La forma correcta de preocuparse

14 de mayo de 2021

Daron Acemoglu[1]

Después de un año en el que COVID-19 ha suspendido la vida económica normal en todo el mundo, la humanidad ha adquirido una nueva apreciación del riesgo. Pero el simple hecho de reconocer las amenazas potenciales es simplemente el comienzo del proceso; el verdadero desafío consiste en decidir qué problemas merecen nuestra atención y en qué orden.

Toby Ord, El precipicio: riesgo existencial y el futuro de la humanidad, Bloomsbury, Londres, 2020.

CAMBRIDGE - El reinado de los dinosaurios terminó hace 65 millones de años por un asteroide que se estrelló contra lo que ahora es la ciudad de Chicxulub en México. Aunque este trozo de roca y metal no era particularmente grande, probablemente de unos diez kilómetros (seis millas) de diámetro, golpeó la Tierra a más de 60.000 kilómetros por hora (37.000 millas por hora), generando una explosión miles de millones de veces mayor que la de la bomba atómica cayó sobre Hiroshima y mató a toda la vida en un radio de 1.000 kilómetros.

Más inquietantemente, la explosión envió una enorme nube de polvo y cenizas a la atmósfera superior, bloqueando el sol durante los próximos años. Esto impidió la fotosíntesis y condujo a temperaturas muy reducidas, razón por la cual los científicos calculan que fue este polvo atmosférico y aerosoles de sulfato los que finalmente mataron a los dinosaurios y muchas otras especies.

Si un asteroide o cometa similar se estrellara contra la Tierra hoy, causaría otro evento de extinción masiva, acabando con la mayoría de las especies y la civilización humana tal como la conocemos. Esta posibilidad lejana es un ejemplo de un riesgo existencial natural: un evento no causado por humanos que lleva a la extinción o casi a la extinción de nuestra especie.

Pero también existen riesgos existenciales antropogénicos, creados por humanos. Como sostiene el filósofo de la Universidad de Oxford Toby Ord en su nuevo libro que invita a la reflexión, El precipicio: riesgo existencial y el futuro de la humanidad, son estos riesgos los que más deberían preocuparnos ahora y en el próximo siglo.

Ord reconoce que la ciencia y la tecnología son las herramientas más potentes de la humanidad para resolver problemas y lograr la prosperidad. Pero nos recuerda que siempre existen peligros asociados con tal poder, particularmente cuando se coloca en las manos equivocadas o se ejerce sin preocuparse por las consecuencias no deseadas a largo plazo.

---

[1] Daron Acemoglu, profesor de economía en el MIT, es coautor (con James A. Robinson) de Why Nations Fail: The Origins of Power, Prosperity and Poverty y The Narrow Corridor: States, Societies, and the Fate of Liberty.

Más concretamente, Ord sostiene que el riesgo existencial antropogénico ha alcanzado un nivel alarmantemente alto, porque hemos desarrollado herramientas capaces de destruir a la humanidad sin la sabiduría necesaria para reconocer el peligro en el que nos encontramos. Señala que el eminente astrónomo del siglo XX Carl Sagan emitió una advertencia similar en su libro de 1994, Pale Blue Dot, escribiendo:

“Muchos de los peligros a los que nos enfrentamos de hecho surgen de la ciencia y la tecnología, pero, más fundamentalmente, porque nos hemos vuelto poderosos sin ser proporcionalmente sabios. Los poderes que alteran el mundo que la tecnología nos ha entregado en la cabeza ahora requieren un grado de consideración y previsión que nunca antes se nos había pedido”.

Para Ord, esta brecha entre el poder y la sabiduría podría decidir el futuro de la humanidad. Por un lado, podríamos desaparecer por completo o sufrir un colapso que borre la mayoría de las señas de identidad de la civilización (desde las vacunas y los antibióticos hasta el arte y la escritura). Pero, por otro lado, Ord ve en la humanidad el potencial para el florecimiento a largo plazo a escala cósmica: con sabiduría e ingenio tecnológico, los humanos podrían sobrevivir a este planeta y lanzar nuevas civilizaciones a través del espacio.

Esta visión de gran alcance del florecimiento pesa mucho en los cálculos de Ord, porque reconoce que puede que no haya otras formas de vida inteligente en el universo. Si estamos realmente solos, un evento de extinción masiva que acabó con todos en este planeta también eliminaría todo el potencial de existencia inteligente y con propósito en todas partes.

Con base en este razonamiento, Ord llega a lo que matemáticos y economistas llamarían un "orden de preferencia lexicográfico". En una situación en la que nos preocupan los criterios múltiples, un orden lexicográfico asigna una importancia abrumadora a un criterio para proporcionar claridad cuando se comparan dos opciones. Por ejemplo, en un orden lexicográfico entre comida y refugio, siempre se preferiría la opción que ofrezca más comida, independientemente de cuánto más refugio ofrezca la otra opción.

La postura filosófica de Ord es equivalente a un orden lexicográfico porque implica que debemos minimizar el riesgo existencial, sean cuales sean los costos. Un futuro en el que el riesgo existencial se ha minimizado triunfa sobre cualquier futuro en el que no se haya minimizado, independientemente de cualquier otra consideración. Después de establecer esta jerarquía básica, Ord procede con una visión experta de los diferentes tipos de riesgo existencial antropogénico, concluyendo que la mayor amenaza proviene de una superinteligencia artificial que ha evolucionado más allá de nuestro control.

CUANDO EL PROGRESO NO ES PROGRESO

Se puede fechar el riesgo existencial impulsado por la ciencia al menos con las reacciones nucleares en cadena controladas que permitieron las armas atómicas. Probablemente Ord tenga

razón en que nuestra sabiduría (social) no ha aumentado desde este fatídico acontecimiento, que culminó antes con los bombardeos de Hiroshima y Nagasaki. Aunque hemos establecido algunas instituciones, herramientas reguladoras, normas y otros mecanismos de internalización para asegurarnos de no hacer un mal uso de la ciencia, nadie diría que son suficientes.

Ord sugiere que el marco institucional inadecuado de hoy puede ser un fenómeno temporal que podría abordarse a su debido tiempo, siempre y cuando sobrevivamos al próximo siglo más o menos. “Porque nos encontramos en un momento crucial en la historia de nuestra especie”, escribe. “Impulsado por el progreso tecnológico, nuestro poder ha crecido tanto que por primera vez en la larga historia de la humanidad, tenemos la capacidad de destruirnos a nosotros mismos ...” Y, de hecho, al escribir su libro, Ord “aspira a empezar a cerrar la brecha entre nuestra sabiduría y poder, permitiendo a la humanidad una visión clara de lo que está en juego, para que tomemos las decisiones necesarias para salvaguardar nuestro futuro ”.

Sin embargo, no veo evidencia de que esto sea realmente factible. Tampoco hay ninguna señal de que nuestra sociedad y nuestros líderes hayan mostrado sabiduría cuando se trata de controlar el poder destructivo de la tecnología.

Sin duda, uno podría argumentar a favor del optimismo de Ord sobre la base de lo que el sociólogo alemán Norbert Elias llamó el famoso "proceso civilizador". Según Elías, el proceso de desarrollo económico y el surgimiento de instituciones estatales para la resolución de conflictos y el control de la violencia desde la Edad Media han llevado a la adopción de modales y comportamientos propicios a la convivencia en sociedades de masas. El caso matizado de Elias de por qué las personas en las economías avanzadas se han vuelto menos violentas y más tolerantes fue popularizado recientemente por el psicólogo cognitivo y lingüista de la Universidad de Harvard Steven Pinker en su libro superventas The Better Angels of Our Nature: The Decline of Violence in History and Its Causes. Ambos autores ofrecen argumentos de por qué deberíamos seguir esperando un fortalecimiento de las normas e instituciones necesarias para controlar el mal uso de la ciencia y la tecnología.

Pero incluso si ese proceso civilizador actúa sobre las normas de comportamiento individuales y las relaciones sociales de manera más amplia, no parece haber afectado a muchos líderes políticos o científicos y tecnólogos. El proceso de civilización debería haber estado en pleno apogeo en la primera mitad del siglo XX; y, sin embargo, el químico Fritz Haber, ganador del Premio Nobel, utilizó con entusiasmo su conocimiento científico para inventar y luego vender armas químicas al ejército alemán en la Primera Guerra Mundial.

El impacto del proceso civilizador tampoco fue muy evidente en el pensamiento de los líderes estadounidenses que ordenaron los ataques contra Hiroshima y Nagasaki, o en las actitudes de otros líderes políticos que abrazaron con entusiasmo las armas nucleares después de la Segunda Guerra Mundial. Algunos pueden encontrar esperanza en el hecho de que no hemos tenido una repetición de la Primera Guerra Mundial o la Segunda Guerra Mundial en los últimos 75 años.

Pero esta visión optimista ignora muchos fallos cercanos, entre ellos la Crisis de los Misiles Cubanos en 1962 (el episodio con el que Ord abre su libro).

Se pueden identificar muchos más ejemplos que contradicen la idea de que nos estamos volviendo más "civilizados", y mucho menos mejor en el control de los riesgos antropogénicos o en el cultivo de la sabiduría colectiva. En todo caso, controlar nuestro mal comportamiento y adaptarnos a los constantes cambios provocados por los descubrimientos científicos y la innovación tecnológica seguirá siendo una lucha constante.

Esto plantea problemas para el resto del argumento de Ord. ¿Por qué debería darse una prioridad suprema a tratar de eliminar los riesgos existenciales futuros sobre todos los demás esfuerzos para mejorar los males y el sufrimiento que nuestras elecciones actuales están generando ahora y en el corto plazo?

En aras del argumento, supongamos que pudiéramos reducir significativamente la probabilidad de nuestra propia extinción esclavizando a la mayoría de la humanidad durante los próximos siglos. Bajo el orden lexicográfico de Ord, tendríamos que elegir esta opción, porque minimiza el riesgo existencial y al mismo tiempo preserva el potencial de la humanidad para florecer plenamente en algún momento del futuro lejano.

Este argumento no convencerá a todo el mundo. Cuenteme entre los que no han sido persuadidos.

¿LA ERA DE LAS MÁQUINAS DEMÓNICAS?

Para aclarar aún más la elección, considere el principal riesgo existencial en el que se centra Ord: el posible uso indebido de la inteligencia artificial. Ord estima que hay una posibilidad entre seis de que la humanidad sea presa de una superinteligencia malvada (que él llama, eufemísticamente, "IA no alineada") en los próximos 100 años. Por el contrario, su riesgo existencial estimado para la humanidad por el cambio climático es uno en 1.000, y uno en un millón en el caso de colisiones con asteroides o cometas.

Incluso si muchos otros expertos no asignarían una probabilidad tan alta a la amenaza de la superinteligencia, Ord no es el único que se preocupa por las implicaciones a largo plazo de la investigación de la IA. De hecho, estas preocupaciones se han convertido en algo común entre muchas luminarias de la tecnología, desde Stuart Russell de la Universidad de California en Berkeley hasta el fundador de Microsoft, Bill Gates, y el fundador de Tesla, Elon Musk.

Todas estas cifras creen que, a pesar de los riesgos existenciales, la IA traerá muchos beneficios netos. Pero si bien Ord está lo suficientemente bien informado sobre estos debates como para saber que incluso esta última propuesta es en realidad bastante inestable, su postura lexicográfica lo lleva a ignorar la mayoría de los riesgos no existenciales asociados con la IA.

Pero si uno acepta que nuestro alcance de atención es finito, esta ponderación de prioridades es problemática. Mi propia evaluación es que la probabilidad de que surja una superinteligencia en el corto plazo es baja, y que el riesgo de que una superinteligencia maligna destruya nuestra civilización es aún menor. Como tal, preferiría que el debate público se centre mucho más en los problemas que la IA ya está creando para la humanidad, en lugar de los intrigantes pero improbables riesgos de cola.
VOLVER AL AHORA

Como he argumentado aquí y en otros lugares, la trayectoria actual del diseño y despliegue de la IA nos está llevando por mal camino, causando una amplia gama de problemas inmediatos (aunque prosaicos). Lejos de ser inevitables o reflejar alguna lógica inherente de la tecnología, estos problemas reflejan elecciones que están tomando (e impuestas sobre nosotros) las grandes empresas tecnológicas, y específicamente por un pequeño grupo de ejecutivos, científicos y tecnólogos dentro de estas empresas (o dentro de sus orbita).

Uno de los problemas más visibles que está causando la IA es la automatización incesante, que desplaza a los trabajadores, aumenta la desigualdad y aumenta el espectro del desempleo futuro para grandes sectores de la fuerza laboral. Peor aún, la obsesión por la automatización se ha producido a expensas del crecimiento de la productividad, porque ha llevado a ejecutivos y científicos a pasar por alto usos más fructíferos y complementarios a las personas de la tecnología innovadora.

La IA también se está diseñando y utilizando de otras formas problemáticas, ninguna de las cuales inspira esperanza para el progreso moral de la humanidad. La política democrática ha sido contaminada no solo por una explosión de información errónea amplificada algorítmicamente, sino también por las nuevas tecnologías de inteligencia artificial que han empoderado a los gobiernos y las empresas para monitorear y manipular los comportamientos de miles de millones de personas.

Este desarrollo representa un doble golpe. La política democrática es el medio principal por el cual una sociedad puede frenar la mala conducta de las élites políticas y económicas, pero es precisamente este proceso el que está siendo socavado. Si no podemos responsabilizar a las élites por el daño que están causando porque la democracia misma se ha visto afectada, ¿cómo podemos escapar de nuestra situación actual?

No estamos indefensos. Los costos que está infligiendo la IA pueden abordarse porque, a diferencia de los riesgos existenciales en los que se centra Ord, son tangibles y fáciles de reconocer. Pero primero hay que llamar más la atención sobre el problema con el fin de generar presión en gobiernos y empresas para que reconozcan los riesgos que se están materializando ahora. Además, un sector tecnológico que se inclina por la automatización y la manipulación y vigilancia antidemocráticas no es una buena base sobre la cual abordar los riesgos a largo plazo.

Aunque no debemos descartar más riesgos especulativos para la humanidad, no podemos permitirnos ignorar las amenazas que tenemos frente a nosotros. Puede que lleguemos a un precipicio en algún momento, pero ya nos deslizamos por una pendiente resbaladiza.