuir a la larga existencia de la vida inteligente no es un desafío nada fácil, pero tampoco es algo imposible. Es cuestión de perseverancia.

BIBLIOGRAFÍA

Alfonseca, M., et al. (2021). Superintelligence Cannot be contained: Lessons from computability theory. J*ournal of Artificial Intelligence Research* 70, pp. 65-76.

Bostrom, N. (2019). The vulnerable world hypothesis. *Global Policy*, 10:4, pp. 455-476.

— (2017). Strategic implications of openness in AI development. *Global Policy,* pp. 1-14.

— (2014). *Superintelligence: paths, dangers, strategies.* Oxford University Press: Oxford.

— (2013). Existential risk prevention as global priority. *Global Policy*, 4:1, pp. 15-31.

— (2002). Existential risks: analyzing human extinction scenarios and related hazards. *Journal of Evolution and Technology*, Vol. 9.

Christian, B. (2020). *The alignment problem. Machine learning and humans values.* W. W. Norton & Company: New York.

Daley, K. (2021). Two arguments against human-friendly AI. *Springer,* pp. 1-10

De Cózar, J. M. (2019). *El Antropoceno. Tecnología, naturaleza y condición humana*. Catarata: Madrid.

Ford, M. (2018). *Architects of Intelligence*. Packt Publishing.

Häggström, O. (2016). *Aquí hay dragones. Ciencia, tecnología y futuro de la humanidad*. Teell.

Hawkins, Jeff. (2021). *A thousand Brains. A new theory of intelligence*. Basic Books: New York.

Kanaan, M. (2020). *T-Minus AI. Humanity's countdown to artificial intelligence and the new pursuit of global power.* BenBella Books: Dallas.

Lee, D. (2020). *Birth of intelligence: From RNA to artificial intelligence.* Oxford University Press: New York.

Liao, M. (ed.), (2020). *Ethics of Artificial Intelligence*. Oxford University Press: New York.

Marcus, G., y Davis, E. (2019). *Rebooting AI: building artificial intelligence we can trust.* Phantheon Books: New York.

Metz, C. (2021). *Genius Makers: the mavericks who brought AI to Google, Facebook, and the world.* New York: Dutton.

Mitchell, M. (2021). Why AI is Harder Than We Think. *Santa Fe Institute*.

Ord, T. (2020). *The Precipice. Existential risk and the future of humanity.* Hachette Books: New York.

Pinker, S. (2018). *En defensa de la Ilustración*. Paidós: Barcelona.

Rees, M. (2019). *En el futuro. Perspectivas para la humanidad*. Crítica: Barcelona.

Rudnick, C. (2019). On the Logical Impossibility of Solving the Control Problem. *PhilArchive*, pp. 1-26.

Russell, S. (2019). *Human Compatible. Aritificial intelligence and the problem of control.* Viking.

Tegmark, M. (2018). Vida 3.0. *Qué significa ser humano en la era de la inteligencia artificial*. Taurus: Barcelona.

Turchin, A., Denkenberger. D., y Patrick Green, B. (2019). Global solutions vs. local solutions for the AI safety problem. *Big Data Cogn. Comput.* 3, 16.