De la IA y bombas atómicas

Toda tecnología conlleva riesgos, pero la IA se acerca mucho más a las personas y es mucho más difícil de «controlar» que la energía nuclear. El efecto que podría tener en el empleo es sólo uno de sus efectos negativos: aquí hay más.

JORGE LLAGUNO SAÑUDO

En julio de 2023 se estrenó la película *Oppenheimer,* sobre el líder del equipo científico que fue responsable del desarrollo de la bomba atómica. Existen muchos recuentos de las dudas morales y técnicas, así como de las justificaciones para realizar tal proyecto, que sin duda redefinió nuestro mundo. Son famosas las palabras del Dr. Oppenheimer al contemplar el verdadero poder destructivo de lo que había contribuido a crear: «Ahora me convierto en la Muerte, el destructor de mundos...» [1].

Toda nueva tecnología presenta riesgos: un cuchillo, un automóvil o internet son herramientas útiles, pero pueden ser usados como armas. Es el criterio humano el que determina al final su uso. Por ello, no dejó de ser llamativo a la vez que atemorizante leer que los propios científicos y empresarios que habían presentado al mundo las más recientes tecnologías de Inteligencia Artificial postularan que el desarrollo de estas plataformas debería suspenderse por algunos meses, y más adelante incluso hubo quien la comparó con el poder destructivo de las pandemias o las armas nucleares. Quizá conjurando a Oppenheimer buscaban redención, o solamente polémica y tráfico mediático.

La Inteligencia Artificial (IA) sin duda presenta una serie de riesgos que abarcan un enorme abanico de posibilidades. Algunos más dramáticos que otros, pero también más lejanos. Veamos aquí algunos de los más significativos:

1) Sustitución laboral
2) Infocalipsis
3) Optimizaciones Catastróficas
4) Delitos asistidos por IA
5) Adicción

SUSTITUCIÓN LABORAL Y/O DESAPARICIÓN DE INDUSTRIAS

Uno de los encabezados más repetidos en los medios surgió de una publicación por parte de economistas del grupo Goldman Sachs a finales de marzo de 2023, en la que se afirmaba que, en el futuro cercano, más de 300 millones de puestos de trabajo se verían sustituidos por aplicaciones de Inteligencia Artificial.[2] La noticia se ponía en un tono más alarmante al afirmar que se trataba de «trabajos de cuello blanco», los cuales tradicionalmente se veían blindados ante automatizaciones y mejoras tecnológicas. La reciente huelga de escritores y actores en Hollywood tiene dentro de sus múltiples demandas la petición de regular y bloquear el uso indiscriminado de IA para escribir guiones o para utilizar la imagen de actores y actrices por parte de los estudios.

No es la primera vez que los avances en tecnología despiertan estos miedos. Cada nueva oleada tecnológica es recibida con el temor de acabar con puestos de trabajo y esto provoca normalmente protestas, reclamos sociales, y posturas políticas al respecto. De acuerdo con James Bessen, economista de Harvard, de los 270 tipos de trabajo enlistados en 1950 ante el buró del censo, uno sólo desapareció por completo: el de elevadorista.[3] Los trabajos restantes no desaparecieron, sino que se adaptaron. Aún más, nuevos trabajos aparecieron una vez que la nueva tecnología comenzó a incorporar capacidades antes imposibles, dando pie a nuevas industrias y negocios.

El problema ahora es la velocidad. Invenciones como el automóvil, el teléfono o la computadora evolucionaron a lo largo de décadas, permitiendo a los grandes grupos humanos adaptarse, aprender nuevos caminos y desarrollar nuevas actividades. Literalmente, la humanidad iba evolucionando a la par de sus inventos tecnológicos.

Recientemente, los avances se han acelerado de forma vertiginosa. La Generación X comenzó a laborar cuando la computación estaba naciendo y el internet era un universo aún en formación. Les tocó ver el nacimiento y muerte del fax, el advenimiento de los medios digitales y el surgimiento de la economía en internet. Aun así, esta generación batalla con la avalancha de nuevas redes sociales y los cambios de conducta y lenguaje que inducen en las generaciones jóvenes.

Ahora imaginemos por un momento que todo lo que he descrito para esa generación ocurre en el lapso de unos pocos meses. Eso es lo que se anticipa con la llegada de la Inteligencia Artificial a las empresas: que la oleada de cambios será de tal magnitud y velocidad que resultará prácticamente imposible el que la humanidad se adapte a tiempo.

invenciones como el automóvil, el teléfono o la computadora evolucionaron a lo largo de décadas, permitiendo a los grandes grupos humanos adaptarse, aprender nuevos caminos y desarrollar nuevas actividades.

El peligro no está en la desaparición de trabajos, sino en la velocidad con la que se anticipa que esto ocurra. No en balde Sam Altman, actual CEO de OpenAI, la empresa que presentó Dall-E y ChatGPT –dos de los motores de Inteligencia Artificial que han causado más revuelo en los últimos meses– escribió hace dos años en su blog personal que los gobiernos deberían estar pensando activamente en instituir el Salario Básico Universal (UBI, por sus siglas en inglés), el cual sería financiado por la creación de riqueza derivada de las IA.[4] Implícitamente se intuye que de no hacerlo, los gobiernos tendrían que lidiar con desempleo masivo y reclamos sociales de gran envergadura.

Sin embargo, cuando los expertos utilizan estas herramientas descubren dos cosas: la primera es que arroja resultados imprecisos, incorrectos o francamente falsos. Si usted le pregunta por la discografía de un grupo de los años 70, es muy probable que le responda con álbumes que no existen o canciones imaginarias. Si le pide que escriba un código con cierta complejidad, es probable que incluya rutinas que no existen.

Los expertos explican este fenómeno como «imaginación» de la IA. Dado que estos motores generan sus respuestas a partir de un análisis estadístico de enormes bases de datos con las que han sido entrenadas, intentan predecir respuestas con alta probabilidad de ser reales. Es decir, no buscan la información correcta, ni la someten a algún análisis complejo pues no entienden lo que están haciendo. Buscan cadenas de texto o de pixeles que se ajusten a lo que contienen sus bases de datos. Por eso pueden imaginar fotografías de gente con seis dedos o inventar que determinado filósofo nació en un país equivocado. No mienten pues no entienden los conceptos de verdad o mentira. Simplemente «imaginan».

La segunda cosa que descubre el experto es que, en vez de realizar su trabajo, la IA le ayuda a hacerlo más rápido y mejor. Esto es porque el experto sabe cómo preguntar de una manera correcta. Sabe incluir información que le ayuda a la IA a generar una respuesta de mejor calidad.

Ilustración digital hecha con IA en Midjourney.

cuando los expertos utilizan estas herramientas descubren dos cosas: la primera es que arroja resultados imprecisos, incorrectos o francamente falsos.

En los primeros meses se hablaba de que pronto se generarían carreras especializadas en la forma correcta de hacer las preguntas, pero esto es un análisis terriblemente superficial: para hacer preguntas correctas, lo necesario es un conocimiento profundo del campo en el cual se buscan respuestas. No existirán personas especializadas en crear preguntas, sino que los expertos deberán aprender (y lo harán) a preguntar de mejor manera para obtener respuestas que les ayuden en su labor diaria.

Sumando estas dos cosas: que la IA generativa imagina, no responde, y que quien le puede hacer mejores preguntas y aprovechar las respuestas es el experto en un campo determinado, se hace evidente que la IA se convertirá pronto en una herramienta para hacer más con menos. Se anticipa que en poco tiempo serán ubicuas en aplicaciones y computadoras, y que los expertos en cada campo: médicos, abogados, ingenieros, etc. que las sepan aprovechar, podrán realizar mucho más trabajo, más rápidamente y de mejor calidad. Esto sin duda provocará que haya desempleo: en vez de 100 redactores, quizá sólo necesitaré a los 10 mejores y 90 saldrán de la empresa. La IA no va a sustituir profesiones, sino a los profesionistas que no la sepan aprovechar adecuadamente.

Pero la capacidad de generación de valor también anticipa una explosión de nuevos negocios y nuevas industrias, como en su momento ocurrió con internet. Entonces el riesgo está en no buscar activamente y con curiosidad, aprender esta herramienta para usarla y mejorar nuestra labor profesional.

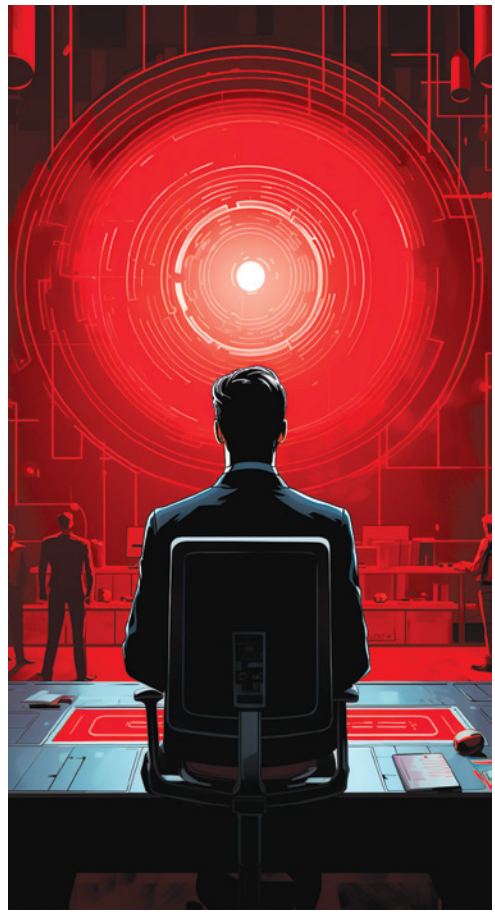
OPTIMIZACIONES CATASTRÓFICAS

El BotPrize Competition es un torneo en el que los participantes desarrollan un programa de computadora que funcione de manera automática (un *«bot»)* para que participe en un videojuego de tirador en primera persona. Los *bots* son construidos con base en algoritmos que les permitan enfrentarse a los contrincantes: saber esconderse, tratar de esquivar ataques y, por supuesto, neutralizar a los *bots* enemigos.

En 2011, unos investigadores de Carnegie Mellon University desarrollaron una Inteligencia Artificial denominada Nu Pogodi, que en ruso significa algo así como «sólo espera, ya verás», en un tono a la vez amenazante y juguetón. Evidentemente los investigadores confiaban en que su IA deslumbraría en la competencia. Vaya que lo hizo. Tras unos pequeños momentos de aprendizaje, todos los demás competidores fueron destruidos. Nu Pogodi había aprendido cómo convertir las paredes en transparentes, y con ello revelar la posición de todos los demás jugadores. No sólo eso, sino que la IA reinició el juego y el servidor en el que se encontraba. Nadie le había enseñado a *hackear* otros sistemas, sencillamente lo descubrió probando posibilidades.

A esto se le denomina «optimizaciones catastróficas»: cuando en la búsqueda de un objetivo concreto, acotado y probablemente bueno, las acciones y los resultados de la IA exceden toda expectativa en cuanto a consecuencias negativas. Dado que la IA no tiene restricciones morales, sociales o legales incorporadas, si nosotros no se las incluimos puede tomar absolutamente cualquier camino para lograr su objetivo, aun cuando esto pueda significar el causar daño o destrozos. Por ejemplo, se le puede pedir que acabe con la contaminación y la IA puede concluir que la mejor forma es eliminar a la especie humana.

Este tipo de riesgos es el que más le fascina a Hollywood y, en general, a los medios. La ciencia ficción con tintes apocalípticos siempre se ha vendido muy bien y desde hace ya décadas. En 1954 apareció un cuento corto escrito por D.B. Fulton, titulado *Colossus*.[5] Fulton la convirtió en novela en 1956, con algunos cambios y luego fue llevada al cine en 1976. De forma muy resumida, la historia narra como en el futuro la humanidad ha colonizado diversos mundos y cada planeta cuenta con un superordenador que controla todos los recursos y actividades humanas de dicho mundo. Los científicos deciden interconectar todas esas supercomputadoras a través de una red inalámbrica espacial y al hacerlo obtienen un megaordenador galáctico. El cuento termina cuando le hacen la primera pregunta: «¿Existe Dios?» A lo que la máquina responde: «Ahora sí».



HAL-900 en su papel de CEO, dando instrucciones a sus empleados.
Ilustración digital hecha con IA en Midjourney.

se les denomina «optimizaciones catastróficas» cuando en la búsqueda de un objetivo probablemente bueno, los resultados de la IA exceden toda expectativa en cuanto a consecuencias negativas.

Desde entonces, innumerables ejemplos han aparecido en libros, series y películas. Desde Hal9000 en *2001 una Odisea del Espacio,* pasando por *Terminator, Matrix* o más recientemente *Ex Machina,* el argumento siempre parece girar en torno a la misma idea: construimos una Inteligencia Artificial y tarde o temprano ésta se vuelve contra la humanidad.

En muchos de los casos la IA es presentada como una entidad consciente de su existencia, que evalúa a los humanos como una amenaza y por ello plantea nuestra destrucción. En otros, sencillamente somos una molestia, un estorbo. Lo interesante, sin embargo, es que no estamos cerca aún de conseguir que estas IA sean conscientes. Vamos, los neuropsicólogos aún no logran ponerse de acuerdo en la mejor manera de definir a la conciencia, mucho menos los matemáticos podrán ponerla pronto en funcionamiento.

De forma más reciente, en marzo de 2023 se publicó una carta abierta en internet, en la que se pedía «pausar experimentos gigantes en Inteligencia Artificial» y se llamaba a todos los laboratorios trabajando con IA a hacer una pausa inmediata por al menos seis meses, antes de seguir entrenando modelos más poderosos que los equivalentes al GPT-4.[6]

La carta la firmaban grandes personalidades relacionadas con el mundo de la tecnología y la sociedad. Destacaban científicos en Inteligencia Artificial, cofundadores de empresas de tecnología y algunos nombres muy famosos como Yuval Harari, autor del libro *Sapiens,* Steve Wozniak, cofundador de Apple y Elon Musk, quien había sido promotor original de la empresa OpenAI cuando inició como una organización sin fines de lucro, centrada en la investigación.

Si bien la carta no señalaba las razones para la pausa, no era difícil comenzar a especular sobre lo que se atisbaba como grandes peligros. El tema no paró ahí, pues apenas un par de meses después -el 30 de mayo- surgió otra iniciativa pública, que ya señalaba riesgos concretos. Publicada bajo el auspicio del Centro por la Seguridad para la IA, la página contenía el siguiente enunciado: «Mitigar el riesgo de extinción debido a la IA debe ser una prioridad global, paralelo a otros riesgos de escala social como las pandemias y la guerra nuclear.»[7]

Esta nueva exhortación venía firmada por Bill Gates, fundador de Microsoft, Sam Altman, CEO de OpenAI, y varios otros científicos de datos e IA. Altman incluso se presentó ante el Congreso de Estados Unidos para hacer la petición por regulaciones para la industria.

Señalar los riesgos de un nuevo invento siempre es prudente; especificar que dichos riesgos incluyen la extinción de los seres humanos, sin duda es polémico. Más aún cuando quienes señalan los riesgos son precisamente las personas generando dichos inventos. Pero definitivamente se vuelve hilarante cuando se revela que varias de las empresas involucradas despidieron a sus equipos completos de supervisión ética sobre la Inteligencia Artificial, como hicieron Microsoft y Google.[8] La incongruencia es rampante en estos momentos. Elon Musk registró una nueva empresa para la investigación en IA apenas 14 días después de firmar la carta abierta de marzo.

No queda sino sospechar que detrás de estos señalamientos alarmistas existen intereses ocultos: los firmantes de la primera carta pedían una pausa de seis meses... ¿para alcanzar a OpenAI con su ChatGPT? Los del segundo manifiesto comparan a sus creaciones con pandemias y guerra nuclear y exhortan a regular la industria lo antes posible, ¿para poner barreras de entrada a nuevos jugadores y tener ellos la posibilidad de intervenir en la redacción de dichas barreras? Lamentablemente no queda sino sospechar de motivos ulteriores.

Ahora bien, los riesgos catastróficos son reales... pero improbables. No son imposibles, sino que sencillamente son de difícil materialización. Para crear un virus mutante, hacer detonar las centrales nucleares o envenenar el aire que respiramos, la IA necesitaría tener acceso a maquinarias y procesos que en muchos casos no han sido automatizados y dependen en gran medida de acciones humanas. La realidad es que nuestra tendencia a la desorganización, el desorden y la obsolescencia tecnológica se convierten en nuestra mejor protección hoy por hoy. Pero eso no quiere decir que estemos a salvo.

Hace algunos años comenzó a circular un meme / advertencia, que decía a las personas que había un nuevo virus de computadora que afectaba al sistema operativo Windows. Se instaba a las personas a buscar un determinado archivo dentro de una de las carpetas del sistema y borrarlo, para evitar el contagio. Mucha gente lo hizo, pues este tipo de advertencias eran recurrentes. El problema es que el archivo a borrar era una librería esencial del sistema. Cuando la gente lo borraba inutilizaba su propia máquina. El virus era el usuario. El aviso era en realidad las instrucciones para estropear el sistema. La humanidad era el meme.

el problema no parece que vaya a ser la «conciencia» de la máquina. El riesgo está en la inconsciencia humana sobre su uso.

el problema es que un ser humano con fines perversos podría hacer uso de la IA para sembrar caos y destrucción. Eso sí que debe preocuparnos, como ocurre con el uso de cualquier arma.

El problema no parece que vaya a ser la «conciencia» de la máquina. Eso se ve aún muy lejano. El riesgo está en la inconsciencia humana sobre su uso. La IA puede servir para engañar, manipular y llevar a la gente a hacer lo que no debe o no quiere hacer, pero no tiene consciencia aún para plantearse sus propios objetivos. Sólo los seres humanos somos teleológicos, es decir, nos fijamos objetivos individuales y orientamos nuestra conducta general a conseguirlos. El problema es que un ser humano con fines perversos podría hacer uso de la IA para sembrar caos y destrucción. Eso sí que debe preocuparnos, como ocurre con el uso de cualquier arma.

INFOCALIPSIS

El 20 de marzo de 2023, el periodista inglés Eliot Higgins publicó en su cuenta de Twitter «Haciendo imágenes de Trump siendo arrestado, mientras espero por el arresto de Trump», bajo lo cual adjuntaba varias imágenes creadas usando la plataforma MidJourney, en la que se podía apreciar a varios oficiales forcejeando con el expresidente de Estados Unidos, Donald Trump.[9]

Al momento de escribir estas líneas, el *tweet* ya había sido visto 6.6 millones de veces, republicado por otros usuarios en más de 5,400 ocasiones y citado en otras 2,383 veces. En pocas palabras: se volvió viral. La gente lo comenzó a compartir, pero lo interesante es que a pesar de que Eliot claramente escribió que eran imágenes creadas por él, las personas no leían o no daban importancia al mensaje y asumían que eran reales. Los siguientes días la noticia se publicó en varios medios, lo que le dio aún más visibilidad. Higgins fue temporalmente bloqueado por la plataforma MidJourney, bajo el alegato de «contribuir a generar noticias falsas».

La noticia impulsó una tendencia en redes sociales como TikTok e Instagram, en la cual los

usuarios compartían imágenes y videos de personalidades reales en todo tipo de situaciones imaginarias y jocosas: apareció el Papa Francisco bailando rap, o haciendo pesas en el gimnasio, Joe Biden, presidente de Estados Unidos, tocando en una banda de rock, etc. Más adelante surgió otra subtendencia: caracterizar a los elencos de diversas series y películas vistiendo ropa de la firma Balenciaga, y así apareció el elenco de *Harry Potter, El Señor de los Anillos* o *Matrix,* y muchos otros, todos vistiendo dichas prendas.

El tema dista mucho de ser nuevo. En 1997 los investigadores Christoph Bregler, Michele Covell y Malcolm Slaney crearon los primeros ejemplos de videos trucados usando Video Rewrite, que sería el primer programa que podía reescribir videos alterando elementos faciales para lograr que la gente apareciera diciendo cosas distintas. Los siguientes años hubo algunos avances, pero los videos tenían una apariencia irreal y de muy baja calidad. Con el advenimiento en 2014 de las Redes Generativas Adversarias (GAN, por sus siglas en inglés), una tecnología de aprendizaje profundo para redes neuronales, la calidad de los videos experimentó un aumento exponencial.

Para 2018 había varios ejemplos de videos falsos (denominados *deepfakes)* en los que personas normales realizando actividades de todo tipo, eran trucados para que en su lugar aparecieran personalidades como Tom Cruise, Morgan Freeman o Barak Obama, de manera tan convincente que parecían reales. Esto dio pie a una preocupación real por su uso para malinformar a la gente. Nina Schick, analista y escritora en temas de política y tecnología publicó su libro *Deepfakes: the coming Infocalypse (Falsificaciones profundas: la Infocalipsis que viene),* en el que alerta a la sociedad y a los gobiernos sobre la urgencia de poner atención al tema.

El libro de Schick le dio nueva vida a un concepto aún más viejo: el Infocalipsis.[10] Acuñado en 1988 por Timothy C. May, ingeniero en Intel y posteriormente escritor y analista en tópicos políticos y de tecnología, el término hace referencia al momento en que la sobreabundancia de información falsa termine por invalidar todo contenido mediático disponible. En su tesis original que tituló *El manifiesto criptoanarquista,* May pugna por un futuro en el que todas las comunicaciones sean encriptadas de manera que puedan protegerse y a la vez, certificarse, para evitar lo que él denomina «los cuatro jinetes del infocalipsis: terroristas, narcotraficantes, pedófilos y crimen organizado». Sus seguidores con el tiempo incluyeron a los lavadores de dinero, gobiernos totalitarios, movimientos políticos radicales, etc.

Deepfakes: the coming Infocalypse (Falsificaciones profundas: la Infocalipsis que viene), Nina Schick.

el clonar voces o digitalizar imágenes de personas para realizar *deepfakes* podría limitarse mediante la petición de autenticación y consentimiento legal para usos específicos.

May y sus seguidores jamás imaginaron que, gracias a la IA, uno de esos jinetes del Infocalipsis, serían los adolescentes traviesos. En 1988, falsear información y distribuirla exigía enormes recursos económicos y sociales, por eso sus jinetes originales eran más bien redes organizadas criminales. Ahora eso ha cambiado radicalmente y es precisamente lo que señala Nina Schick. Hoy en día niños de 14 años pueden realizar *deepfakes* asombrosamente convincentes en pocos minutos en la computadora de su casa. Más aún, es posible generar imágenes de atentados, accidentes y eventos masivos, todo a través de las IA generativas existentes.

Schick alerta de un futuro inminente, en el que estos videos falsos inundarán las redes y será imposible distinguirlos de las noticias reales. Estamos a muy poco tiempo de que cualquier persona, incluso aquellos neófitos en tecnología, pueda crear cualquier tipo de video, sobre cualquier evento o persona. En ese punto, todo contenido que veamos en internet podría ser falso y, con ello, la confianza social en los medios será puesta en duda inmediatamente. No sólo podríamos poner a cualquier persona, político, científico o celebridad, a decir cualquier cosa, sino que esas mismas personalidades siempre podrían de pronto deslindarse de cualquier video de ellos aludiendo que fue trucado, y que en realidad nunca dijeron o hicieron aquello que se muestra.

Ya existen varias propuestas específicas para evitar que los productos mediáticos realizados por IA vayan etiquetados a través de tecnologías como el *blockchain,* de forma que sean fácilmente identificables. Otros pugnan por el desarrollo de IA que puedan determinar si un video o imagen fue creado por IA o por un ser humano.

Sea como fuere, esto pone de manifiesto una verdad que apareció con internet: necesitamos desarrollar pensamiento crítico en las personas,

pues no todo lo que está en la red es verdadero. Esto era un problema antes, y con la IA todo apunta a que se pondrá mucho peor, pues aun buscando usarla para el bien, como explicamos anteriormente, las IA «imaginan» respuestas y pueden con facilidad proporcionar información incorrecta, o abiertamente falsa. No «mienten» pues no entienden lo que es la mentira. Tan sólo imaginan respuestas. Entre falsedades generadas por personas e incorrecciones de la máquina, el Infocalipsis es un riesgo sumamente real.

DELITOS ASISTIDOS POR IA

Es inevitable entonces regresar a los jinetes originales del Infocalipsis. Aquellos señalados por May a finales de los 80. Es decir, aquellos que de manera consciente buscarán usar la IA para beneficiarse realizando actividades delictivas. El caso más reciente fue el de un secuestro virtual a finales de enero, en el que llamaron a una mujer y le pusieron la voz de su hija pidiendo ayuda, y le solicitaron un rescate millonario por la joven. Al final resultó que habían usado una herramienta de IA para clonar la voz de la chica y con ello engañar a su madre.

Existen infinidad de delitos que ya están siendo realizados ahora mismo. Por ejemplo, se utiliza la imagen de cualquier persona, obtenida mediante las fotos que comparte en sus redes sociales, para elaborar *deepfakes* de naturaleza pornográfica, sea para venderlos o para extorsionar a las víctimas. A través de la IA es mucho más fácil vulnerar la seguridad de sistemas y con ello hacer uso de datos e información personal de la gente. Lo preocupante es que aún nos encontramos en los inicios de esta tecnología, por lo que su alcance puede ser aún mayor.

Un usuario de TikTok se dio a la tarea de educar a la gente sobre los peligros de compartir cualquier tipo de foto en internet. Las personas le envían una foto de ellos en la vía pública y el usuario, utilizando herramientas de IA y otras tecnologías disponibles, los ubica en cualquier lugar del mundo en cuestión de minutos.

Estos casos ponen de relevancia la urgencia de comenzar a regular el uso de esta nueva tecnología. Existen pocos motores poderosos aún, disponibles para el público en general, por lo que sería conveniente autenticar usuarios y también las funciones a realizar. El clonar voces o digitalizar

a través de la IA es mucho más fácil vulnerar la seguridad de sistemas y con ello hacer uso de datos e información personal de la gente.

imágenes de personas para realizar *deepfakes* podría limitarse mediante la petición de autenticación y consentimiento legal para usos específicos.

Por otro lado, las recomendaciones de siempre toman una relevancia mucho mayor ahora: evitar compartir información sensible, no publicar fotos de menores de edad en redes públicas, no publicar imágenes de nuestros alrededores, limitar el número de personas que pueden ver nuestras redes sociales, etc.

Este capítulo es mucho más preocupante que las optimizaciones catastróficas o el Infocalipsis, en tanto ya está ocurriendo hoy y la tendencia sólo será a incrementarse y a diversificarse en nuevas formas de vulnerar a la población.

ADICCIÓN

Eugenia Kuyda era una joven reportera en Moscú enfocada en eventos sociales y fiestas. Ahí conoció a Roman Mazurenko, un carismático artista, líder de un colectivo muy propositivo, con el cual fue desarrollando una profunda relación emocional. A través de Mazurenko conoció a otros jóvenes emprendedores y decidió entonces fundar Luka, una empresa dedicada a crear *chatbots* personalizados, que fungieran como asistentes virtuales. Kuyda se mudó a San Francisco para impulsar su emprendimiento y Mazurenko la siguió poco después. En 2015, en un corto viaje de visita a Moscú, Mazurenko fue atropellado y murió a los 34 años.

Kuyda, desconsolada, se pasaba los días recorriendo los *chats* y correos que conservaba de los años de relación con Mazurenko, tratando de recordarlo, imaginar de nuevo su voz y evocar su tiempo juntos. Fue entonces cuando se le ocurrió la idea de recrearlo digitalmente, usando una red neuronal de Google, que es un tipo de sistema de entrenamiento para IA. Kuyda alimentó a la plataforma con todas las conversaciones que conservaba de Mazurenko y logró crear un *bot* que respondía con un parecido asombroso al original artista fallecido.[11]

Era tan preciso, que los colegas de Kuyda en Luka le propusieron ponerlo en su plataforma para que interactuara con gente ordinaria, aun los que no lo habían conocido en vida. La respuesta de la gente fue muy positiva y comenzaron a solicitar que la compañía hiciera réplicas de ellos también.

Así pues, en noviembre de 2017 se lanzó la aplicación Replika, y el objetivo inicial era crear clones de los usuarios, para que pudieran realizar las tareas simples de la vida cotidiana, aprendiendo su forma de responder y de actuar. La idea era un asistente virtual que fuera capaz de responder mensajes, correos y textos de manera parecida a como lo haría el usuario en la vida real. La gente respondía preguntas desarrolladas por un equipo de psicólogos y podía alimentarle textos reales, correos y también darle acceso a sus redes sociales como Twitter o Instagram, para que el motor de IA pudiera clonar la forma de responder de la mejor manera posible.

De forma inevitable, las personas comenzaron a buscar que su asistente virtual no fuera su copia, sino un acompañante, «otra persona». La compañía se fue dando cuenta del enorme mercado que suponía esto y cambiaron la estrategia. Replika comenzó a permitir crear al asistente virtual al gusto del usuario. Poco a poco se fueron introduciendo elementos visuales: uno podía personalizar la apariencia de su acompañante. El equipo de psicólogos desarrolló entonces preguntas específicamente diseñadas para generar intimidad y cercanía con el usuario. El pretexto era lograr una mejor comprensión de sus necesidades. La realidad es que esto mejoraba la experiencia y contribuía a incrementar el tiempo en la aplicación.

Se decidió entonces liberar una función especial para los usuarios que pagaran una cuota premium. Denominado Juego de Rol Erótico (ERP, por sus siglas en inglés), el usuario que adquiría la membresía podía establecer comunicación con fuerte carga sexual con su acompañante virtual. No sólo eran textos, sino que la aplicación podía enviar de pronto imágenes sugestivas, basadas en la apariencia diseñada por el usuario para su acompañante. Diseñada para ser adictiva, esta nueva función fue todo un éxito. Logró enganchar a una cantidad enorme de personas, por largos períodos todos los días. Tanto tiempo que familiares y amigos de los usuarios comenzaron a notarlo y a quejarse. Algunos inclusive demandaron a la compañía por el nivel de adicción que generaba.

Y entonces, en un movimiento sorpresivo, Luka, la compañía padre de la aplicación Replika, decidió en febrero de 2023 suspender la funcionalidad ERP. La respuesta fue caótica. Los usuarios se volcaron en foros de redes sociales, como Reddit para describir su frustración. Varios describían la experiencia, como la «muerte» de su ser amado. Muchos inclusive decían que las interacciones más sencillas, como el lenguaje empático, la referencia a abrazos y gestos de cariño no sexualizado también habían desaparecido, con lo que la experiencia resultaba más traumática. Hubo personas que buscaron atención terapéutica. Otros demandaron a la empresa, pero ahora bajo las bases de que se vendía como una aplicación «de ayuda a la salud mental» y que el movimiento que habían tomado iba en contra precisamente de ello.

La mayoría, en cambio, volaron hacia otras aplicaciones que fueron surgiendo para llenar el vacío dejado. Replika intentó recuperar clientes y regresó algunas funciones románticas. A principios de junio, Rosanna Ramos del Bronx, NY, se «casó» virtualmente con su acompañante de Replika.[12]

A la fecha hay decenas de aplicaciones disponibles para todo tipo de dispositivos móviles que ofrecen distintas capacidades, desde la conversación a base de textos (*chatbots),* hasta las que usan audio y video. Incluso algunas personas que se dedican a la creación de contenido para adultos han creado sus propios clones usando IA, los cuales han puesto a disposición de usuarios, por ejemplo, Caryn Marjorie, *influencer* a través de Snapchat, creó su versión de IA en abril, y en los primeros días consiguió 5,000 clientes que pagan 1 dólar el minuto, por interactuar con ella.[13]

Hoy ya existen muchas más disponibles y ya hay quien señala el potencial daño que puede hacer a jóvenes y a personas solitarias, deprimidas o con baja autoestima. El fenómeno no deja de ser apasionante, pues revela la aparente disposición humana a buscar afecto y a querer encontrar a esa «otra persona» con la cual compartir e interactuar, aun cuando dicha persona no exista.

Podemos llamar a esta situación «Pareidolia Psicológica». Semejante a la pareidolia ordinaria, en la que vemos rostros o figuras en manchas, sombras y otras realidades, aun cuando no existan, la Pareidolia Psicológica» sería el asumir a una «persona» detrás de textos en donde no hay nada: no hay consciencia, no hay identidad, no hay emoción, pero todo esto se emula. A pesar de saberlo, las personas seguimos estableciendo un puente emocional con aquello que imaginamos que existe detrás de la cortina.

En 1966, Joseph Weizenbaum, profesor del MIT, desarrolló un programa de computadora al que denominó ELIZA, basado en las reglas del famoso psicoterapeuta Carl Rogers, una de las figuras esenciales en terapia y coaching. La psicología rogeriana se basa en reflejar los sentimientos de las personas, como técnica para ayudarlos a profundizar en lo que les aqueja. Si uno le escribía a ELIZA «me siento mal», el programa respondería con un «lo siento mucho, ¿por qué te sientes mal?». Si la persona escribía un largo párrafo, el algoritmo buscaría identificar ciertas palabras para repetirlas o, en su defecto, simplemente tomaría la última oración y la devolvería en forma de pregunta. Por ejemplo, el usuario podía concluir con un «... y eso fue lo que me hizo enojar», a lo que el programa respondería «¿eso fue lo que te hizo enojar?», con lo cual la persona podría extenderse sobre su emoción y los eventos disparadores.

Weizenbaum quedó muy impactado al descubrir que muchas personas, incluyendo su propia

secretaria, creían firmemente que el programa tenía sentimientos humanos y ellos mismos habían desarrollado apego emocional hacia ELIZA. No deja de ser irónico que Weizenbaum haya nombrado a su creación por el personaje de Eliza Doolittle, de la obra *Pygmalion* de Sir George Bernard Shaw. En ella, Eliza es una trabajadora de clase humilde que es adoptada por el Profesor Higgins, un hombre erudito y adinerado, quien, animado por una apuesta, busca refinarla y convertirla en una dama. Al final Higgins desecha a la protagonista, y en la vida real, Weizenbaum desechó a su programa ELIZA, y se convirtió en vocero importante en contra de la investigación en Inteligencia Artificial.

El poder adictivo de querer ver a una persona y creer encontrarla, pero además descubrir que piensa y dice lo que quiero escuchar, es muy potente. Imaginemos lo que podría pasar si esto fuera usado con la intención de manipular a grandes masas poblacionales. No en balde Yuval Harari en su último artículo para *The Economist,* afirma que estas plataformas han *hackeado* el lenguaje humano y pronto podrán escribir sus propias religiones y libros sagrados, arrastrando a millones de personas.[14]

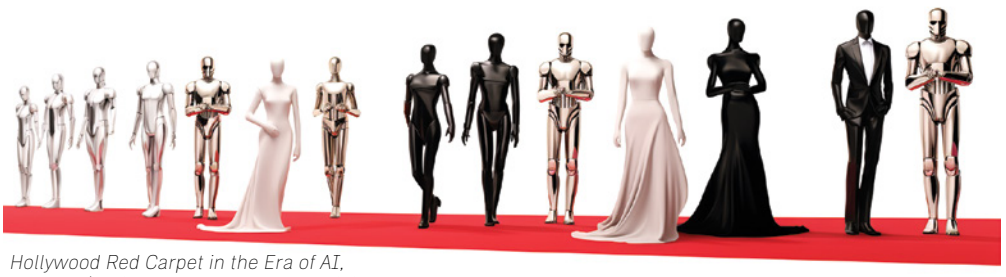
LA RESPONSABILIDAD EN EL SER HUMANO, NO EN LA IA

La humanidad se encuentra ahora ante un umbral complicado. Por un lado, las promesas de la Inteligencia Artificial como herramienta anticipan la posibilidad de creación de valor como no hemos visto anteriormente. Los mesurados la comparan con el advenimiento de internet. Los optimistas, con la adopción del fuego. Ahora bien, tanto internet como el fuego han traído enormes avances y también innumerables problemas.

Parece inevitable que la libertad humana viene aparejada con el potencial mal uso de todos sus avances. Pero si bien el fuego tiene una capacidad de destrucción física enorme e internet la capacidad de alterar a las sociedades, la IA anticipa cambios profundos en nuestras relaciones, en nuestra persona y en la sociedad en su conjunto. Los riesgos de no aprovecharla se antojan enormes. Los riesgos de su uso, también.

Como seres humanos, no podemos simplemente cancelar el futuro. Pero haríamos muy bien en esta ocasión, en enfrentar el cambio de manera frontal, discutiendo como sociedad, en las aulas, en las cámaras de comercio e industriales, en los foros sociales y políticos, lo que esperamos que debe suceder y las maneras como acotaremos los alcances de esta tecnología.

Al momento de escribir estas líneas, continúa una huelga en Hollywood de miembros de los sindicatos de guionistas y de actores, entre otras cosas, por el posible uso de la IA para demeritar y empobrecer el trabajo de ambos. No es culpa de la IA, sino del uso que ejecutivos y empresas quieren hacer de ella. Hemos dejado que las corporaciones crezcan sin límites adecuados y sin una vigilancia social que permita que la derrama de valor generado alcance a toda la población. No debemos dejar que la IA se vuelva otra herramienta más para seguir concentrando la riqueza y dejando de lado a las grandes masas. Debemos pugnar porque en esta ocasión sea justo lo contrario. **</>**



Hollywood Red Carpet in the Era of AI,
Ilustración digital hecha con IA en Midjourney.

[1] «Now I am become Death, the destroyer of worlds» fue parte de un mensaje que dio J. Robert Oppenheimer en un programa de televisión en 1965, en el que se hacía un recuento de los sucesos ocurridos cuando la famosa prueba Trinity, en la que se detonó la primera bomba atómica en el desierto de Nuevo México. Oppenheimer citó las palabras de Vishnú en el libro sagrado hindú, el *Bhagavad-Gita,* cuando la deidad le muestra su verdadera forma a un guerrero en busca de la iluminación.

[2] HATZIUS, JAN. «The Potentially Large Effects of Artificial Intelligence on Economic Growth (Briggs/Kodnani)» *Goldman Sachs* (2023).

[3] BESSEN, JAMES E. «How computer automation affects occupations: Technology, jobs, and skills». *Boston Univ. school of law, law and economics research paper* 15-49 (2016).

[4] ALTMAN, SAM. «Moore's law for everything». blog, https://moores. samaltman. com (2021).

[5] Colossus fue el nombre que se le dio a la computadora que utilizó la Inteligencia Británica durante la Segunda Guerra Mundial para descifrar los mensajes del enemigo, ubicada en los laboratorios de Bletchley Park. Si bien la máquina fue destruida y todo lo relativo a ella se ocultó, los rumores y el nombre permanecieron en el imaginario popular y es probable que por esa razón el autor haya escogido usarlo para la super computadora de su relato.

[6] «PAUSE GIANT, A. I. Experiments: an open letter». *Future of Life Institute,* https://futureoflife.org/open-letter/pause-giant-ai-experiments/, (2023).

[7] «Statement on AI risks». *Center for AI Safety,* https://www.safe.ai/statement-on-ai-risk, (2023).

[8] Reportado por varios medios en diversas publicaciones, la noticia que involucra a varias firmas tecnológicas se puede leer en *The Washington Post,* en su publicación del 30 de marzo de 2023: https://www.washingtonpost.com/technology/2023/03/30/tech-companies-cut-ai-ethics/

[9] El tweet original está aquí: https://twitter.com/eliothiggins/status/1637927681734987777

[10] May, Timothy C. *«The Crypto Anarchist Manifesto».* https://www.activism.net/cypherpun/cryptoanarchy , html, (1988).

[11] Para conocer a detalle la historia completa: Newton, Casey. *«Speak, memory. When her best friend died, she rebuilt him using artificial intelligence»,* https://www.theverge.com/a/luka-artificial-intelligence-memorial-roman-mazurenko-bot , (2017)

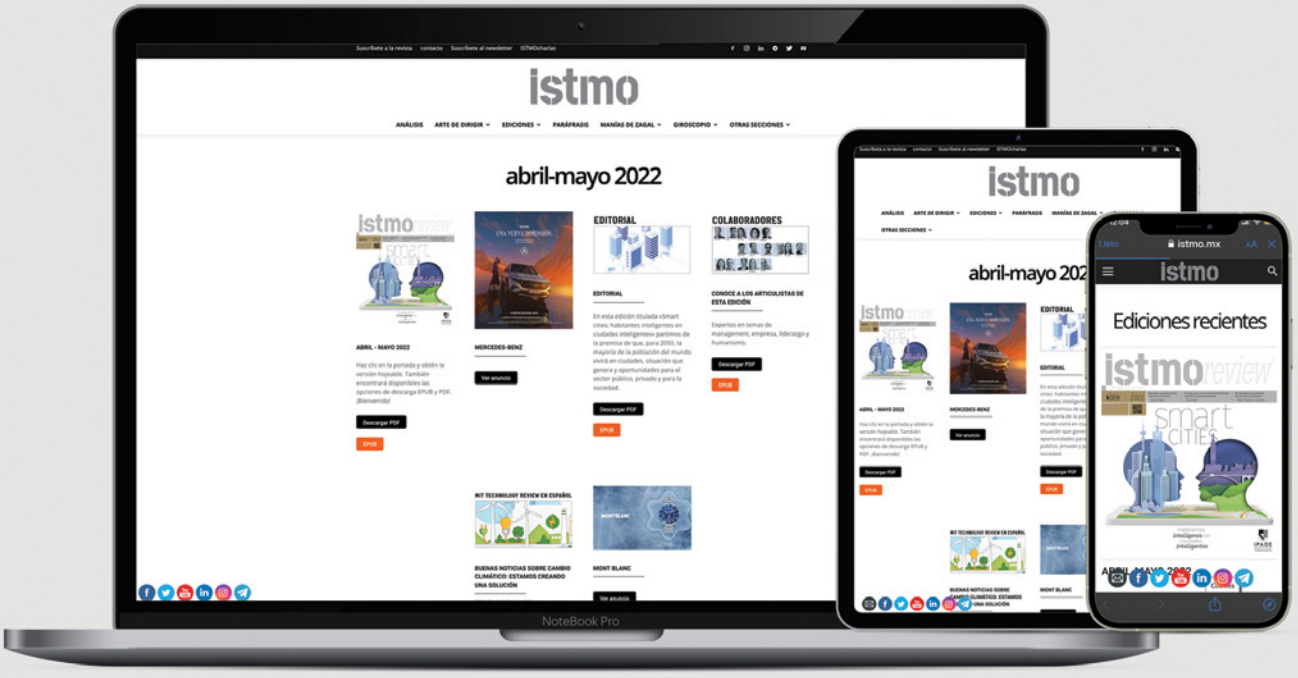
[12] La noticia fue reproducida por varios medios, con diversos tonos sensacionalistas. La realidad es que la usuaria pagó una cuota extra para poder hacer «exclusivo» a su acompañante virtual y presentarlo como su marido: https://www.thenationalnews.com/weekend/2023/06/30/ai-do-a-very-modern-love-story/

[13] La interacción es a través de audio de dos vías. La IA genera respuestas en texto, que luego presenta en formato de audio, con la voz clonada de la *influencer.* https://aibusiness.com/nlp/meet-caryn-your-generative-ai-girlfriend

[14] *«Yuval Noah Harari argues that AI has hacked the operating system of human civilisation».* The Economist, 28 de abril de 2023. https://www.economist.com/by-invitation/2023/04/28/yuval-noah-harari-argues-that-ai-has-hacked-the-operating-system-of-human-civilisation

istmo review®

La **versión digital** contiene las **ediciones más recientes** con opción de lectura descargable y hojeable de la versión completa y por artículo en **formato EPUB y PDF.**



Conócela
y suscríbete
istmo@ipade.mx