Deloitte Review
No. 22, enero 2018

Más inteligentes juntos.

Por qué la inteligencia artificial necesita diseño centrado-en-lo-humano♦

Por James Guszcza
Ilustración por Barry Downard

Deloitte. Insights

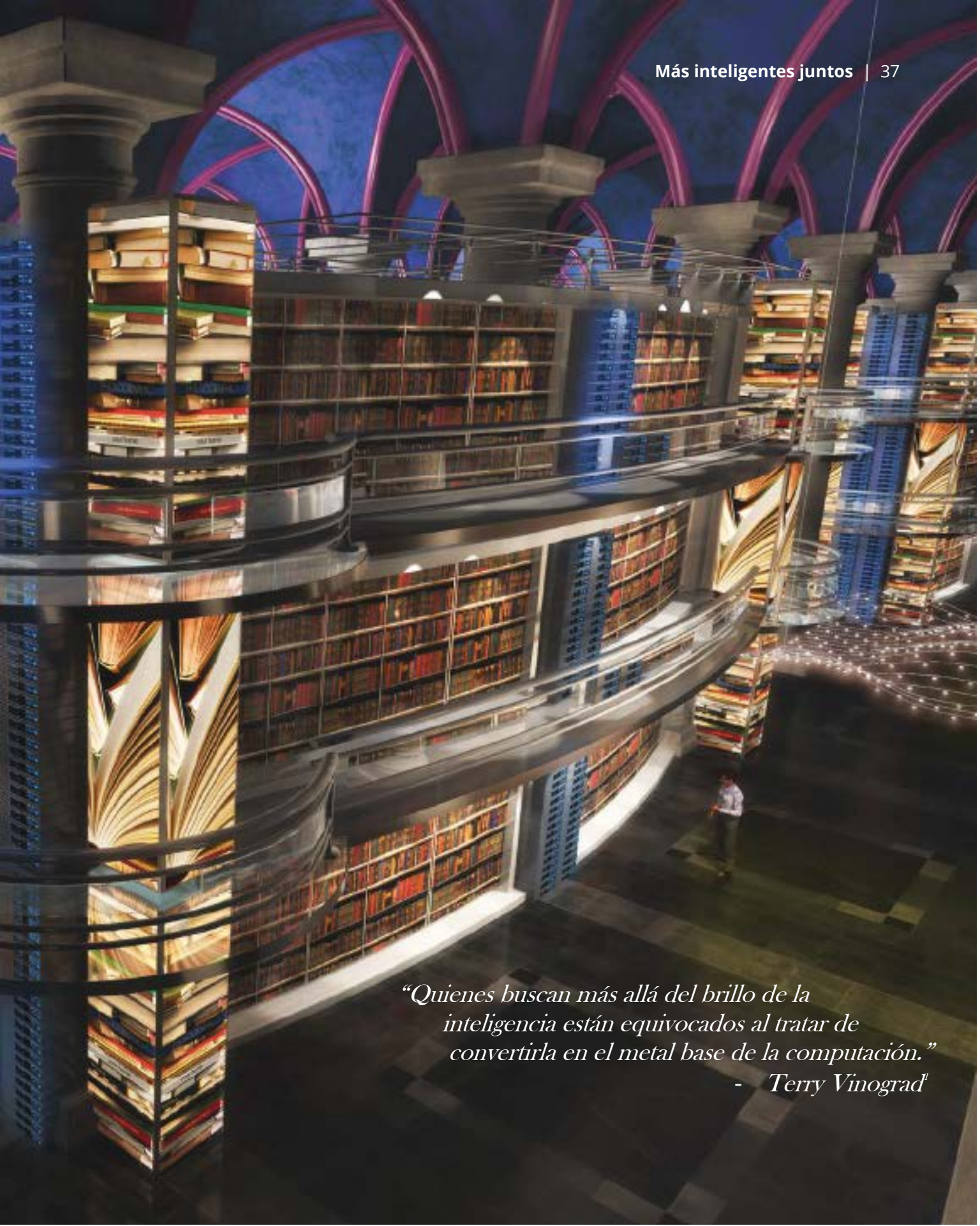
♦ Documento original: "*Smarter together. Why artificial intelligence needs human-centered design*", Deloitte Review, Issue 22, January 2018. By James Guszcza. Ilustration by Barry Downard.
https://www2.deloitte.com/insights/us/en/deloitte-review/issue-22/artificial-intelligence-human-centric-design.html.
Traducción realizada por Samuel A. Mantilla, asesor de investigación contable de Deloitte &Touche Ltda., Colombia, con la revisión técnica de César Cheng, Socio Director General de Deloitte & Touche Ltda., Colombia.

MÁS INTELIGENTES JUNTOS.

POR QUÉ LA INTELIGENCIA ARTIFICIAL NECESITA DISEÑO CENTRADO-EN-LO-HUMANO

Por James Guszcza

ILUSTRACIÓN POR BARRY DOWNARD



"Quienes buscan más allá del brillo de la inteligencia están equivocados al tratar de convertirla en el metal base de la computación."

- *Terry Vinograd*[ii]

LA INTELIGENCIA ARTIFICIAL (IA) ha surgido como un problema que caracteriza nuestro tiempo, configurado para remodelar los negocios y la sociedad. La emoción está garantizada, pero también hay preocupaciones. A nivel del negocio, los proyectos grandes de "grandes datos" e IA a menudo fallan en entregar. Muchos de los culpables son familiares y persistentes: ajustar las clavijas tecnológicas en agujeros redondos estratégicos, sobreestimar la suficiencia de los datos disponibles o subestimar la dificultad de organizarlos en formas utilizables, dando pasos insuficientes para asegurar que los resultados algorítmicos resultan en el resultado de negocio deseado. A nivel de la sociedad, los titulares están dominados por el problema del desempleo tecnológico. Aun así, crecientemente está quedando claro que la tecnología digital puede codificar sesgos sociales, distribuir conspiraciones y promulgar noticias falsas, amplificando las cámaras del eco de la opinión pública, secuestrando nuestra atención, e incluso menoscabando nuestro bienestar mental.[2]

Abordar de manera efectiva tales problemas requiere una concepción realista de la IA, la cual a menudo es promocionada como el surgimiento de "mentes artificiales" en un camino exponencial hacia generalmente dejar fuera-del-pensamiento a los humanos.[3] En realidad, las aplicaciones de IA de hoy han estado en desarrollo durante décadas, pero están siendo implementadas en computadores considerablemente más poderosos y preparadas a partir de conjuntos más grandes de datos. En sentidos estrechos son "inteligentes," no de la manera general como los humanos son inteligentes. En términos funcionales, es mejor verlas no como "máquinas de pensamiento," sino como prótesis cognitivas que pueden ayudarles a los humanos a pensar mejor.[4]

En otras palabras, los algoritmos de IA son "herramientas de la mente," no mentes artificiales. Esto implica que las aplicaciones exitosas de la IA dependen más que de solo grandes datos y algoritmos poderosos. El *diseño centrado-en-lo-humano* también es crucial. Las aplicaciones de IA tienen que reflejar concepciones realistas de las necesidades del usuario y de la sicología humana. Parafraseando a Don Norman, pionero del diseño centrado-en-lo-humano, la IA necesita "aceptar el comportamiento humano tal y como es, no la manera como quisiéramos que fuera."[5]

Este ensayo explora la idea de que las *tecnologías* inteligentes es improbable que engendren *resultados* inteligentes a menos que estén diseñados para promover la *adopción* inteligente de la parte de los humanos como usuarios finales. Muchos de nosotros hemos experimentado el efecto aparentemente paradójico de agregar a un equipo un individuo altamente inteligente, solo para presenciar que la efectividad del equipo – su "QI colectivo" – disminuye. Análogamente, la tecnología de IA "inteligente" puede de manera inadvertida resultar en "estupidez artificial" si es pobremente diseñada, implementada, o adaptada al contexto social humano. Los factores humanos, organizacionales, y sociales son cruciales.

Una estructura de IA

Es común identificar la IA con máquinas que piensan como humanos o simulan aspectos del cerebro humano (para una discusión de esos puntos de partida potencialmente engañosos, vea el recuadro, "Los significados pasados y presentes de la 'IA'," en la pg. 43). Quizás aún más común es la identificación de la IA con varias *técnicas* de aprendizaje de máquina. Es verdadero que el aprendizaje de máquina aplicado a grandes datos permite poderosas aplicaciones de IA que varían desde carros sin conductor hasta asistentes personales facilitados-por-voz. Pero no todas las formas de IA involucran aprendizaje de máquina que esté siendo aplicado a grandes datos. Es mejor comenzar con una definición *funcional* de IA. "Cualquier programa puede ser considerado IA si hace algo que nosotros normalmente pensaríamos de ello como inteligente en humanos," escribe el científico de computadores Kris Hammond. "Cómo el programa hace esto no es el problema, solo si es capaz de hacerlo. Esto es, ello es IA si es inteligente, pero no tiene que ser inteligente como nosotros."[6]

Según esta definición expansiva, la automatización de la rutina del computador, explícitamente definida como tareas de "procesos robóticos" tales como cobrar cheques y formas de pre-poblados de Recursos Humanos cuentan como IA. También lo hace la aplicación perspicaz de productos de la ciencia de datos, tales como usar un algoritmo predictivo del árbol de decisión para la asignación de grados de urgencia de los pacientes en la sala de urgencias. En cada caso, el algoritmo realiza una tarea previamente hecha solo por humanos. Aun así, es obvio que ningún caso involucra imitar la inteligencia humana, ni aplicar aprendizaje de máquina a conjuntos masivos de datos.

Comenzando con la definición de Hammond, es útil adoptar una estructura que distingue entre IA para *automatización* y AI para *aumentar* lo humano.

IA para automatización

IA ahora es capaz de automatizar tareas asociadas con conocimiento humano tanto *explícito* como *tácito*. El primero es el conocimiento del "libro de texto" que puede ser documentado en manuales y libros de reglas. Es crecientemente práctico capturar tal conocimiento en códigos de computador para lograr la automatización robótica de procesos [robotic process automation (RPA)]: construir "robots" de software que realicen tareas aburridas, repetitivas, propensas a error, o consumidoras de tiempo, tales como procesamiento de cambios de direcciones, reclamos de seguros, facturas de hospitales, o formularios de recursos humanos. Dado que RPA disfruta tanto de riesgo bajo como de retorno económico alto, a menudo es un punto natural de partida para las organizaciones que desean lograr eficiencias y ahorros de costos mediante IA. Idealmente, también puede liberar valioso tiempo humano para tareas más complejas, significativas, u orientadas-al-cliente.

El conocimiento tácito puede ingenuamente ser visto como impermeable a la automatización de la IA: es automático, "saber cómo" intuitivo que es aprendido mediante hacer, no solamente mediante estudio o seguimiento de reglas. La mayoría del conocimiento humano es conocimiento tácito: una enfermera intuye que un muchacho tiene grupa, un bombero que con instinto siente que una edificación incendiada está a punto de colapsar, o un científico de datos que intuye que una variable refleja una relación cercana sospechosa. Aun así, la capacidad de las aplicaciones de IA para automatizar tareas asociadas con conocimiento humano tácito está progresando rápidamente. Los ejemplos incluyen reconocimiento facial, detección de emociones, conducción de carros, interpretación de lenguaje hablado, lectura de texto, escritura de reportes, calificar documentos de estudiantes, e incluso preparar personas para citas. En muchos casos, las formas más nuevas de IA pueden realizar tales tareas más exactamente que los humanos.

La misteriosa calidad de tales aplicaciones hace tentador concluir que los computadores están implementando – o acercándose rápidamente a – un tipo de inteligencia humana en el sentido de que "entienden" lo que están haciendo. Ello es una ilusión. Los algoritmos "demuestran el conocimiento tácito similar-al-humano" solo en el sentido débil de que están construidos o entrenados usando datos que codifican el conocimiento tácito de un número grande de humanos que trabajan detrás de las escenas. El término "aprendizaje de máquina humano-en-el-lazo" a menudo es usado para connotar este proceso.[7] Si bien los grandes datos y el aprendizaje de máquina permiten la creación de algoritmos que puedan capturar y transmitir significado, esto es muy diferente de entender u originar significado.

Es tentador concluir que los computadores están implementando – o acercándose rápidamente a – un tipo de inteligencia humana en el sentido de que "entienden" lo que están haciendo. Ello es una ilusión.

Dado que la automatización elimina la necesidad de involucramiento humano, ¿por qué los sistemas autónomos de IA requieren diseño centrado-en-lo-humano? Hay varias razones:

Relevancia de la meta: Los productos de la ciencia de datos y las aplicaciones de la IA son más valiosos cuando perspicazmente están diseñados para satisfacer las necesidades de los humanos usuarios finales. Por ejemplo, teclear "área de Polonia" en el motor de búsqueda de Bing regresa la respuesta literal (129,728 millas cuadradas) junto con la nota: "Casi igual al tamaño de Nevada." La respuesta numérica es más exacta, pero la respuesta intuitiva a menudo será más útil.[8] Esto ejemplifica el punto más amplio de que "óptimo" desde la perspectiva de los algoritmos de computador no necesariamente es igual a "óptimo" desde la perspectiva del usuario final o de la sicología.

Manos libres. Muchos sistemas de IA pueden operar la mayor parte del tiempo con "piloto automático," pero requieren intervención humana en situaciones excepcionales o ambiguas que requieren sentido común o entendimiento contextual. El diseño centrado-en-lo-

humano es necesario para asegurar que este "manos libres" del computador para con el humano ocurra cuando se deba, y que vaya sin problemas cuando ello ocurra. Aquí hay un ejemplo personal de, admitámoslo, bajas apuestas de cómo la IA puede dar origen a "estupidez artificial" si el libre de manos no va bien. Recientemente pedí un taxi para un viaje que requería solo sentido común y una minúscula cantidad de conocimiento local – conducir hacia un bulevar principal. Aun así, el conductor se perdió porque estaba siguiendo las indicaciones (como se vio después, distorsionadas) de la aplicación del teléfono inteligente. Una alarma de "baja confianza" o de "potencialmente interferencia alta" puede haber llevado al conductor a repensar sus acciones más que eliminar su sentido común a favor del indicador algorítmico.

Esto ilustra el problema general conocido como "la paradoja de la automatización":[9] Entre más confiados nos volvemos respecto de la tecnología, menos preparados estamos para asumir el control en los casos excepcionales cuando la tecnología falla. El problema es espinoso porque las condiciones bajo las cuales los humanos tienen que asumir el control requieren *más*, no menos, habilidad que las situaciones que pueden ser manejadas por algoritmos – y las tecnologías de la automatización precisamente pueden erosionar las habilidades que se necesitan en tales escenarios. Mantener las habilidades humanas suficientemente frescas para manejar tales situaciones puede algunas veces involucrar confiar en la automatización menos que lo que la tecnología hace práctico. Una vez más, "óptimo" desde una perspectiva tecnológica estrecha puede diferir de "óptimo" para un sistema humano-computador.

> "La tecnología es la parte fácil. La parte dura es descifrar las estructuras sociales e institucionales alrededor de la tecnología."

Lazos de retroalimentación. Las decisiones algorítmicas automatizadas pueden reflejar y amplificar patrones indeseables en los datos que los entrenan. Un ejemplo reciente vívido es Tay, un robot de charlas diseñado para aprender acerca del mundo mediante conversaciones con sus usuarios. El robot de charlas tuvo que ser apagado 24 horas después que bromistas lo entrenaron para pronunciar declaraciones racistas, sexistas, y fascistas.[10] Otros ejemplos de algoritmos que reflejan y amplifican sesgos sociales indeseables son por ahora ubicuos. Por tales razones, hay un creciente pedido para que el diseño de robots de chateo y motores de búsqueda se optimicen no solo por velocidad y exactitud algorítmica, sino también por el comportamiento del usuario y los sesgos sociales codificados en los datos.[11]

Impacto sicológico. Así como el comportamiento del usuario puede impactar los algoritmos, también los algoritmos pueden menoscabar el comportamiento del usuario. Dos problemas contemporáneos serios ilustran. Primero, se está volviendo crecientemente claro que las aplicaciones de entretenimiento y medios de comunicación social facilitadas-por-IA pueden menoscabar el bienestar humano en una serie de maneras. La verificación compulsiva del correo electrónico puede causar que las personas se estafen a sí mismas en el sueño y se distraigan en el trabajo; el uso excesivo de los medios de comunicación social ha sido vinculado con sentimientos de infelicidad y "miedo de perderse"; y el personal interno de Silicon Valley crecientemente se preocupa porque las mentes de las personas sean "secuestradas" por tecnologías adictivas.[12]

Segundo, hay creciente preocupación porque el filtrado colaborativo de noticias y comentarios pueda llevar a "filtrar las burbujas" y a "comunidades de opinión de compuertas epistémicas." En su libro reciente *#Republic*, el erudito legal Cass Sunstein argumenta que esto puede exacerbar la polarización del grupo y menoscabar la deliberación razonada, un pre-requisito del bien funcionamiento de la democracia. Sugiere que los motores de recomendación de los medios de comunicación social tengan imbuida una forma de diseño centrado-en-lo-humano: los descubrimientos espontáneos, casuales, de relatos alternativos de noticias y piezas de opinión para ayudar a alejar la polarización y el pensamiento de grupo.[13] Sunstein hace analogía de esto con los encuentros casuales de modificación-de-la-perspectiva y los descubrimientos característicos de vivir en un entorno humano denso, diverso, transitable.

En resumen, puede ser contraproducente desplegar sistemas autónomos de IA tecnológicamente sofisticados, sin un enfoque correspondientemente sofisticado para el diseño centrado-en-lo-humano. Tal y como John Seely Brown de manera precisa comentó, "La tecnología es la parte fácil. La parte dura es descifrar las estructuras sociales e institucionales alrededor de la tecnología."[14]

Con todo, la automatización es solo parte de la historia. Los algoritmos también pueden ser usados para aumentar las capacidades humanas cognitivas – tanto el "pensar rápido" del sistema 1, como el "pensar lento" del sistema 2. Es posible lograr formas de inteligencia colectiva humano-computador – provisto que adoptamos un enfoque centrado-en-lo-humano para la IA.

IA para el pensamiento lento aumentado

Desde hace tiempo los sicólogos han sabido que incluso algoritmos sencillos pueden supera los juicios del experto en tareas predictivas que varían desde hacer diagnósticos médicos hasta estimar las probabilidades que quien esté en libertad condicional reincidirá, hasta explorar jugadores de béisbol para que suscriban riesgos de seguros. El campo fue iniciado en 1954 con la publicación del libro *Clinical Versus Statistical Prediction* por el sicólogo y filósifo Paul Meehl.

Meehl fue un héroe para el joven Daniel Kahneman, el autor de *Thinking, Fast y Slow* [Pensando, rápido y lento],[15] cuyo trabajo con Amos Tversky descubrió la sorprendente tendencia de la mente humana hacia confiar en narrativas intuitivamente coherentes pero predictivamente dudosas, más que en las valoraciones lógicas de la evidencia. Economistas comportamentales tales como Richard Thaler señalan que esta característica sistemática de la sicología humana resulta en mercados y procesos de negocio, persistentemente ineficientes, que pueden ser racionalizados mediante el uso de toma de decisiones basada-en-algoritmos – "jugando Moneyball."[16] Así como las gafas compensan la visión miope, los datos y los algoritmos pueden compensar la miopía cognitiva.

El trabajo de Meehl y de Kahneman implica que, en muchas situaciones, los algoritmos deben ser usados para *automatizar* decisiones. Los humanos demasiado seguros tienden a pasar por alto los algoritmos predictivos más a menudos que como deben.[17] Cuando es posible, es por lo tanto mejor emplear juicio humano en el diseño de algoritmos, y remover a los humanos de la toma de decisiones caso-por-caso. Pero esto no siempre es posible. Por ejemplo, la justicia procedimental implica que sería inaceptable reemplazar al juez que toma decisiones sobre libertad condicional por los outputs mecánicos de un algoritmo de predicción de reincidencia. El segundo problema es de naturaleza epistémica. Muchas decisiones, tales como hacer un diagnóstico médico complejo, suscribir un riesgo de seguro raro, tomar una decisión importante, y similares, no están asociados con un cuerpo suficientemente rico de datos históricos para permitir la construcción de un algoritmo predictivo suficientemente confiable. En tales escenarios, un algoritmo imperfecto puede ser usado no para *automatizar* decisiones, sino para generar puntos de anclaje para *aumentar y mejorar* las decisiones humanas.

¿Cómo puede funcionar esto? Una ilustración sugestiva viene del mundo del ajedrez. Varios años después que el Deep Blue de IBM derrotó al campeón mundial de ajedrez Garry Kasparov, se realizó una competencia de "ajedrez de estilo libre," en la cual podría competir cualquier combinación de jugadores de ajedrez humanos y computarizados. La competencia terminó con una victoria molesta que Kasparov subsiguientemente discutió:

> Se reveló que el ganador no es un gran maestro con un PC del estado-de-arte sino un par de jugadores de ajedrez americanos aficionados que usaron tres computadores al mismo tiempo. Su habilidad para manipular y "entrenar" sus computadores para mirar muy profundamente las posiciones de manera efectiva contrarrestaron el entendimiento superior de ajedrez de sus grandes maestros oponentes y del enorme poder computacional de los otros participantes. Humano débil + máquina + mejor proceso fue superior a un solo computador fuerte y, más destacado aún superior a humano fuerte + máquina + proceso inferior… La orientación humana estratégica combinada con la agudeza práctica de un computador fue abrumadora.[18]

La idea de que humano débil + máquina + mejor proceso supera a humano fuerte + máquina + proceso inferior ha sido denominada la "ley de Kasparov." El corolario es que el diseño centrado-en-lo-humano es

necesario tanto para la creación como para el despliegue de algoritmos que tienen la intención de mejorar el juicio experto. Así como una ciclista se puede desempeñar mejor con una bicicleta que diseñada para ella y que ella ha sido entrenada para usar, una experta puede tomar mejores decisiones con un algoritmo construido con sus necesidades en mente, y para el cual ella haya sido entrenada para usar.[19]

Con ese fin, los algoritmos de IA centrados-en-lo-humano deben reflejar confiablemente la información, las metas, y las restricciones que quien toma la decisión tiende a sopesar cuando llega a una decisión; los datos deben ser analizados desde una posición de dominio y conocimiento institucional, y un entendimiento del proceso que los genera; el diseño del algoritmo debe anticipar las realidades del entorno en el cual va a ser usado; debe evitar los predictores socialmente molestos; debe ser revisado-por-pares o auditado para asegurar que sesgos no-deseados no han sido inadvertidamente deslizados en él; y debe estar acompañado por medidas de confianza y mensajes de "por qué" (idealmente expresados en lenguaje intuitivo) explicando por qué un cierto indicador algorítmico es lo que es. Por ejemplo, uno no desearía recibir un indicador algorítmico de caja negra de las posibilidades de una enfermedad seria sin la capacidad para investigar las razones por las cuales el indicador es lo que es.

Pero incluso esos tipos de consideraciones del diseño del algoritmo no son suficientes. El *entorno* general de decisión – que incluye tanto el algoritmo como a los humanos que toman decisiones – debe similarmente estar bien diseñado. Así como los ganadores de ajedrez de estilo libre triunfaron a causa de sus profundas familiaridad y experiencia con tanto el ajedrez como sus programas de ajedrez, los usuarios finales del algoritmo deben tener un entendimiento suficientemente detallado de su herramienta para usarla efectivamente. Los supuestos, limitaciones, y características de los datos del algoritmo deben por consiguiente ser comunicados de manera clara mediante información escrita y visual. Además, guías y las reglas de negocio deben ser establecidas para cubrir las predicciones en las prescripciones y sugerir cuándo y cómo el usuario final puede ya sea pasar por alto el algoritmo o complementar sus recomendaciones con otra información. Los usuarios finales también deben ser entrenados para "pensar lento," más que como estadísticos. Los sicólogos Philip Tetlock y Barbara Mellors han encontrado que quienes toman decisiones de entrenamiento en razonamiento probabilístico y evitan sesgos cognitivos mejoran sus capacidades para la elaboración de pronósticos.[20] Construir algoritmos exactos no es suficiente; también es esencial el diseño centrado-en-el-usuario.

3D: Datos, digital y diseño para el pensamiento rápido aumentado

El valor económico proviene no de los algoritmos de IA, sino de los algoritmos de IA que hayan sido apropiadamente diseñados para, y adaptados a, los entornos humanos. Por ejemplo, considere el "problema de la última milla" de los algoritmos predictivos: ningún algoritmo rendirá valor económico a menos que de la manera apropiada se actúe en él para orientar resultados. Si bien esto es una perogrullada, también es una de las cosas más fáciles para que las organizaciones se equivoquen. Un estudio reciente estimó que el 60 por ciento de los proyectos de "grandes datos" fallan en que no se pueden operacionalizar.[21]

Un buen ejemplo de operacionalización del modelo es el algoritmo predictivo usado para clasificar todas las edificaciones en New York City en orden de peligrosidad. Antes del despliegue del algoritmo, casi el 10 por ciento de las inspecciones de edificaciones resultaron en una orden de desocupación. Después del despliegue, el número se elevó al 70 por ciento.[22] Este es un ejemplo clásico de analíticas predictivas usadas para mejorar la toma de decisiones del "Sistema 2," tal y como se discutió en la sección anterior. Aún más valor puede derivarse mediante la aplicación de lo que los economistas comportamentales llaman *arquitectura de elección*, alias "empujones."[23] Considere los riesgos que sean ambiguos o no suficientemente peligrosos (todavía) para requerir la visita de los cuadros limitados de la ciudad de inspectores de edificaciones. Tales riesgos menores podrían ser incitados a "auto-curarse" mediante, por ejemplo, cartas de empuje que hayan sido probadas-en-el-campo y optimizadas usando pruebas aleatorias controladas [randomized controlled trials (RCT)]. Estrategias análogas de "empuje [fuertemente] lo peor, empuje [suavemente] el resto" pueden ser adoptadas para los algoritmos diseñados para identificar restaurantes antihigiénicos, programas ineficientes, lugares de trabajo inseguros, episodios de residuos, fraude, abuso, o gasto o no-cumplimiento con la política tributaria.

En ciertos casos, la aplicación de la arquitectura de elección será crucial para el éxito económico y la aceptabilidad social de un proyecto de IA. Por ejemplo, el estado de Nuevo Méjico recientemente adoptó un algoritmo de aprendizaje de máquina diseñado para señalar los beneficiarios de seguros de desempleo que *relativamente* sea probable estén recaudando de manera inapropiada grandes beneficios de seguros de desempleo [unemployment insurance (UI)]. La palabra "relativamente" es importante. Si bien los casos de puntuación más alta fueron muchas veces más probables que el promedio para estar recaudando de manera inapropiada beneficios de UI, la mayoría fueron (inevitablemente) falsos positivos. Este resultado contra intuitivo es conocido como la "paradoja del falso positivo."[24] La implicación crucial es que usar ingenuamente el algoritmo para recortar beneficios dañaría a un número grande de ciudadanos que genuinamente los necesitan. Más que adoptar esta estrategia ingenua, el estado por consiguiente probó-en-el-campo una serie de mensajes emergentes [pop-up] de empuje en las pantallas del computador de los beneficiarios de UI realizando sus certificaciones semanales. El más efectivo de tales mensajes recortó en la mitad los pagos inapropiados: informó a los beneficiarios que "99 de 100 personas en <su país> reportan de manera exacta las ganancias cada semana."[25]

SIGNIFICADOS PASADOS Y PRESENTES DE "IA"

Si bien el término "IA" ha hecho un regreso importante, ha llegado a significar algo bastante diferente de lo que sus fundadores tuvieron en mente. Las tecnologías de IA de hoy no son generalmente máquinas de pensamiento inteligente; son aplicaciones que les ayudan a los humanos a pensar mejor.

El campo de la inteligencia artificial se remonta a un lugar y tiempo específicos: una conferencia tenida en Dartmouth University en el verano de 1956. La conferencia fue convocada por John McCarthy, quien acuñó el término "inteligencia artificial" y lo definió como la ciencia de crear máquinas "con la capacidad para lograr metas en el mundo."[26]

La definición de McCarthy todavía es muy útil. Pero quienes asistieron a la conferencia – incluyendo figuras legendarias como Marvin Minsky, Alan Newell, Claude Shannon, y Herbert Simmon – aspiraron a una meta mucho más ambigua: implementar una versión completa del pensamiento y el lenguaje humano con tecnología de computador. En otras palabras, desearon crear inteligencia artificial *general*, modelada en inteligencia humana general. Su propuesta señaló:

> El estudio es para proceder con base en la conjetura de que *cada* aspecto del aprendizaje o cualquier otra característica de la inteligencia puede en principio ser tan precisamente descrito que una máquina pueda simularlo. Se intentará encontrar cómo hacer que las máquinas usen lenguaje, a partir de abstracciones y conceptos, resuelvan tipos de problemas ahora reservados para humanos, y se mejoren a sí mismos.[27]

La propuesta procedió a señalar, "Nosotros pensamos que un avance importante se puede hacer en uno o más de esos problemas si grupos de científicos cuidadosamente seleccionadas trabajan en él durante un verano." Este optimismo puede verse con sorpresa en retrospectiva. Pero vale la pena recordar que los autores estaban escribiendo en el apogeo de tanto la sicología conductista de B. F. Skinner como de la escuela de filosofía de la lógica positivista. En este clima natural, era natural asumir que el pensamiento humano en últimas era una forma de cálculo lógico. Nuestro entendimiento tanto de la sicología humana como de los desafíos de la codificación del conocimiento en lenguajes perfectamente lógicos ha evolucionado considerablemente desde los años 1950.

Es una reveladora nota histórica que Minsky subsiguientemente asesoró al director Stanley Kubrick durante la adaptación de la película de la novela de Arthur C. Clarke *2001: A Space Odyssey* [2001: Una odisea en el espacio]. El personaje más memorable de esa historia fue HAL – una máquina sapiente capaz de pensamiento conceptual, razonamiento de sentido común, y un comando fluido del lenguaje humano. Minsky y los otros asistentes a la Darmounth Conference consideraron que tales computadores generalmente inteligentes estarían disponibles para el año 2001.

Continúa...

SIGNIFICADOS PASADOS Y PRESENTES DE "IA" (continuación)

Hoy, IA denota una colección de tecnologías que, parafraseando la definición original de McCarthy, sobresale en *tareas* específicas que previamente solo podrían ser desempeñadas por humanos. Si bien para los comentaristas es común señalar que tecnologías tales como que el sistema de reconocimiento facial DeepFace o el AlphaGo de DeepMind están "modelados en el cerebro humano" o pueden "pensar igual a como los humanos lo hacen," son declaraciones que conducen a engaño. Un punto obvio es que las tecnologías de IA de hoy – y todas las que están en el horizonte previsible – son soluciones puntuales de AI, *estrechas*. Un algoritmo diseñado para conducir un carro es inútil para diagnosticar un paciente, y viceversa.

Además, tales aplicaciones están lejos de la visión popular de los computadores que implementan (súper) pensamiento humano. Por ejemplo, los algoritmos de redes neurales de aprendizaje profundo pueden identificar tumores en rayos-X, etiquetar fotogramas con frases en inglés, distinguir entre razas de animales, y distinguir personas que genuinamente se están riendo de las que están fingiendo – a menudo más exactamente que como nosotros podemos.[28] Pero esto no involucra representar algorítmicamente conceptos tales como "tumor," "pinscher," y "sonrisa." Más aún, los modelos de redes neurales de aprendizaje profundo están entrenados en números grandes de fotografías digitalizadas que ya han sido etiquetadas por humanos.[29] Tales modelos ni imitan ni simulan el cerebro. Son modelos predictivos – similares a modelos de regresión – típicamente entrenados en millones de ejemplos y que contienen millones de parámetros no-interpretables. La tecnología puede realizar tareas hasta ahora desempeñadas solo por humanos; pero ello no resulta en emular el cerebro humano o imitar la mente humana.

Si bien tales aplicaciones de IA orientadas-a-datos tienen aplicaciones prácticas masivas y potencial económico, también son "rígidas" en el sentido de que carecen de conciencia contextual, entendimiento casual, y capacidades de razonamiento de sentido común. Una implicación crucial es que ellas no pueden confiar en escenarios de "cisne negro" o entornos significativamente diferentes a los cuales están entrenados. Así como un algoritmo de calificación den crédito entrenado en datos acerca de consumidores de los Estados Unidos no rendiría un puntaje confiable para un inmigrante de otro país, un carro auto-dirigido entrenado en Palo Alto no necesariamente se desempeñaría igual de bien en Pondicherry.

La naturaleza centrada-en-lo-humano de la arquitectura de la elección puede por lo tanto permitir aplicaciones de IA que sean tanto económicamente benéficas como pro-sociales.[30] Además, el caso para la arquitectura de la elección es más fuerte que nunca en nuestra era de grandes datos y tecnologías digitales ubicuas. Los datos comportamentales de grano fino de grandes poblaciones crecientemente pueden permitir intervenciones personalizadas apropiadas para casos individuales. Empapar nuestras tecnologías digitales siempre presentes con la arquitectura de la elección puede mejorar tanto el compromiso como los resultados. Los utilizables de salud son un ejemplo familiar. Expertos prominentes de salud comportamental señalan que tales dispositivos son facilitadores – pero no orientadores – de mejores comportamientos de salud.[31] Usar tales utilizables para solamente obtener datos y generar reportes de información simplemente no es suficiente para solicitar que la mayoría de nosotros sigamos y cambiemos nuestros comportamientos. Una estrategia más comprometedora es usar los datos obtenidos mediante utilizables para señalar, informar y personalizar tales tácticas de empuje como comparaciones de pares, contratos de compromiso, intervenciones de gamificación, y programas de formación de hábitos.[32]

Esto ilustra un principio general que puede ser denominado "3D": los *datos* y la tecnología *digital* son facilitadores, el *diseño* sicológicamente informado también es necesario para orientar mejores compromisos y resultados. El pensamiento de 3D puede permitir productos innovadores y modelos de negocio. Considere, por ejemplo, los datos telemáticos que emanan de los carros conectados a Internet de las Cosas, los cuales los aseguradores ya usan para fijar más precisamente el precio de los contratos de seguros de personas y de vehículos comerciales. Estos datos también pueden ser usados para estimular la prevención de pérdida; un conductor que es mujer joven puede obtener un descuento en su póliza de seguros de

automóviles si sigue las prescripciones generadas-por-datos para mejorar sus comportamientos de conducción. La arquitectura de selección permite una idea adicional: las herramientas de generación de lenguaje natural podrían ser usadas para automáticamente producir reportes ricos-en-datos que contengan tanto consejos útiles como también mensajes de empuje de comparación de pares. Por ejemplo, ser informados de que su conducción-en-autopista es más riesgosa que la mayoría de sus paros puede ser una manera altamente efectiva, de bajo costo, para promover la conducción segura. Tales estrategias pueden permitir que los aseguradores estén menos centrados-en-el-producto y más centrados-en-el-cliente de una manera que beneficie a la compañía, al tomador de la póliza, y a la sociedad en su conjunto.

Ya sea que se tenga la intención de que sean para la automatización o para la aumentación humana, los sistemas de IA es más probable que rindan beneficios económicos y aceptabilidad social si las necesidades del usuario y sus factores sicológicos son tenidos en cuenta. El diseño puede ayudar a cerrar la brecha entre los *resultados* del algoritmo de IA y los *resultados* mejorados mediante permitir mejores modos de colaboración humano-computador. Por consiguiente, es adecuado darle la última palabra a Garry Kasparov, en su libro reciente, *Deep Thinking*: "Muchos trabajos continuarán siendo perdidos por la automatización inteligente. Pero si usted está buscando un campo que continuará en auge durante muchos años, vaya a la colaboración humano-máquina y a la arquitectura y diseño de procesos."

Tanto figurativa como literalmente, la última palabra es: diseño.

JAMES GUSZCZA es el US Chief Data Scientist de Deloitte, con sede en Santa Monica, California.

MÁS INTELIGENTES JUNTOS.

página 36

[1] Vea Terry Winograd, "Thinking machines: Can there be? Are we?," originally published in James Sheehan and Morton Sosna, *The Boundaries of Humanity: Humans, Animals, Machines* (Berkeley: The University of California Press, 1991).

[2] Vea, por ejemplo, Matthew Hutson, "Even artificial intelligence can acquire biases against race and gender," *Science*, April 13, 2017; "Fake news: You ain't seen nothing yet," *Economist*, July 1, 2017; Paul Lewis, "'Our minds can be hijacked': The tech insiders who fear a smartphone dystopia," *Guardian*, October 6, 2017; Holly B. Shakya and Nicholas A. Christakis, "A new, more rigorous study confirms: The more you use Facebook, the worse you feel," *Harvard Business Review*, April 10, 2017.

[3] El destacado investigador de IA, Yann LeCun, discute la promoción y la realidad de la IA en una entrevista con James Vincent, "Facebook's head of AI wants us to stop using the Terminator to talk about AI," *The Verge*, October 26, 2017.

[4] Para discusión adicional de este tema, vea James Guszcza, Harvey Lewis, and Peter Evans-Greenwood, "*Cognitive collaboration: Why humans and computers think better together*," Deloitte University Press, January 23, 2017.

[5] Vea Don Norman, *The Design of Everyday Things* (New York: Basic Books, 2013).

[6] Vea Kris Hammond, "What is artificial intelligence?," *Computerworld*, April 10, 2015.

[7] Vea Lukas Biewald, "Why human-in-the-loop computing is the future of machine learning," *Computerworld*, November 13, 2015.

[8] Búsqueda realizada en Bing en octubre 11, 2017. Este es el resultado del trabajo liderado por los científicos cognitivos Dan Goldstein y Jake Hoffman sobre ayudar a las personas a comprender mejor los números grandes (vea Elizabeth Landau, "How to understand extreme numbers," *Nautilus*, February 17, 2017).

[9] Debo este punto a Harvey Lewis. Para una discusión, vea Tim Harford, "Crash: How computers are setting us up for disaster," *Guardian*, October 11, 2016.

[10] Vea Antonio Regalado, "The biggest technology failures of 2016," *MIT Technology Review*, December 27, 2016.

[11] Many examples of algorithmic bias are discussed in April Glaser, "Who trained your AI?," *Slate*, October 24, 2017.

[12] Vea las referencias en la primera nota final. En su reciente libro *The Distracted Mind: Ancient Brains in a High-Tech World* (MIT Press, 2016), Adam Gazzaley y Larry Rosen exploran las razones evolutivas y neurocientíficas por las cuales nosotros estamos "cableados para" ser distraídos por la tecnología digital, y exploran las intervenciones cognitivas y comportamentales para promover relaciones más saludables con la tecnología.

[13] *#Republic* by Cass R. Sunstein (New Jersey: Princeton University Press, 2017).

[14] Vea John Seely Brown, "Cultivating the entrepreneurial learner in the 21st century," March 22, 2015. Available at www.johnseelybrown.com.

[15] Vea Daniel Kahneman, *Thinking, Fast and Slow* (Farrar, Straus and Giroux, 2011).

[16] Vea Richard Thaler and Cass Sunstein, "Who's on first: A review of Michael Lewis's 'Moneyball: The Art of Winning an Unfair Game,'" University of Chicago Law School website, September 1, 2003.

[17] Vea Berkeley J. Dietvorst, Joseph P. Simmons, and Cade Massey, "Algorithm aversion: People erroneously avoid algorithms after seeing them err," *Journal of Experimental Psychology*, 2014. En un trabajo reciente, Dietvorst y Massey han encontrado que los humanos están más propensos a acoger la toma de decisiones algorítmica si incluso se les da una pequeña cantidad de latitud para ocasionalmente pasarla por alto.

[18] Garry Kasparov, "The chess master and the computer," *New York Review of Books*, February 11, 2010.

[19] Esta analogía no es accidental. Steve Jobs describió los computadores como "bicicletas para la mente" en el documental de 1990 *Memory & Imagination: New Pathways to the Library of Congress*. Clip disponible en "Steve Jobs, 'Computers are like a bicycle for our minds.' - Michael Lawrence Films," YouTube video, 1:39, posted by Michael Lawrence, June 1, 2006.

[20] Esto fue discutido por Philip Tetlock in *Superforecasting: The Art and Science of Prediction* (Portland: Broadway Books, 2015).

[21] Vea Gartner, "Gartner says business intelligence and analytics leaders must focus on mindsets and culture to kick start advanced analytics," press release, September 15, 2015.

[22] Vea Viktor Schönberger and Kenneth Cukier, "Big data in the big apple," *Slate*, March 6, 2013.

[23] La arquitectura de elección puede ser vista como el diseño centrado-en-lo-humano de entornos de elección. La meta es hacer que para nosotros sea fácil tomar mejores decisiones mediante el pensamiento automático del "Sistema 1," más que mediante la fuerza de voluntad o el esfuerzo del pensamiento del Sistema 2. Por ejemplo, si una persona es automáticamente matriculada en un plan de ahorros para retiro, pero puede salir de él, es considerablemente más probable que ahorren más

para retiro, pero puede salir de él, es considerablemente más probable que ahorren más para el retiro que si por defecto se establece el no-matricularse. Desde la perspectiva de la economía clásica, tales detalles menores en la manera como la información es proporcionada o se organizan las elecciones - Richard Thaler, Nobelist y co-autor de *Nudge: Improving Decisions About Health, Wealth and Happiness* (Yale University Press, 2008) las denomina "factores supuestamente irrelevantes" – deben tener poco o ningún efecto en las decisiones de las personas. Aun así, la perspectiva central de la arquitectura de la selección es que la larga lista de peculiaridades cognitivas y comportamentales

humanas sistemáticas pueden ser usadas como elementos de diseño para los entornos de elección que hagan fácil y natural que las personas tomen las elecciones que tomarían si tuvieran recursos ilimitados de fuerza de voluntad y cognitivos. En su libro *Misbehaving* (W. W. Norton & Company, 2015), Thaler reconoció la influencia del enfoque de Don Norman para el diseño centrado-en-lo-humano al escribir *Nudge* [empujar].

[24] Cuando la tasa base a nivel-de-población de un evento es baja, y la prueba o el algoritmo usado para señalar el evento es imperfecto, los falsos positivos pueden superar numéricamente los verdaderos positivos. Por ejemplo, supóngase que el 2 por ciento de la población tiene una enfermedad rara y que la prueba usada para identificarla es 95 por ciento exacta. Si un paciente específico proveniente de esta población prueba que es positivo, esa persona tiene aproximadamente un 29 por ciento de posibilidad de tener la enfermedad. Esto resulta de la aplicación simple del Teorema de Bayes. Un sesgo cognitivo bien conocido es "negligencia de la tasa base" – muchas personas asumirían que la posibilidad de tener la enfermedad no es 29 por ciento sino 95 por ciento.

[25] Para más detalles, vea Joy Forehand and Michael Greene, "Nudging New Mexico: Kindling compliance among unemployment claimants," *Deloitte Review* 18, January 2016.

[26] Vea Mitesh Patel, David Asch, and Kevin Volpp, "Wearable devices as facilitators, not drivers, of health behavior change," *Journal of the American Medical Association* 313, no. 5 (2015).

[27] Para discusión adicional de los temas contenidos en esta sección, vea, see James Guszcza, "The last-mile problem: How data science and behavioral science can work together," *Deloitte Review* 16, January 2015.

[28] McCarthy definió inteligencia artificial como "la ciencia y la ingeniería de elaborar máquinas inteligentes, especialmente programas de computador inteligentes," y definió inteligencia como "la parte computacional de la capacidad para lograr metas en el mundo." Observó que #tipos y grados diversos de inteligencia ocurren en personas, muchos animales, y algunas máquinas." Vea John McCarthy, "What is artificial intelligence?," Stanford University website, accessed October 7, 2017.

[29] La propuesta original se puede encontrar en John McCarthy, Marvin L. Minsky, Nathaniel Rochester, and Claude E. Shannon, "A proposal for the Dartmouth Summer Research Project on artificial intelligence," *AI Magazine* 27, no. 4 (2006).

[30] Este tema es explorado en Jim Guszcza, David Schweidel, and Shantanu Dutta, "The personalized and the personal: Socially responsible uses of big data," *Deloitte Review* 14, January 2014.

[31] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun, "Delving deep into rectifiers: Surpassing human-level performance on ImageNet classification," arXiv, February 6, 2015.

[32] Es común que tales algoritmos fallen en ciertos casos ambiguos que pueden ser correctamente etiquetados por expertos humanos. Esos nuevos puntos de datos pueden luego ser usados para entrenar los modelos, resultando en exactitud mejorada. Este círculo virtuoso de etiquetado humano y aprendizaje de máquina es denominado "computación humano-en-el-lazo." Vea, por ejemplo, Biewald, "Why human-in-the-loop computing is the future of machine learning."