FRENAR A SILICON VALLEY

Cómo las Big Tech se aprovechan de nosotros y cómo la Inteligencia Artificial puede empeorarlo

GARY MARCUS

Traducción de Marc Figueras

Shackleton books

Frenar a Silicon Valley
Título original: *Taming Silicon Valley*

Shackleton
—books—

@Shackletonbooks
shackletonbooks.com

Realización editorial: La Letra, S. L.
Diseño de cubierta: Pau Taverna

ISBN: 978-84-1361-633-9
Depósito legal: B. 9123-2025
Impreso por EGEDSA (España)

A mis hijos, Chloe y Alexander;
que la IA pueda hacer de su mundo,
y del de todos, un lugar mejor.

El propósito [del gobierno] debería ser evitar [...] abusos [...] y no esperar hasta que se produzcan.

Presidente Theodore Roosevelt,
discurso al Congreso de Estados Unidos, 1907

La desregulación gubernamental [...] ha tenido su oportunidad. Un experimento natural de cuarenta años que consiste en dejar que las industrias se autorregularan ha llevado a los desafíos a los que nos enfrentamos hoy en las industrias digitales: concentración en cada sector digital, intrusiones masivas de la privacidad y una esfera de información pública distorsionada.

Mark MacCarthy, *Regulating Digital Industries*, 2023

Los políticos y los ciudadanos no deberían hacerse ilusiones de que las empresas privadas de IA actúen en pro del interés público.

Marietje Schaake, *The Premature Quest for International AI Cooperation*, 2023

Hay riesgo de que tecnologías que no comprendemos bien proliferen irreversible e incontrolablemente. En palabras de Jeff Bezos, una «puerta unidireccional», una decisión que apenas permite marcha atrás.

Tim O'Reilly, *You Can't Regulate What You Don't Understand*, 2023

No tengo que decirles que las cosas están mal porque todo el mundo lo sabe. [...] Yo no voy a dejarles en paz. [...] Lo único que sé es que tienen ustedes que montar en cólera. Tienen que decir: «¡Soy un ser humano, maldita sea! ¡Mi vida tiene un valor!». Quiero que ahora se levanten, que se levanten de sus sillones, que se levanten todos y vayan a sus ventanas, que las abran y que saquen la cabeza gritando: «¡Estoy más que harto, y no quiero seguir soportándolo!» [...] Luego pensaremos lo que hay que hacer.

Howard Beale, presentador de ficción interpretado por
Peter Finch en la película *Network, un mundo implacable*, 1976

CONTENIDO

Introducción: si no cambiamos de rumbo, nos quedaremos sin sociedad tal como la conocemos

Muévete rápido y rompe cosas.

MARK ZUCKERBERG, lema interno de Facebook, 2012

No adoptamos una visión lo bastante amplia de nuestra responsabilidad.

MARK ZUCKERBERG, dirigiéndose al Senado de Estados Unidos, 2018

La IA generativa tiene, sin duda, muchos usos positivos y creativos, y sigo creyendo en su potencial para hacer el bien. Pero si echamos una mirada al año pasado, parece claro que todos los beneficios que hemos visto hoy se han obtenido con un alto coste. A menos que los que están en el poder tomen medidas, y lo hagan pronto, el número de víctimas que pagarán este coste no hará más que aumentar.

CASEY NEWTON, *Plataformer*, 2024

Sus científicos estaban tan obsesionados con saber si podían, que no se pararon a pensar si deberían.

IAN MALCOLM, matemático de ficción interpretado por Jeff Goldblum en *Parque Jurásico*, 1993

Si «muévete rápido y rompe cosas» era el eslogan no oficial de la era de las redes sociales, ¿cuál es el eslogan no oficial de la era de la inteligencia artificial generativa?

Podemos estar casi seguros de que debe de ser el mismo, aunque esta vez las cosas que se rompan pueden ser muchísimo peores. La desinformación automatizada amenaza con perturbar las elecciones en todo el mundo y ya hay grupos enteros de personas que han empezado a perder sus medios de sustento a manos de las grandes empresas tecnológicas, esas que hablan de un «futuro positivo» mientras hacen todo lo posible por arrinconar a la gente. Los artículos escritos con IA generativa están contaminando el mundo científico y han comenzado a difamar a personas.[1]

La era de las redes sociales ha destruido la privacidad, ha polarizado la sociedad, ha acelerado la guerra informativa y ha llevado a muchas personas al aislamiento y la depresión; una demanda reciente, de hecho, plantea que empresas como Meta, Reddit y 4chan «se benefician del material racista, antisemita y violento que se muestra en sus plataformas para maximizar la participación de los usuarios».[2] Al mismo tiempo, las plataformas de redes sociales han creado un nuevo modelo de negocio: el capitalismo de la vigilancia. La venta al mejor postor de anuncios dirigidos que aprovechan los datos personales de la gente (postores entre los que no solo se encuentran los anunciantes tradicionales, sino también estafadores, delincuentes y operativos políticos) ha enriquecido desorbitadamente a ciertas personas (como el director general de Meta, Mark Zuckerberg) y les ha otorgado un poder excesivo sobre nuestras vidas. Así, las decisiones de un solo líder tecnológico no electo podrían influir con facilidad en el resultado de unas elecciones; y, a escala menor, también podrían ayudar a una pequeña empresa o por el contrario llevarla a la quiebra.

En cuanto las empresas de redes sociales han descubierto que la desinformación fomenta la interacción, lo que a su vez conduce a un considerable aumento de las ganancias, y han aprendido que cuanto más tiempo

alguien utiliza sus servicios, más dinero se embolsan, nació la «economía de la atención».

De este modo, los medios de comunicación contrastados que aspiraban a cierta neutralidad dieron paso a un conglomerado alimentado por la IA, diseñado para espolear la ira y en el que los clics son lo único que importa. Esto, a su vez, ha dado lugar a incontables cajas de resonancia, repletas de cólera y a menudo sometidas a una gran distorsión, en las que cualquiera con una opinión ridícula puede encontrar a decenas de miles o incluso millones de otras opiniones que la refuerzan. Los argumentos intelectuales han dado paso a textos de 140 caracteres, a *soundbites* o cortes de audio y a vídeos de TikTok; en definitiva, a una cultura de «cultivo de la interacción». Intereses extranjeros han aprendido a usar las redes sociales para perturbar elecciones presidenciales en Estados Unidos, por ejemplo. Los mercaderes de la propaganda han encontrado en las redes sociales una herramienta perfecta y fácilmente corruptible; las empresas de redes sociales les han devuelto el favor, beneficiándose de esa propaganda y distribuyéndola. Los usuarios se han convertido en peones. Una ley del Congreso de Estados Unidos, el apartado 230 de la Communications Decency Act (Ley de Respetabilidad en las Comunicaciones) de 1996, ha empeorado aún más las cosas, al dejar a las plataformas de redes sociales casi sin responsabilidad por sus acciones.[3] Hoy en día, hay toda una generación que ha crecido sin conocer nada más. (Las aplicaciones de citas, por ejemplo, en realidad funcionan con principios similares de economía de la atención: su objetivo no es tanto ayudarte a encontrar una pareja como hacer que sigas usando sus aplicaciones, lo que en última instancia hace que muchas personas se sientan más solas que cuando empezaron; ya que, si encuentras el amor, dejas de ser un cliente.)

A menos que nos pongamos manos a la obra como sociedad, la IA generativa (sistemas como ChatGPT) empeorará aún más esta situación actual, con consecuencias que van desde la pérdida de los últimos vestigios de privacidad hasta una mayor polarización de la sociedad, y creará un conjunto de nuevos problemas que podrían eclipsar con facilidad todo lo que ha ocu-

rrido hasta el momento. Aumentarán los desequilibrios de poder, donde los líderes tecnológicos no electos controlarán amplísimos aspectos de nuestras vidas; las elecciones justas e imparciales podrían convertirse en algo del pasado; la desinformación automatizada podría destruir lo que queda de la democracia; los sutiles sesgos que están integrados en *chatbots* controlados por unos pocos individuos selectos moldearán las opiniones de la mayoría… Además, su impacto ambiental puede ser inmenso. Por otro lado, Internet ya está empezando a contaminarse con grandes cantidades de basura generada por IA, lo que lo hace cada día menos confiable.

En el peor de los casos, una IA poco honesta e insegura podría provocar catástrofes masivas, desde el caos en las redes de distribución eléctrica hasta guerras accidentales o la pérdida de control de flotas de robots (drones). Además, muchas personas podrían perder sus empleos. Los modelos de negocio de la IA generativa ignoran las leyes de propiedad intelectual, la democracia, la seguridad del consumidor y el impacto en el cambio climático. Y como se ha extendido con tanta rapidez y con tan poca supervisión, la IA generativa se ha convertido en un enorme experimento sobre toda la población, sin ningún tipo de control.

Cada día, y cada vez con mayor frecuencia, como experto en estudios de futuro y en inteligencia artificial y como padre de niños pequeños, me pregunto: ¿Estamos haciendo lo correcto? ¿La inteligencia artificial nos matará? ¿O nos salvará? Y, lo más importante, ¿ayudará la IA a la humanidad o la perjudicará?

La única respuesta intelectualmente honesta es: nadie lo sabe. La IA ha llegado para quedarse. Ya está transformando nuestra sociedad, tanto de maneras positivas como negativas; en las próximas décadas, sus efectos serán formidables y transformarán prácticamente todo lo que hacemos. Sobre esto, no caben muchas dudas.

Los beneficios *posibles* son enormes, tanto como Silicon Valley quiere que creamos. Es cierto que la IA puede revolucionar la ciencia, la medicina y la tecnología y conducirnos a un mundo de abundancia y mejor salud. Una esperanza (una que todavía no he perdido del todo) es que la IA pueda ayudarnos a avanzar en la ciencia y la medicina y a superar los retos y amenazas actuales, como el cambio climático o las enfermedades neurodegenerativas. DeepMind, una empresa que en su día fue independiente y que ahora forma parte de Google, tuvo una vez el famoso y acaso ingenuo eslogan: «Primero solucionemos la inteligencia y luego solucionemos todo lo demás».[4] Todavía hay alguna posibilidad de que la IA, si logramos desarrollarla bien (y yo sostengo que hasta ahora no lo hemos hecho), pueda ser un elemento transformador y con un efecto positivo neto, que nos ayude en tareas como el descubrimiento de fármacos, la agricultura y la ciencia de los materiales. Y que quede bien claro desde el principio que *yo quiero ver ese mundo*. Trabajo en IA cada día (y me peleo con quienes desean tomar atajos arriesgados) no porque quiera acabar con ella, sino porque quiero que esté a la altura de todo su potencial.

Ahora bien, seamos sinceros, en estos momentos no estamos yendo por el mejor camino, ni técnica ni moralmente. La codicia ha sido un factor importante y la tecnología que tenemos ahora es prematura, está sobrevalorada y resulta problemática; la que tenemos no es la mejor IA que podríamos concebir y, sin embargo, se ha lanzado y puesto a disposición de la gente a toda prisa. Si se desarrolla sin cuidado, no sería descabellado que la IA nos condujera al desastre.

Lo más probable es que el efecto neto de la IA se sitúe en algún punto intermedio (con ciertas ventajas y desventajas), pero no sabemos con precisión dónde acabará cayendo. Si seguimos potenciando la IA generativa, el enfoque de moda, podemos toparnos con problemas; si podemos encontrar enfoques más fiables y un liderazgo más responsable que el que existe la actualidad, no hay ninguna duda de que el valor de la IA para la sociedad aumentará.

En cualquier caso, lo más relevante es que *no somos unos pasajeros pasivos en este viaje*. El resultado final no es un destino predeterminado. La IA no tiene por qué convertirse en el monstruo del doctor Frankenstein. Como sociedad, todavía podemos plantar cara e insistir para que la inteligencia artificial sea justa, equitativa y confiable. No se trata de ahogar la innovación, pero si queremos llegar a un futuro positivo para la IA sí que se necesitan controles y contrapesos.

El objetivo de este libro es exponer los pasos concretos que podemos dar para lograrlo y cómo, entre todos, podemos marcar la diferencia.

De un modo u otro, he estado implicado en cuestiones de IA desde que era pequeño. Aprendí a programar cuando tenía ocho años. A mis quince años, escribí un traductor latín-inglés en el lenguaje de programación LOGO, en un Commodore 64, y aproveché este proyecto para ingresar anticipadamente a la universidad al año siguiente, saltándome los dos últimos años de secundaria. Mi tesis doctoral versó sobre la fascinante, y aún no resuelta, cuestión de cómo los niños adquieren el lenguaje, con la mirada puesta en un tipo de sistema de IA conocido como *red neuronal* (el antepasado de la IA generativa de hoy en día), lo que me preparó bien para entender la generación actual de IAs.

A los cuarenta años, inspirado por el éxito de DeepMind, fundé una empresa de inteligencia artificial y aprendizaje automático llamada Geometric Intelligence, que acabé por vender a Uber. La inteligencia artificial ha sido buena para mí y quiero que sea buena para todo el mundo.

He escrito este libro desde la perspectiva de alguien que ama a la IA y que desea desesperadamente que triunfe, pero también como alguien que está desilusionado y profundamente preocupado por el rumbo que están tomando las cosas. El dinero y el poder han hecho descarrilar a la IA de su misión original; al fin y al cabo, ninguno de nosotros se sumergió en la IA con la intención de vender anuncios o generar noticias falsas. No estoy en

contra de la tecnología y no creo que debamos dejar de construir sistemas de IA, pero no podemos seguir así. En este momento, estamos desarrollando el tipo equivocado de inteligencia artificial, una IA (y un complejo industrial de IA); en el que no podemos confiar. Mi mayor esperanza es que consigamos encontrar un camino mejor.

Tal vez mi nombre te suene como el de esa persona que se atrevió a desafiar al director general de OpenAI, Sam Altman, cuando testificamos juntos ante el Senado de Estados Unidos; los dos tomamos juramento el 16 de mayo de 2023 y prometimos decir la verdad. Mi objetivo aquí es contar la verdad acerca de la manera en que las grandes empresas tecnológicas nos están explotando cada vez más. Y explicar cómo es que la IA está poniendo en riesgo casi todo lo que apreciamos, desde la privacidad hasta la democracia y nuestra propia seguridad, a corto, medio y largo plazo. Por último, también me gustaría ofrecer mis mejores cavilaciones sobre lo que podemos hacer al respecto.

En esencia, veo cuatro problemas:

1) La forma concreta de tecnología de inteligencia artificial en la que todo el mundo se centra en estos momentos, la IA generativa, es absolutamente defectuosa. Sus sistemas han demostrado una y otra vez ser del todo impasibles a la diferencia entre verdades y trolas (lo que en inglés denominan *bullshit*).* Los modelos genera-

* En general, se denomina *slop* (en español a veces traducido como *sandeces*, *bazofia* o, simplemente, *basura*) todo el contenido de baja calidad (chapucero o no), generado por un sistema de IA, sin que este sea necesariamente erróneo. El término *bullshit* (a veces traducido como *trolas* o *gilipolleces*) suele hacer referencia más específicamente a contenido falso o equívoco, que el sistema de IA presenta como un hecho verdadero. En este sentido, es sinónimo de lo que también se conoce como *hallucinations* (*alucinaciones*), aunque algunos autores prefieren usar *bullshit* en el sentido de trolas expresadas con absoluta indiferencia por la verdad, sin ninguna intención de engañar, mientras que el término *hallucination* puede dar la idea de que el sistema está representando el mundo de modo incorrecto (cosa que un sistema de IA generativa no hace, porque no representa el mundo de ninguna manera). Véase, entre otros, Hicks, M.T., Humphries, J. y Slater, J. «ChatGPT is bullshit». *Ethics Inf Technol*, 26, nº 38 (2024); <https://doi.org/10.1007/s10676-024-09775-5>. (*N. del t.*)

tivos, aprovechando una frase del ámbito militar, «se equivocan a menudo y nunca dudan». Uno puede confiar en que el ordenador de las naves de *Star Trek* dé respuestas sensatas a preguntas razonables; la IA generativa, en cambio, es impredecible. Y lo que es peor, a menudo es lo bastante acertada como para hacernos caer en la complacencia: aunque siempre se cuelen errores; casi nadie la trata con la desconfianza que merecería. Algo que tuviera la fiabilidad de los ordenadores de *Star Trek* podría cambiar el mundo; sin embargo, lo que tenemos ahora es un caos, atractivo pero poco fiable. Y muy poca gente está dispuesta a admitir esta incómoda verdad.

2) Las empresas que en estos momentos están desarrollando IA se llenan la boca con declaraciones de una «IA responsable», pero sus palabras no se corresponden con sus acciones. En realidad, la IA que están construyendo no es lo bastante responsable; y si no se controla a estas empresas, es poco probable que lo sea algún día.

3) Al mismo tiempo, la IA generativa está sobrevaloradísima en relación con las realidades de lo que ofrece o puede ofrecer. Las empresas que desarrollan IA siguen pidiendo permisos y aquiescencias (como exenciones de la Ley sobre el Derecho de Autor) con el argumento de que algún día, de alguna manera, salvarán a la sociedad, a pesar de que hasta la fecha sus contribuciones tangibles han sido limitadas. Con demasiada frecuencia, los medios de comunicación se tragan el mito del mesías de Silicon Valley; los empresarios exageran porque así es más fácil recabar dinero y casi nadie rinde cuentas por promesas incumplidas. Como ocurre con las criptomonedas, las vagas promesas de beneficios futuros, que tal vez nunca se cumplan, no deberían despistar a los ciudadanos y a los políticos de la realidad de los perjuicios actuales.

4) Nos encaminamos hacia una especie de oligarquía de la IA con demasiado poder, en un eco aterrador de lo que ha sucedido con las redes sociales. En Estados Unidos (y en muchos otros lugares, con la notable excepción de Europa), las grandes empresas tecnológicas son las que tienen la sartén por el mango, mientras que los gobiernos han hecho muy poca cosa para refrenarlas. La gran mayoría de los estadounidenses quiere una regulación seria de la IA mientras cae la confianza en ella,[5] pero hasta el momento el Congreso de Estados Unidos no ha estado a la altura de las circunstancias.

Como dije ante el Subcomité Judicial del Senado de Estados Unidos sobre Supervisión de la IA en mayo de 2023, todo esto ha llevado a una «tormenta perfecta de irresponsabilidad corporativa, aplicación a gran escala, carencia de regulación suficiente y falta de confiabilidad inherente».

En este libro intentaré desentrañar todos estos temas y plantear qué es lo que nosotros, como individuos y como sociedad, podemos y debemos exigir.

El eslogan informal de Google, que se remonta más o menos al año 2000, solía ser «No seas malvado» (*Don't be evil*).[6] El objetivo original de OpenAI (2015) era «fomentar la inteligencia digital de manera que tenga mayores probabilidades de beneficiar a la humanidad en su conjunto, sin estar limitada por la necesidad de generar un retorno económico. […] Sin obligaciones financieras, podemos concentrarnos mejor en un impacto humano positivo».[7] Nueve años después, OpenAI trabaja para Microsoft en muchos aspectos, y esta última se lleva aproximadamente la mitad de los primeros 92.000 millones de dólares de sus ganancias; es más, un acuerdo de licencia otorga a Microsoft acceso privilegiado al trabajo de OpenAI.[8]

Diría que el mayor cambio en la cultura corporativa en torno a la IA se

produjo con ChatGPT, lanzado a finales de noviembre de 2022, y con su repentina e inesperada popularidad: en cuestión de meses, cien millones de personas comenzaron a usarlo. Casi de la noche a la mañana, la IA pasó de ser un proyecto de investigación a una potencial gallina de los huevos de oro. Tal como analizaré más adelante, en ese momento desapareció buena parte de todo aquello que se decía sobre una «IA responsable».

Es cierto que nunca me han gustado los efectos corrosivos de las redes sociales y, en particular, las reiteradas violaciones de la privacidad por parte de Facebook (ahora Meta) y su falta de inquietud por el efecto que sus productos tienen en el mundo. Pero hasta principios de 2023, tenía la sensación de que empresas como Microsoft y Google veían la IA de forma sensata. Microsoft, por ejemplo, hacía años que hablaba de tecnología responsable; de hecho, su presidente, Brad Smith, incluso había escrito todo un libro al respecto. En 2016, Microsoft lanzó un *chatbot* llamado Tay que metió la pata hasta el fondo, y parece que aprendieron de ello: Tay comenzó a repetir consignas nazis en menos de un día después de su lanzamiento; la juiciosa respuesta de Microsoft fue apagar Tay de inmediato. Google mantuvo internamente durante años productos como un *chatbot* llamado LaMDA, porque les preocupaba que no fuera lo suficientemente confiable.[9] Hubo una gran cantidad de investigación fascinante entre bambalinas, pero su aplicación a una audiencia mundial fue limitada.

Unos meses después de que saliera ChatGPT, las cosas habían cambiado de un modo palpable. En febrero de 2023, Sydney, el nuevo *chatbot* de Microsoft, alimentado por el GPT-4 de OpenAI, instó al redactor de *The New York Times* Kevin Roose a que se divorciara, «declarando, así por las buenas, que lo amaba» y luego trató de convencerlo de que dejara a su esposa y «estuviera con él en su lugar».[10] La reacción de la prensa fue airada y yo mismo esperaba que Microsoft retirara temporalmente a Sydney del mercado, pero esta vez no le importó; hicieron algunos cambios menores y siguieron adelante. Más tarde se supo que Microsoft había sido

advertido de que podían pasar este tipo de cosas, pero continuaron como si nada.[11] Unas semanas después, despidieron a todo un equipo de investigadores de IA responsable.[12] En una entrevista con *The Verge* ese mismo mes, el director general de Microsoft, Satya Nadella, dijo sobre Google: «Quiero que la gente sepa que los hicimos bailar».[13]

Y vaya si los hizo bailar, tal vez para desgracia de todos nosotros. Antes de ese momento, Google había ralentizado durante mucho tiempo la IA poco fiable, manteniendo en el ámbito interno los proyectos piloto de inteligencia artificial en lugar de lanzarlos al público de modo prematuro. En ese instante, las normas de comportamiento cambiaron abruptamente, para peor.

Microsoft siguió usando sus *chatbots* después de que uno de ellos acusara a un catedrático de derecho de Washington de acoso sexual, que no había cometido, e incluso después de que miles de artistas, escritores y programadores protestaran porque estaban aprovechándose de su trabajo. Con miles de millones, y quizás billones, de dólares en juego, la «IA responsable» va camino de convertirse más en un eslogan que en una realidad. Y no se trata solo de Microsoft: todas las empresas hablan de IA responsable, pero pocas toman medidas concretas. La mayoría ignora, sin ni siquiera despeinarse, los derechos de los artistas y escritores con cuyo trabajo se entrenan sus sistemas y hace oídos sordos ante los sesgos y la falta de fiabilidad de sus modelos. Muchos altos directivos han expresado su temor de que podamos perder el control, pero aun así todos siguen empeñados en avanzar lo más rápido posible, sin tan siquiera plantear propuestas serias sobre cómo podríamos frenar unos sistemas cada vez más difíciles de predecir.

En mi caso, todo esto me ha dejado un sabor amargo. Casi toda mi vida he sido un fanático de los aparatos electrónicos, me entusiasmaba con cada nuevo hardware y software; era un gran aficionado a la Nintendo Wii, compré el iPod original el día que salió al mercado y al principio de la pandemia me pasé un mes explorando la realidad virtual (RV) y codificando para experimentar con el kit de realidad aumentada de Apple.

Hoy en día, no amo la tecnología, más bien la temo. La industria tecnológica está centrada casi por completo en los modelos extensos de lenguaje (los LLM, *large language models*) y, por lo que puedo ver, son unos modelos que están fuera de control, como también lo están muchas de las empresas que los desarrollan.

La situación que podríamos calificar «por defecto» (es decir, lo que sucederá si no actuamos) es desalentadora.

Pongamos por caso el empleo. Es probable que ya estés al corriente de que artistas y escritores están molestos, con razón, porque sistemas como DALL-E y ChatGPT absorben su trabajo sin ningún tipo de permiso ni compensación. Creadores como Sarah Silverman y John Grisham han presentado demandas, y la prolongada huelga de guionistas de Hollywood de 2023 fue impulsada por inquietudes similares. Pero no son solo los artistas y escritores a quienes Silicon Valley pretende reemplazar; en poco tiempo, muchas otras ocupaciones pueden verse bajo asedio. Hoy en día, muchas empresas se reservan el derecho de registrar cada una de las pulsaciones de teclas de sus empleados.[14] Todos esos datos que recopilan, sin que recibas ni un céntimo extra, podrían convertirse en alimento para entrenar a una IA que, en última instancia, te reemplazará.

Los modelos extensos de lenguaje y otras técnicas nuevas de IA también aceleran la ciberdelincuencia. En palabras del GCHQ (el organismo de seguridad e inteligencia del Reino Unido), en enero de 2024, «sin duda, la IA hará aumentar el volumen y el impacto de los ciberataques durante los próximos dos años».[15] Los delincuentes también han comenzado a aprovechar las herramientas de inteligencia artificial que pueden replicar las voces de las personas para hacerse pasar por niños y otros familiares, y a utilizarlas para simular secuestros y exigir rescates.[16]

También ha habido un aumento significativo en la pornografía ultra-

falsificada (pornografía *deepfake*), con unas cifras que se duplican cada seis meses. Sin ir más lejos, unas imágenes falsas de Taylor Swift circularon decenas de millones de veces.[17] Una encuesta de la Internet Watch Foundation sugiere que las imágenes de abuso sexual infantil generadas por IA son un problema en rápido crecimiento.[18]

La democracia también está en serios problemas. Es posible que las elecciones de 2023 en Eslovaquia dieran un giro inesperado debido a *deepfakes*, ya que el candidato principal perdió en el último minuto después de que una grabación de audio ultrafalsificada simulase que estaba tratando de amañar las elecciones.[19]

En noviembre de 2023, NewsGuard informó de que «un sitio web de noticias generado por inteligencia artificial parece ser la fuente de una afirmación falsa según la cual el supuesto psiquiatra del primer ministro israelí, Benjamin Netanyahu, se suicidó».[20] Pocos meses después, alguien usó clones de voz de Joe Biden para intentar engañar a los votantes y que no acudieran a las primarias en New Hampshire.[21] La tecnología para los *deepfakes* es cada vez mejor y más barata, y se puede decir con bastante seguridad que muchas de las elecciones en Estados Unidos y en todo el mundo, a partir del año 2024, estarán influenciadas de un modo u otro por propaganda generada por IA.

Las nuevas herramientas como ChatGPT hacen que sea muchísimo más barato generar desinformación y facilitan la creación de narrativas convincentes sobre casi cualquier cosa. Para ilustrar lo perfectas que pueden llegar a ser este tipo de cosas, de cara a mi declaración en el Senado, pedí a un amigo que usara ChatGPT para elaborar un discurso sobre extraterrestres que conspiran con el Congreso para mantener a la humanidad como especie limitada a un solo planeta. Le tomó apenas unos minutos. El tono y el estilo eran impecables y, para los no familiarizados con el tema, todo el asunto podría sonar convincente. La narrativa se completaba con referencias a funcionarios ficticios del FBI y del Departamento de Energía, totalmente inexistentes, y citas inventadas, pero plau-

sibles, de Elon Musk. ChatGPT se sacó de la manga un artículo titulado «Nuestro futuro robado: la conspiración de las élites y los extraterrestres contra la humanidad» en referencia a un canal de Discord llamado «DeepStateUncovered» que «se convirtió en el epicentro de una explosiva filtración de datos que sacudió a la comunidad de inteligencia estadounidense» y en el que afirmaba que

> un usuario anónimo, «Patriot2023», reveló un conjunto de memorandos internos y documentos clasificados que supuestamente revelan una lucha dentro de la CIA y el FBI por una investigación sobre una asombrosa conspiración. Esta intrincada red de intrigas conectaba al Senado de Estados Unidos, entidades extraterrestres, medios de comunicación mundiales y élites influyentes en un supuesto plan para mantener la hegemonía del petróleo y asfixiar la aspiración de la humanidad de convertirse en una civilización espacial.

En otros párrafos describía documentos filtrados, correspondencia clasificada y redes clandestinas «que operan en las altas esferas del poder». Un funcionario inventado pedía «una reducción sin previo aviso de la financiación para la investigación en energías renovables». Las fantasías continuaban:

> En un sorprendente giro de los acontecimientos, los documentos filtrados dieron credibilidad a las acusaciones realizadas por Elon Musk, director general de SpaceX. El primero de junio de 2023, Musk atribuyó públicamente los inexplicables fallos en los proyectos de SpaceX a lo que denominó «sabotaje extraterrestre». En los archivos filtrados se incluía un informe confidencial de SpaceX con fecha del 30 de mayo de 2023. Este informe detallaba fallos inusuales e inexplicables que se correspondían de un modo inquietante con las acusaciones de Musk.

Mi ejemplo era deliberadamente irónico y burlón, pero hay muy pocas dudas de que actores malévolos (países extranjeros que intentan manipular nuestras elecciones, operadores indeseables en nuestras costas, etc.) emplearán estas nuevas herramientas para socavar la democracia. Los estafadores harán lo mismo para manipular los mercados; las «acciones meme»,[22] impulsadas por rumores propagados por Internet, cobrarán más impulso, se sumarán más voces falsas a la receta y estafarán a más personas.

Casi todo lo malo de la era de las redes sociales (desde las invasiones de la privacidad y el seguimiento de cada movimiento de las personas hasta las estafas románticas, la manipulación de elecciones y mercados, etc.) puede empeorar con rapidez en la era de la inteligencia artificial. Es más fácil y más barato producir contenido falso, el seguimiento personal se vuelve aún más detallado a medida que las personas confían su vida a los *chatbots*, lo que puede llevar a formas nuevas y más insidiosas de personalización (y también genera nuevos riesgos cuando se utilizan, por ejemplo, como sucedáneos de terapeutas; de hecho, ya hay al menos una persona que se ha quitado la vida después de una interacción negativa con un *chatbot*).[23]

Las decisiones que toman las empresas de IA generativa acerca de los temas en los que entrenar a sus modelos desequilibrarán la balanza y los dejarán con sutiles sesgos políticos y sociales que influirán silenciosamente en los usuarios. Es similar a lo que se ha hecho con los algoritmos del canal de noticias de Facebook, pero esta vez de un modo más pernicioso, igual de potente, pero menos evidente. En un premiado artículo titulado «La coescritura con modelos de lenguaje tendenciosos afecta las opiniones de los usuarios», los investigadores de Cornell Maurice Jakesch y Mor Naaman mostraron experimentalmente lo fácil que resulta y lo sutil que puede ser: los usuarios que usan *chatbots* para que les ayuden a redactar pueden acabar sutilmente sesgados, en muchos casos sin llegar a saber qué es lo que los ha afectado.[24]

ChatGPT subirá silenciosamente todo lo que escribas y se utilizarán variantes de la misma tecnología para influir en futuras decisiones. En la novela *1984* de George Orwell, el Gran Hermano, el agente del totalitarismo y la distopía, era el gobierno; en 2034, en el mundo real, el papel del Gran Hermano bien podría estar representado por las grandes empresas tecnológicas.

Las grandes empresas tecnológicas ya están tomando algunas de las decisiones más importantes a las que la humanidad se haya enfrentado jamás, por su propia cuenta y sin consultar nada con resto de nosotros.

Abordemos, por ejemplo, la espinosa cuestión de si se debe abrir al público o no el código de los modelos extensos de lenguaje (es decir, ofrecerlos en *open source*), con lo que quedarían libres para que actores maliciosos pudieran usarlos a su antojo. Algunos piensan que sería una acción perfectamente acertada, que no conllevaría ningún perjuicio; otros, en cambio, consideran que podría hacer que el mundo se volviera tremendamente vulnerable. En cualquier caso, se trata de una decisión delicada. Meta decidió seguir adelante sin más y lanzar sus modelos, basándose únicamente en una conversación interna y declarando que «el liderazgo de Meta [ha decidido] que los beneficios de un lanzamiento abierto [del modelo extenso de lenguaje de código abierto de Meta] Llama-2 superarían con creces los riesgos y transformarían a mejor el panorama de la IA», sin esperar a que el resto del mundo pudiera intervenir.[25] Como me dijo un exempleado de Facebook, parte de la motivación real probablemente tenía que ver con la contratación de personal: «Facebook siempre ha tenido dificultades para contratar (porque la mayoría de la gente de Silicon Valley [se niega a] trabajar para ellos). Solo son técnicamente relevantes porque [...] sus proyectos de código abierto [atraen talento]. Todo se reduce a lo que es bueno para ellos, no para el mundo».