LA ÉTICA DEL FUTURO

CENNYDD BOWLES

Traducido por
SILVIA CALVET y GABY PRADO

First published in 2018 by NowNext Ltd, Studio 19 Unit 50 Uplands B, Blackhorse Lane, London E17 5QJ, United Kingdom.

Cover image by Bernard Hermant. Typeset in Iowan Old Style.

ISBN 978-1-9996019-4-2 (paperback) / 978-1-9996019-5-9 (Kindle) / 978-1-9996019-6-6 (Apple Books)

Calon lân yn llawn daioni,
Tecach yw na'r lili dlos:
Dim ond calon lân all ganu,
Canu'r dydd a chanu'r nos.

Corazón puro lleno de bondad,
más justo que el bello lirio:
Solo un corazón puro puede cantar,
cantar de día y cantar de noche.

(Calon Lân, himno galés)

ÍNDICE

PREFACIO

La ética se ha ganado con razón su reputación de tema ridículamente aburrido. Gran parte del canon existente sobre ética suena al oído moderno como teólogos debatiendo el número de ángeles en el cielo. O no es relevante, o su prosa es túrgida y prolija, o ambas cosas. Los practicantes contemporáneos encuentran poca tracción en el mundo del pensamiento ético convencional. Sin embargo, hay mucho de valor, si tan sólo pudiera formularse de forma relevante. Y eso es exactamente lo que tiene en sus manos. Cennydd Bowles ha logrado mucho en este volumen único, legible y cercano.

La creciente ubicuidad en todos los aspectos de nuestro mundo del software y sus datos ha abierto la caja de Pandora llena de cuestiones éticas. Las preguntas no son nuevas - se han debatido durante siglos - pero están adoptando nuevas formas en sus manifestaciones digitales, y están adquiriendo una nueva magnitud, mucho mayor, en las esferas sociales, económicas y políticas contemporáneas.

Al igual que la Ley de Moore, el mundo del software crece a un ritmo exponencial, en lugar de lineal. Crecer a esa velocidad significa que las señalizaciones lejanas nos pasarán por delante mucho antes de lo que nuestra intuición nos sugeriría. Las cuestiones éticas que plantean las nuevas capacidades del software son numerosas y exigen respuestas ya.

En los círculos tecnológicos, los "éticos de sillón" discuten el problema del tranvía. Se trata de un ejercicio hipotético en el que uno

debe decidir -cambiando de vía- si un tranvía fuera de control debe matar a una joven madre o a un anciano. Es cierto que son los fabricantes de coches autoconducidos los se enfrentan a un problema real, pero los profesionales que se dedican al diseño y desarrollo de software cotidianos no suelen ser conscientes de las decisiones éticas que toman inconscientemente. Sin embargo, estas decisiones pueden tener ramificaciones comparables a las que tuvieron que afrontar los inventores del gas tóxico o la bomba atómica. La frecuencia y las consecuencias de las cuestiones éticas son hoy en día mucho más grandes que en el pasado.

En la era industrial, permitíamos que los vendedores nos atrajeran con el diseño de la publicidad, la colocación de productos y la señalización. La lentitud inherente al mundo atómico nos hizo creer que este tipo de persuasión era inofensiva. Hoy, sin embargo, analizando tus compras y gustos, Amazon puede adaptar su sitio a tu personalidad y hacerte comprar más cosas. Aunque esto es conceptualmente lo mismo que los anuncios de cartón en el pasillo del supermercado, su magnitud lo convierte en todo un nuevo universo de moralidad. Algunas de las cuestiones más fundamentales sobre el comportamiento aceptable se están sometiendo a escrutinio, y muchas de las respuestas tienen que cambiar. Y cuando el producto que se vende es un partido político o, peor aún, un golpe totalitario basado en el miedo, el cambio es imperativo.

Un programa informático de un tercero puede recopilar datos sobre ti sin que te des cuenta. Cada vez más, puede hacerlo sin que ni siquiera lo utilices. ¿Es legal? ¿Es correcto? ¿A quién pertenecen esos datos? ¿Qué pueden hacer con ellos? ¿Qué derechos tienes tú? ¿Y si esos datos son erróneos? ¿Y si terceros toman decisiones duraderas con esos datos? ¿Cómo puedes saberlo?

La afirmación de que la tecnología en sí misma es moralmente neutra es una reconfortante excusa esgrimida por muchos tecnólogos, pero las pruebas en contra de esa idea han alcanzado magnitudes asombrosas, y muchos pensadores expertos en tecnología digital se enfrentan a estos dilemas morales sin aparentes precedentes con nuevos ojos.

Delegar la autoridad en el software sin establecer mecanismos sólidos de retroalimentación es un punto común de fracaso. Claro, el algoritmo nos da una respuesta, pero ¿cómo sabemos que es correcta

y cómo la corregimos si no lo es? Para que los circuitos de retroalimentación sean eficaces, tienen que ser oportunos y procesables, y hay que comprometerse a actuar en consecuencia. En la tecnología predigital, la fuerte fuerza amortiguadora de la participación humana en los bucles de retroalimentación ayudaba a estabilizar el proceso. Los bucles de retroalimentación digitales pueden ser demasiado eficientes, haciendo que el sistema oscile y falle. Pero, en la mayoría de los casos, los mecanismos de retroalimentación son inexistentes.

Con las herramientas tecnológicas actuales, el software suele tomar decisiones basándose en datos recopilados por una aplicación lejana, propiedad de una empresa no afiliada, en un momento muy lejano. Incluso a las organizaciones más concienzudas les puede resultar imposible verificar que las decisiones que toman sus algoritmos son correctas, y mucho menos actualizarlas para que lo hagan mejor en el futuro. La víctima, por lo general una persona que solicita un préstamo, un empleo o ayuda a los veteranos, simplemente cae en la trampa.

Se está gestando una reacción adversa contra el mundo digital, y podríamos beneficiarnos de las lecciones de la historia. Hoy más que nunca necesitamos principios claros y útiles. A medida que nuestros artefactos digitales maduran, van mostrando capacidades inimaginables en ámbitos imprevistos. Los productos destinados a vender libros se han transformado en árbitros de nuestro entorno construido. Los productos destinados a que los adultos jóvenes se conozcan se han transformado en herramientas de lavado de cerebro y tambores tribales que nos llevan a la guerra.

Peor que una reacción violenta es lo contrario. Empresas como Amazon, Facebook y Cambridge Analytica están aplicando con entusiasmo herramientas digitales sobre todo el mundo. Estas organizaciones están dispuestas a utilizar tu información para obtener beneficios sin tener en cuenta tus intereses. Y los medios de comunicación, cuyos canales de distribución de la era industrial permitían cierto grado de distanciamiento, son ahora esclavos virtuales del clic. Incluso los principales medios de comunicación, como *The New York Times*, *The Atlantic* y *CNN*, tienen que explotar las "noticias" sensacionalistas para sobrevivir.

Los gobiernos también están jugando a este arriesgado juego. China está experimentando con un sistema de calificación social,

como una calificación crediticia, salvo que califica si eres o no una buena persona. El potencial de abuso es evidente. Los próximos pasos serán robots policía que se asomen de forma autónoma a tu habitación y decidan si te comportas correctamente.

Al comienzo de este libro, Cennydd nos presenta un panorama ético. Articula estos importantes principios éticos en el contexto de la tecnología moderna, para que tengan sentido para el profesional contemporáneo. A continuación, expone amplios ejemplos de los crecientes obstáculos éticos en el mundo real y nos muestra cómo aplicar el marco. Y lo que es más, lo hace interesante y aplicable.

Como no se trata de un problema determinista, no hay respuestas en blanco y negro. Pero hay muchas técnicas útiles que el diseñador debe dominar. El lector dispondrá de un sentido de la misión, herramientas conceptuales útiles y un mapa del camino a seguir.

Ahora más que nunca necesitamos buenos libros sobre ética. Los profesionales necesitan orientación sobre cómo pensar éticamente, cómo detectar opciones éticas y cómo resolver dilemas éticos. Eso es exactamente este libro, destinado a convertirse en un clásico.

—Alan Cooper,
21 Agosto 2018,
Petaluma, California.

PRÓLOGO A LA VERSIÓN EN ESPAÑOL

La Ética del Futuro es un libro que propone una introducción accesible a la ética del diseño y la tecnología dirigida a profesionales de estos campos. No se trata de un volumen para académicos, sino de una aproximación a la cuestión en la que prima la reflexión aplicada sobre las implicaciones de la tecnología en nuestras vidas y sobre el planeta en su totalidad, ahora y en el futuro, pero también sobre el propio acto de diseñar. La propuesta que hace Cennydd Bowles es viable para comenzar este recorrido en el que desarrollar habilidades y actitudes éticamente responsables.

Todos estos términos, ética, diseño y tecnología, albergan múltiples significados y connotaciones que Bowles se encarga de esclarecer de una manera practica y accesible para un público interesado en extender el campo de exploración crítica sobre su disciplina y en fortalecer la calidad de la reflexión sobre la misma, yendo más allá de las intuiciones, las dicotomías de blanco o negro y los lugares comunes. La exploración no es solo interesante por ser aplicada y práctica, sino también por ser practicable y aplicable.

A pesar de contener opiniones y perspectivas personales, este es un libro sobre sobre ética y no principalmente sobre la ética de su autor. La diferencia es sutil, pero importante y vale la pena detenerse brevemente en ella. Si fuera sólo sobre la ética de su autor, el libro se quedaría en los principios morales de Cennydd Bowles y su manera individual de ver las cosas. Sin embargo, en este libro, como suele ser

frecuente en la filosofía académica, Bowles entiende la ética de una manera más amplia y relacionada con una reflexión sobre los principios y valores que guían el comportamiento humano hacia lo que puede o debe verse como correcto, deseable, preferible o moralmente aceptable. De este modo el libro va más allá de la mera preferencia personal por una corriente ética o teoría filosófica en particular y entra, mediante la argumentación, en un espacio intersubjetivo y plural donde se discuten las distintas perspectivas desde las que abordar dilemas y cuestiones éticas.

Así, *La Ética del Futuro* no busca pontificar acerca de lo que está bien y lo que está mal, sino que busca comprender los fundamentos de los juicios y razonamientos que guían, podrían guiar, o en ciertos casos deberían guiar, nuestra conducta y práctica como profesionales. El libro busca sobre todo plantear preguntas y que seamos capaces de hacernos otras nuevas según la situación. ¿Por qué es mejor hacer esto que aquello? O ¿por qué justamente no deberíamos hacer esto otro? O ¿Por qué aunque no esté bien o no sea bueno, puede estar éticamente justificado hacer esto otro en determinados casos? O, de manera más concreta, ¿en qué casos es aceptable utilizar técnicas persuasivas como como los nudges (empujoncitos) para persuadir a posibles compradores? ¿Qué hay de éticamente problemático en la utilización de datos biométricos y comportamentales para generar predicciones acerca de una persona o colectivo?

Este es un libro sobre la tecnología actual y sobre el diseño de tecnologías emergentes. Diseñamos hoy, pero la materialización de los proyectos ocurrirá en el futuro. Por ello, este libro es necesariamente sobre el futuro. Como profesionales del diseño y desarrollo tenemos cierta capacidad para efectuar cambios en estos ámbitos. Sin embargo, hay muchos otros participantes además de las personas que trabajamos en el diseño y desarrollo.

Un error común al abordar la cuestión de la ética del diseño y la ingeniería es considerar a los equipos de diseño y desarrollo como entidades todopoderosas que consiguen que sus productos y servicios se comporten según lo previsto. Si estos productos o servicios son éticamente cuestionables o provocan consecuencias negativas, la culpa debe recaer en estos equipos, que son los responsables de los daños.

El filósofo de la tecnología Don Ihde habla de la "falacia del dise-

ñador", es decir "la noción de que un diseñador puede diseñar en una tecnología sus propósitos y usos". Ihde se refiere a una falacia porque es bien sabido que el consumo, la adopción y el uso de productos no son actos pasivos, sino procesos dinámicos en los cuales las personas interactúan con los productos de formas distintas a las previstas durante su diseño y desarrollo.

Todas las tecnologías superan nuestras intenciones iniciales y muestran muchas más posibilidades que las originalmente previstas. Estas posibilidades, por lo tanto, no están completamente determinadas antes del lanzamiento y ni siquiera lo están por las propiedades del producto en sí, sino por las personas y colectivos que los usan y consumen, quienes a su vez se hallan inmersos en prácticas arraigadas en contextos culturales y políticos específicos.

Los desafíos a los que los equipos de diseño y desarrollo se enfrentarán solo aumentarán con el tiempo. Por ello la ética debe convertirse en un foco ineludible de reflexión y en un compromiso permanente a lo largo de la carrera profesional de cada una de las personas involucradas en estos procesos.

—Ariel Guersenzvaig,
14 de Noviembre de 2023,
Barcelona.

AGRADECIMIENTOS

Estoy en deuda con:

Mi editor, Owen Gregory, y mis revisores técnicos, Thomas Wendt, Lydia Nicholas y Damien Williams. Ningún argumento débil sin refutar; ninguna afirmación rotunda sin rebatir; ninguna cita intimidatoria sin borrar.

Livia Labate, Eli Schiff, Paul Robert Lloyd y Tom Hume por sus inestimables comentarios como lectores. Lou Rosenfeld, Richard Rutter, Abby Covert, Nick Disabato y Brad Frost por sus consejos editoriales. Marcel Shouwenaar, Jeff Veen, Timo Arnall, y el Future of Life Institute por los permisos fotográficos. Christina Wodtke, Azeem Azhar y Jon Kolko por sus halagadoras reseñas, y a Alan Cooper por su amable prólogo.

A los asistentes al retiro *Juvet AI* por las conversaciones inspiradoras y el asombro infantil ante las Borealis, y a todos los demás que trabajan en el campo de la ética de las tecnologías emergentes. Vuestro trabajo es vital y profundo; espero haber hecho honor a vuestras ideas.

A mi familia y amigos. Gracias en particular a Sascha Auerbach y Andrew Fox por su continuo apoyo (líquido).

Un agradecimiento especial a mis traductoras Silvia Calvet y Gaby Prado, y a Ariel Guersenzvaig por su valioso prólogo para la versión española.

Finalmente, mi esposa, Anna. Mi estrella guía.

CAPÍTULO 1
PROBLEMAS EN EL PARAÍSO

Los sueños utópicos del inicio del ciberespacio no se han hecho realidad. El paraíso se ha cercado y recalificado. Las tiendas y mercadillos se han convertido en colosales centros comerciales; la desinformación y el fraude han contaminado el ágora global. Tras cumplir su promesa de demoler las viejas jerarquías, los tecnólogos erigieron nuevas torres y feudos en su lugar.

Según la ortodoxia del sector - descrita por Richard Barbrook y Andy Cameron en su influyente ensayo "La Ideología californiana"- la tecnología es la solución a cualquier problema. Los defensores de Silicon Valley consideran que la tecnología es intrínsecamente empoderadora, tan cargada de cosas buenas que el daño es impensable. Un espíritu de excepcionalismo corre por las venas de la comunidad tecnológica: los creyentes se ven a sí mismos como betatesters de un mundo feliz y consideran las estructuras sociales, normas y leyes actuales como anacronismos o inconvenientes a sortear. Los tecnólogos hemos aprendido a construir primero y preguntar después. *Lean Startup*, la ideología tecnológica predominante hoy en día, es vehementemente empírica. Considera que estamos tan inmersos en el cambio que es inútil predecir el futuro; en su lugar, debemos dar prioridad a la validación sobre la investigación y aprender construyendo. Construir, medir, aprender, repetir.

Este enfoque ha permitido innovar con audacia en sectores estancados, pero cuando la tecnología se convierte en la respuesta a cual-

quier problema, a nadie le sorprende que el "¿Podemos?" supere al "¿Debemos?". Tal y como prometieron, los tecnólogos nos hemos movido rápido y hemos roto muchas cosas[1]. Los continuos errores cometidos por la industria -algoritmos racistas, abusos contra la privacidad, vista gorda ante el acoso y el odio- han erosionado la fe pública y han llevado a los medios de comunicación a calificar la tecnología como un peligro con la misma frecuencia que como la salvadora. A las personas empleadas en el sector tecnológico puede sorprendernos encontrarnos en el punto de mira. La mayoría de nosotros tenemos buenas intenciones y realmente deseamos contribuir a mejorar la condición humana, o simplemente abordar problemas interesantes. Los problemas del sector se deben sobre todo a la negligencia, no a la malicia.

El despertar ético debería haberse producido hace tiempo. Los tecnólogos estamos empezando, con razón, a cuestionarnos nuestra influencia en un mundo que se desvía de su curso previsto y, a medida que la industria madura, es natural que prestemos atención a cuestiones más profundas de impacto y justicia. Como señala el sociólogo Richard Sennett, "los problemas éticos del oficio hacen su aparición cuando se alcanza la maestría".[2]

Este enfoque coincide con la creciente inquietud pública y el apetito por el cambio ético. Los consumidores quieren apoyar a las empresas que propugnan valores claros: El 87% de los consumidores compraría un producto porque la empresa aboga por un tema que les preocupa.[3] La tecnología emergente incrementa aún más los desafíos. En las próximas décadas, nuestro sector pedirá a los consumidores que nos confíen sus datos, sus vehículos e incluso la seguridad de sus familias. La ciencia ficción distópica ha enseñado a la gente a ser escéptica ante estas solicitudes. A menos que abordemos las cuestiones éticas que están asolando el sector, esta confianza será difícil de ganar.

Instrumentalismo, determinismo y mediación

Como primer paso ético, debemos abandonar la idea reconfortante de que la tecnología es neutra. Esta postura *instrumentalista* sostiene que la tecnología es sólo una herramienta que las personas pueden utilizar para hacer el bien o abusar de ella y causar daños. Los instrumenta-

listas sostienen que, puesto que los malhechores siempre tergiversarán la tecnología para el mal, el único recurso ético es educar y abogar por un uso adecuado. Esto desvía la responsabilidad hacia el usuario, lo que permite a los tecnólogos librarse del peso moral. Todos conocemos un estribillo instrumentalista popular: "Las armas no matan a la personas; las personas matan a las personas".[4]

El punto de vista opuesto, el *determinismo tecnológico*, sostiene que la tecnología es cualquier cosa menos neutral; es tan poderosa que moldea la sociedad y la cultura, actuando más como nuestro patrón que como nuestro sirvo. El determinismo está presente tanto en la ciencia ficción como en el mundo académico, e incluso ha empezado a seducir a los medios de comunicación: portadas de periódicos y noticieros se llenan de informes entusiastas sobre el dominio que la tecnología está ejerciendo sobre la humanidad. Los políticos también empiezan a contagiarse del determinismo, declarando que la tecnología definirá el siglo XXI.

El instrumentalismo es útil para acallar la crítica: si la tecnología es sólo una herramienta inerte, no tiene efectos sociales, políticos o morales. Sin embargo, la industria ha sido obstinada al aferrarse a este punto de vista; el marketing del sector tecnológico sugiere que la industria es muy consciente de sus posibles repercusiones. Los tecnólogos describimos a menudo nuestros elevados objetivos con un lenguaje determinista: —¡Democratizar! ¡Transformar! ¡Disrumpir!— pero después recurrimos a defensas instrumentalistas ante las cuestiones éticas: lamentamos sinceramente este incidente, pero no se nos puede responsabilizar de un uso indebido. En otras palabras, la tecnología cambiará el mundo, pero si el mundo cambia, no nos echen la culpa a nosotros.

Los efectos nocivos de la tecnología hacen insostenible el instrumentalismo; incluso el supuestamente benevolente motor de búsqueda ha reforzado los prejuicios y devaluado las fuentes de información fiables. Oponerse al mito de la neutralidad no es una postura nueva. En 1985, el historiador tecnológico Melvin Kranzberg presentó seis leyes de la tecnología. La primera: "La tecnología no es ni buena ni mala, ni neutral". El rechazo del instrumentalismo no implica necesariamente el determinismo. Situar a los tecnólogos en el centro del universo no es saludable para un sector que necesita imperiosamente humildad, y el determinismo puede minimizar la responsabilidad

ética de los tecnólogos. Si vemos la tecnología como una fuerza social imparable, podríamos llegar a la conclusión de que escapa a nuestro control.

El filósofo tecnológico Peter-Paul Verbeek propone una tercera perspectiva, la teoría de la mediación tecnológica, que combina perfectamente los puntos de vista opuestos del instrumentalismo y el determinismo.[5] Para Verbeek, la tecnología es un medio a través del cual percibimos y manipulamos nuestro mundo. Las gafas nos ayudan a ver y comprender nuestro entorno; los martillos nos ayudan a construir refugios y esculturas; las cámaras nos ayudan a recordar y compartir nuestros recuerdos. Es inútil separar la tecnología de la sociedad. Ni nosotros controlamos totalmente la tecnología, ni ella nos controla totalmente a nosotros. Una anécdota de Kranzberg sobre el violinista Fritz Kreisler muestra cómo funciona esta combinación:

> Una mujer se acercó a [Kreisler] después de un concierto y le dijo efusivamente: "Oh, maestro, su violín hace una música tan maravillosa". Kreisler tomó su violín (nada menos que un Stradivarius), se lo acercó a la oreja y dijo: "No escucho música alguna saliendo de él." Como ven, la maravillosa música que salía del violín no provenía únicamente del instrumento, el hardware; sino que dependía del elemento humano, el software.[6]

Sólo el violinista —un híbrido de violín y ser humano— podía crear una música tan memorable (aunque podríamos culpar a Kreisler de la arrogante ocurrencia).

Barreras a la ética

¿Qué significa eso en términos de ética? Si los seres humanos y la tecnología actúan en tándem, no podemos afirmar que la tecnología sea éticamente inerte, pero tampoco podemos separarla de la acción humana. La ética de la tecnología se convierte en la ética de la vida cotidiana. La ética no siempre despierta entusiasmo como tema de conversación. ¡Todos esos experimentos mentales sin sentido y tantos griegos antiguos!. Por no hablar de la pedantería de las definiciones: ¿cuál es la diferencia entre ética y moral? Tal vez recuerdes los estudios religión o ética del instituto: ¿acaso la ética no se refiere a las

expectativas de la sociedad y la moral a algo más personal e innato? Se abre una profunda madriguera filosófica, pero por fortuna podemos eludirla. La mayoría de los filósofos modernos (aunque no todos) no ven una gran diferencia entre moral y ética, y utilizan los términos indistintamente. Yo también lo haré.

Sea cual sea la etiqueta, la ética es más importante fuera de las aulas. La ética es una necesidad y una realidad, nada menos que el compromiso de tomarnos en serio nuestras decisiones e incluso nuestras vidas. Este compromiso es especialmente importante para los diseñadores. El diseño es ética aplicada. A veces esta conexión es obvia: cuando diseñas alambre de espino, estás diciendo que cualquiera que intente contravenir el derecho de otra persona a la propiedad privada debe resultar herido. Pero sea cual sea el medio o el material, todo acto de diseño es una declaración sobre el futuro. El diseño cambia la forma en que vemos el mundo y cómo podemos interactuar en él; el diseño convierte las creencias sobre cómo deberíamos vivir en objetos y entornos que la gente utilizará y habitará. Al elegir el futuro que aspiran, los diseñadores descartan docenas de realidades alternativas, que se materializan brevemente en prototipos o bocetos, y que desaparecen en la papelera de reciclaje. Una famosa cita proclama: "La ética es la estética del futuro".[7]

Plantear la ética en el lugar de trabajo suele suscitar dos objeciones. La primera consiste en afirmar que la ética no tiene cabida en la industria y que el comportamiento aceptable debe decidirlo el mercado o la ley. Esta idea es política y su punto débil debería ser evidente para cualquiera que cuestione su principio libertario. Un mercado que se autocorrige éticamente requiere una información perfecta y un acuerdo generalizado sobre lo que está bien y lo que está mal. Los clientes sólo pueden castigar las transgresiones éticas si saben y entienden lo que hacen las empresas y si están de acuerdo en que no es ético. Sin embargo, la tecnología actúa de forma invisible, a menudo bajo un consentimiento dudoso y, por lo general, utilizando un dialecto que sólo unos pocos pueden hablar. El público en general no tiene ni idea de qué tipo de actos indeseados están ocurriendo dentro de sus dispositivos. La idea de la autocorrección del mercado es una fantasía.

La afirmación de que la ley es el mejor árbitro ético es especialmente desafortunada; en esencia, sostiene que deberíamos permitir

todos los comportamientos excepto los delictivos. La ética debería consistir en vivir lo mejor posible, no en ver cuán bajo podemos caer. Y las propias leyes pueden ser moralmente erróneas; a veces personas valientes decidieron desobedecer una legislación injusta para provocar un cambio ético: que se lo pregunten a Rosa Parks, la activista afroamericana. Incluso si ignoramos estos razonamientos, para que la ley sea un sustituto apropiado de la ética en la tecnología, tendríamos que encontrar legisladores que entiendan profundamente la tecnología. La historia nos ha demostrado que, lamentablemente, estas personas escasean.

La segunda objeción común a hablar de ética en la empresa es que entorpece la innovación. A veces es cierto. Detenerse a hacer un balance moral puede apagar ideas potencialmente dañinas, pero una empresa progresista debería estar agradecida por una intervención así. La ética no es sólo un lastre para la innovación; bien gestionada, puede abonar nuevas ideas y desterrar las malas.

Este Libro

Aunque es alentador que los tecnólogos por fin nos tomemos en serio la ética, no debemos creer que somos los primeros en llegar a estas costas. Lamentablemente, los filósofos, académicos, escritores y artistas que han estudiado el tema durante décadas todavía no son tomados en serio en nuestra industria; la cultura tecnológica valora la inteligencia, pero es obstinadamente antiintelectual. A su vez, los académicos se lamentan de la arrogancia de los profesionales, que chocan una y otra vez contra los mismos viejos muros a pesar de que se les paga generosamente por ello.

Como diseñador en activo, no como experto en ética, escribo este libro para mis colegas de la industria tecnológica. Aunque el libro debe una profunda gratitud a quienes han allanado el camino y no rehúye las ideas complejas, intentaré siempre traducir la teoría en práctica. Dicho esto, un manual de ética es un oxímoron; si buscas listas de instrucciones, te decepcionarás. Nadie puede responder a los problemas éticos por nosotros; tenemos que reflexionar sobre ellos, y generalmente hay más de una respuesta. Parafraseando a Caroline Whitbeck[8], los problemas éticos son como los proyectos: a menudo hay docenas de soluciones viables, cada una con sus propias contra-

partidas. Sin embargo, esto no significa que no haya respuestas equivocadas. La ética está plagada de trampas y falacias; a continuación destacaremos las más comunes.

Algo de política es inevitable en un libro como éste, ya que la ética y la política están naturalmente entrelazadas. La amplitud de la opinión humana se refleja en la complejidad de la ética; las opiniones morales de las personas tienden a informar sus opiniones políticas, y viceversa. Las personas de izquierdas tienden a inclinarse por posturas morales que dan prioridad al bien social, mientras que las de derechas tienden a preferir perspectivas que apoyen la soberanía y la autonomía individuales. La experiencia personal también deja una fuerte impronta: la víctima de robo probablemente será más sensible al robo en el futuro, conscientemente o no. Sería poco sincero por mi parte disimular mis inclinaciones personales y políticas en aras de una falsa objetividad, aún así prometo evitar hacer propaganda política y, en su lugar, ofrecerte las herramientas necesarias para que tu puedas analizar los argumentos éticos. Puede que incluso descubras que pensar profundamente sobre la ética influye en tus puntos de vista sobre la sociedad en general.

Gracias por interesarte en forjar una industria tecnológica mejor; espero que este libro te aporte tanto la teoría como los conocimientos prácticos que necesitas para conseguirlo. Empecemos.

CAPÍTULO 2
¿NO CAUSAR DAÑOS?

Como gobernantes coloniales de la India, los británicos empezaron a preocuparse por la abundancia de cobras en Delhi. Los gobernadores propusieron un remedio simple y económico: una recompensa por la piel de las cobras. La medida fue un éxito; tanto, que los emprendedores Indios empezaron a criar cobras sólo por la recompensa. Al ver un aumento sospechoso de recompensas pagadas, los británicos cancelaron el plan. Los criadores, en lugar de mantener cobras sin valor, prefirieron liberarlas, causando que la población de cobras salvajes se disparase a niveles superiores a los iniciales, y anulara el objetivo del programa.[1]

Consecuencias no intencionales y externalidades

Incluso los actos más benignos y bienintencionados pueden tener repercusiones inesperadas. El "efecto cobra" no le sorprendería al teórico cultural francés Paul Virilio.

> Inventar el barco es inventar el naufragio; inventar el avión es inventar el accidente aéreo; inventar la electricidad es inventar la electrocución… Cada tecnología lleva consigo su propia negatividad que aparece al mismo tiempo que el progreso técnico[2]

Para Virilio, cada yin de la tecnología tiene su correspondiente

yang, un abanico de consecuencias no intencionadas que nacen cuando la tecnología falla, tiene éxito más allá de lo esperado o simplemente se utiliza de forma inesperada. El filósofo Don Ihde sostiene que las tecnologías no tienen identidades o significados fijos, sino que son *multiestables*: la gente da a la tecnología todo tipo de usos que van más allá de los previstos por su diseñador.[3] El GPS se concibió originalmente para el ejército, pero desde que se puso a disposición del público civil ha generado miles de productos y servicios, cada uno con sus propias consecuencias. Los navegadores por satélite han acabado con los atlas de carreteras y han congestionado las carreteras de los pueblos que, imprudentemente, se ofrecen como atajos. Los programas de seguimiento de personas han simultáneamente mejorado y erosionado la confianza personal, por un lado salvando a menores perdidos y por otro arruinando matrimonios y fiestas sorpresa. Según la *ley de las consecuencias no intencionadas*, siempre habrá resultados que pasemos por alto, pero "no intencionado" no significa imprevisible. Podemos —y debemos— intentar anticipar y mitigar las peores consecuencias potenciales.

Una de las consecuencias no intencionadas es la *externalidad*. Una externalidad es la denominación que dan los economistas al "Problema Ajeno", un efecto que recae sobre alguien ajeno al sistema. Los fumadores pasivos no eligen fumar, sino que son víctimas de una externalidad negativa, perjudicados por el hábito de otra persona. Las externalidades también pueden ser positivas: una de las ventajas del transporte público es que se reducen las muertes de peatones por conductores ebrios.

Las consecuencias no intencionadas afectan a personas conocidas de forma imprevista, mientras que las externalidades afectan a personas que hemos ignorado. En otras palabras, pasamos por alto las consecuencias no intencionadas al no mirar con suficiente profundidad, pero pasamos por alto las externalidades porque miramos en los lugares equivocados.

Las externalidades han sido un problema recurrente a lo largo de la historia de la industria. Un enfoque egoísta y cortoplacista ha tentado a muchas empresas a perjudicar su ecología y su futuro. Hay pruebas, por ejemplo, de que Exxon conocía la amenaza potencial del CO2 para el clima en 1977, pero lo mantuvo en secreto, prefiriendo que la sociedad pagara el coste.[4] Las externalidades también pueden

darse como efecto secundario del diseño centrado en el usuario. Centrarse en cumplir los objetivos y sueños de un usuario individual ha hecho que las empresas tecnológicas pasen por alto los impactos sobre los no usuarios y la sociedad en general.[5] Airbnb es un sueño para anfitriones y huéspedes, pero genera externalidades negativas para el vecindario:

> Al menos a corto plazo, [Airbnb] reduce el stock disponible para el alquiler a largo plazo o la compra [...] Aun así, sigue existiendo una segunda externalidad: el impacto sobre los vecinos. Vivir al lado de un residente permanente es muy diferente a vivir al lado de visitantes que cambian constantemente y que tienen motivos para invertir tiempo en relaciones, en el vecindario o incluso en facilitar una noche de sueño reparador. Dicho de otro modo, no es de extrañar que a los anfitriones y a los huéspedes les encante Airbnb: todo el coste se traslada a la gente que no ve ni un céntimo. —Ben Thompson[6]

La mejor forma de eliminar las externalidades es, por supuesto, internalizarlas. Los economistas, como es su costumbre, suelen sugerir que lo hagamos con impuestos o sanciones. Muchos gobiernos responden a las externalidades medioambientales con el *principio de quien contamina paga*, cargando el coste sobre la parte responsable y anulando la externalidad. Otra alternativa es subvencionar las externalidades positivas, como la financiación de programas de desplazamiento al trabajo en bicicleta que también mejoran la forma física de los ciudadanos. Si Airbnb decidiera dar prioridad al bienestar del vecindario —ya sea por la presión de los consumidores, por la amenaza de multas o como resultado de una inyección de conciencia social— la externalidad desaparecería. La comunidad se convertiría en el problema de Airbnb y no tardarían en aparecer políticas favorables al bienestar del vecindario.

Resolver las externalidades significa que primero tenemos que reconocerlas, pero a menudo permanecen en la sombra, recaen sobre minorías ignoradas o sólo existen en un futuro impreciso.

> Si alguien roba en una tienda, es un delito y el Estado dispone de todo lo necesario para detener al delincuente. Pero cuando alguien roba bienes comunes y del futuro, se considera una actividad empresarial y

> el Estado se alegra y le concede ventajas fiscales en lugar de detenerlo. Necesitamos urgentemente un concepto más amplio de justicia y equidad que contemple una posible hipoteca del futuro. —Ursula Franklin[7]

Sesgo Algorítmico

El *sesgo algorítmico* —aquellos algoritmos supuestamente imparciales que codifican prejuicios implícitos— es un ejemplo clásico de consecuencias no intencionadas. El sesgo se ha convertido en uno de los problemas éticos más conocidos de la tecnología, como demuestran varios ejemplos lamentables: el software prediccion policial que considera que las personas de raza negra corren un mayor riesgo de reincidencia que las de raza blanca; el sistema de recomendación de YouTube que arrastra continuamente a las personas hacia contenidos extremos; las redes que muestran anuncios de trabajo altamente remunerados a los hombres, pero no a las mujeres.

Los algoritmos sesgados son claramente más peligrosos cuando gobiernan sistemas críticos como la justicia o el empleo, pero hasta un algoritmo comercial sesgado puede tener efectos nefastos. Tanto individuos como grupos pueden ser víctimas de la *redlining*, es decir, la denegación de productos y servicios por un software sesgado. La etiqueta *redlining* tiene su origen en la banca de la década de 1930, cuando los prestamistas identificaban en el mapa de la ciudad los barrios (en su mayoría habitados por negros) a los que no concedían préstamos. En la actualidad, el *redlining* es menos calculado, pero puede ser igual de perjudicial. Bloomberg descubrió que el servicio de entrega en el mismo día de Amazon ignoraba los barrios con una mayoría de población negra, como Roxbury, en Boston, a pesar de que todos los suburbios circundantes cumplían los requisitos.[8] Hasta los prejuicios aparentemente menores se acumulan, intensificando diferencias. Al ser denegados la entrega rápida, los residentes de Roxbury se ven obligados a sacrificar tiempo y dinero comprando en establecimientos más caros: otro ladrillo en el muro de la desigualdad.

Ninguna de estas consecuencias estaba prevista, más bien están fuera del alcance de lo que los tecnólogos consideramos. Nadie se fijó en las posibles repercusiones para los usuarios y a nadie se le ocurrió alzar la voz en defensa de las posibles víctimas. Este sigue siendo un

defecto de la industria. En general, el público no puede defenderse de los sesgos algorítmicos ni tampoco hacer reclamaciones. Sin humanos en el bucle, la decisión queda en manos de dioses algorítmicos omniscientes e incuestionables. Si tu suerte algorítmica está echada, no hay mucho que puedas hacer, salvo rezar.

Fuentes de Sesgo

Joanna Bryson, especialista en ética de la IA, afirma que el sesgo algorítmico tiene tres causas principales.[9] La primera, datos de entrenamiento pobres. Los datos incompletos, poco representativos o mal depurados siempre generan puntos ciegos algorítmicos. Un sistema de reconocimiento facial entrenado sólo con rostros blancos garantiza el sesgo racista. Y esto no es sólo inconveniente, sino que además resulta denigrante: no reconocer un rostro es no reconocer la humanidad de alguien.

El sesgo causado por los datos incompletos suele perjudicar especialmente a los más desfavorecidos. Los ricos, que disponen de acceso a la tecnología e historiales financieros detallados, proyectan huella digital amplia; los pobres o los marginados, por lo general, no. Aunque un perfil de datos extenso puede ser a veces un riesgo para los individuos, la "pobreza de datos" sistémica causa progresivamente daños a comunidades enteras. Las subpoblaciones se vuelven algorítmicamente invisibles y, por tanto, reciben un trato injusto; la opresión se reproduce y amplifica digitalmente.

La segunda fuente de sesgo algorítmico que menciona Bryson son los prejuicios intencionados. Los algoritmos ofrecen una forma atractiva de ocultar prejuicios bajo la ilusión de objetividad, y muchos intolerantes en el seno de nuestras propias empresas y de los gobiernos tienen el poder de manipular los algoritmos en favor de sus preferencias. Los prejuicios intencionados son siempre contrarios a la ética, pero a menudo son legales. Los distintos países y estados difieren mucho en cuanto a actitudes y leyes. En la actualidad, en Kansas es legal despedir a alguien por ser gay, pero no en el estado de Colorado. Los prejuicios también pueden venir de fuera del equipo. Tay, el famoso chatbot de Microsoft, se programó para aprender a través de la interacción en X (previamente Twitter), lo que lo hizo vulnerable a la manipulación. Los trolls aprovecharon la oportunidad para incitar a

Tay a hacer comentarios ofensivos y, cuando se corrió la voz de los primeros resultados, el abuso se disparó rápidamente.

Esto apunta a la tercera fuente de sesgo, la más fundamental: incluso el conjunto de datos más completo está impregnado de prejuicios humanos. La teoría de la mediación de Verbeek nos dice que no debemos separar la acción tecnológica de la humana, por lo que las tecnologías reflejarán por defecto los prejuicios sociales. Estos prejuicios son profundos. Bryson y dos colegas entrenaron un sistema básico de *machine learning* en un corpus estándar de texto y descubrieron "todos los prejuicios lingüísticos documentados en psicología que [habían] buscado".[10] Según Bryson, las combinaciones de palabras —esencialmente, los mapeos matemáticos del lenguaje— "parecen saber que los insectos son molestos y las flores hermosas" simplemente porque esos tipos de palabras se emparejan con frecuencia. No es de extrañar, por tanto, que los algoritmos de análisis de sentimientos hayan heredado prejuicios, considerando los nombres europeos (Paul, Ellen) más agradables que los afroamericanos (Malik, Sheeren)[11] y clasificando la palabra "gay" como negativa.[12] Incluso cuando la opinión pública cambie, este sesgo solo se disipará lentamente. Los datos siempre miran hacia atrás, lo que significa que los prejuicios históricos están congelados en el corpus de entrenamiento.

Por supuesto, la desigualdad va más allá del lenguaje: casi cualquier dato puede estar impregnado de prejuicios implícitos. Aunque sea ilegal tener en cuenta la raza de una persona a la hora de calcular su puntuación crediticia, todos los agentes de crédito se fijan en la dirección del solicitante, que está fuertemente correlacionada con la pertenencia racial. Los detractores del algoritmo COMPAS de predicción de la delincuencia afirmaron que otorgaba un peso injusto a las detenciones y condenas previas. Como señaló la catedrática de Derecho Ifeoma Ajunwa, "si miramos el número de condenas de una persona y considerándolo como una variable neutra —bueno, no lo es". [13] Todos sabemos que algunas poblaciones están sometidas a una vigilancia policial excesiva y que el origen étnico influye en la imposición de penas; todos estos factores se filtran en la lógica algorítmica. Incluso decisiones aparentemente inocuas como dónde construir un centro de distribución —presumiblemente la raíz del sesgo de entrega de Amazon— dependen del mercado local, las opciones de transporte, el valor del suelo y otros factores cargados de sesgos implícitos. La

discriminación ya está integrada en el tejido de nuestras herramientas y conjuntos de datos.

Distribución moral

Si el sesgo no es intencionado, ¿por qué es nuestro problema? Seguramente no es tarea de la tecnología arreglar todos los defectos de la psique humana, ¿no? Una vez más, recordemos la teoría de la mediación y su elegante danza de personas y tecnología. En el lugar de un accidente, los investigadores se preguntarán si la bicicleta dio un volantazo o si el conductor estaba manipulando el sistema de entretenimiento, pero también comprobarán los frenos del coche y preguntarán quién hizo la última revisión. Las tecnologías tienden a repartir la responsabilidad moral entre muchas partes.

De momento no culpamos a la tecnología en sí, pero si la tecnología cambia la forma en que los usuarios interpretan el mundo, entonces se puede deducir que los tecnólogos influimos en las decisiones morales de las personas. Cuando las cosas van mal, tanto el usuario como el tecnólogo puede que tengan la culpa. La tecnología es ahora un deporte de equipo: la era de los hackers solitarios quedó atrás. La tecnología moderna está hecha por equipos de ingenieros, diseñadores, gestores de productos y científicos de datos, y se sustenta en múltiples capas subyacentes: aplicaciones de inteligencia artificial en plataformas que utilizan bibliotecas compartidas, conectadas a diversos sistemas operativos y protocolos. Aunque cada capa ha sido creada por equipos u organizaciones diferentes, todos podrían estar moralmente implicados. Un equipo es tan digno de confianza como su miembro más corrupto. Si construyes sobre una plataforma vulnerable, te vinculas a una red social de dudosa reputación o compartes datos con un anunciante abusivo, tendrás parte de culpa, y con justa razón.

Eso no quiere decir que debamos culpar a los tecnólogos de todas las consecuencias no intencionadas: dada la naturaleza multiestable de la tecnología, siempre habrá algunos resultados que no podrían haberse previsto. Parece injusto, por ejemplo, culpar a los arquitectos del GPS de los atascos rurales. Sin embargo, el problema de los prejuicios ya era conocido. En la década de los 1980 ya existía un problema documentado, cuando los médicos de la Facultad de Medicina del

Hospital St George descubrieron que su algoritmo de admisión, basado en las decisiones de la década anterior, era discriminatorio:

> El ordenador utilizaba información [implícita] para generar una puntuación que se utilizaba para decidir qué solicitantes debían ser entrevistados. Las mujeres y los miembros de minorías raciales tenían menos posibilidades de ser entrevistados, independientemente de consideraciones académicas. —Stella Lowry y Gordon Macpherson[14]

Puede que el sesgo algorítmico no sea intencionado, pero es negligente. Es posible que los tecnólogos no lo viéramos venir, pero no nos preocupamos de investigar. Convencidos de que la tecnología es neutral y objetiva, asumimos por error que el sesgo era imposible; y nos preocupamos únicamente por la productividad del usuario principal, pasando por alto el impacto en la sociedad en general.

Al final, la culpa quizás no importe. La responsabilidad causal y la responsabilidad moral no siempre coinciden; quizás sea más útil preguntarse quién tiene el poder de arreglar las cosas. Aunque los tecnólogos no hayamos provocado directamente el desliz ético, seguimos teniendo el deber de intentar resolverlo. Independientemente de la intención, debemos intentar reducir los sesgos dañinos por el bien de la sociedad y, en segundo lugar, por nuestra propia reputación.

Relativismo moral

Aquí nos topamos con un problema ético clásico: hacer lo correcto suena atractivo, pero ¿qué es lo correcto? ¿Qué hace que un algoritmo sea justo? La cuestión se transforma rápidamente en política. ¿Debemos buscar la igualdad de trato o la igualdad de resultados? Si tratamos a todo el mundo igual, no hacemos nada para resolver los problemas sistémicos que perpetúan la desigualdad. Pero si, por el contrario, buscamos resultados más equitativos, se nos acusa de injerencia y discriminación inversa. ¿Deben los algoritmos reflejar simplemente la sociedad actual o ayudarnos a conseguir un mundo más justo? ¿Quién elige? Pensándolo bien, ¿es el punto de vista moral de alguien más válido que el de los demás?

Este es el territorio seductor del *relativismo moral*. Un relativista

sostiene que no existe una única verdad moral, ni una estrella polar que guíe el comportamiento, sino que las normas éticas dependen de las diferencias sociales y varían de una cultura a otra.

El relativismo suele emanar de los bienintencionados principios de tolerancia y diversidad. Responsabilizar a todo el mundo de los limitados valores de una cultura dominante —en otras palabras, atribuir a un único partido, país o religión la custodia de la verdad moral— ha resultado letal históricamente, y los relativistas señalan que las creencias individuales de las personas están moldeadas por la educación y evolucionan con la experiencia. Pero los conservadores suelen denunciar el relativismo moral como una debilidad posmoderna llevada a extremos peligrosos. El filósofo tradicionalista Roger Scruton afirma: "Un escritor que dice que no hay verdades, o que toda verdad es "meramente relativa", te está pidiendo que no le creas. Así que no lo hagas".

La globalización es particularmente desafiante para el relativismo. ¿Debemos aceptar las opciones de otra sociedad que nos repugnan? ¿Debemos hacer negocios con países que discriminan activamente o en los que la corrupción está muy extendida? El relativismo moral sugiere un todo-vale: ¿quiénes somos nosotros para discutir las normas de otra cultura?

El debate filosófico no tiene fin, pero a efectos prácticos el relativismo es un callejón sin salida. Si la gente puede escabullirse del juicio moral alegando que sus acciones son culturalmente aceptables, la propia moralidad se convierte en un concepto cuestionable. *Ad absurdum*, si la bondad está en el ojo del que mira, los dueños de esclavos podrían decidir si la esclavitud es ética. Para progresar moralmente, debemos ser capaces de trazar una línea divisoria entre lo aceptable y lo inaceptable. Afortunadamente, la mayoría de las culturas se muestran de acuerdo sobre lo que está bien y lo que está mal, como el asesinato y el adulterio. Cuarenta y ocho naciones encontraron suficientes puntos en común para codificar los principios morales básicos en la Declaración Universal de los Derechos Humanos.

Si rechazamos el relativismo, tenemos que rechazarlo también en el trabajo. Los magnates de mano dura pueden afirmar que no hay lugar para la moralidad personal en el negocio: las buenas personas llegan en último lugar. Pero si las distintas culturas no pueden dictar

sus propias reglas éticas, tampoco lo lo puede hacer el mundo de los negocios. Es cierto que todos desempeñamos muchos papeles en la vida y adaptamos nuestros comportamientos en consecuencia —una entrada contundente es aceptable en un campo de rugby, pero no en una boda—, pero estos papeles siguen sustentándose en una base moral común, que no puede sustituirse ni apagarse a voluntad. La moralidad no se detiene en la entrada de la oficina; al fin y al cabo, las empresas están formadas por personas.

La trampa de la tecnocracia

Si queremos progresar moralmente, alguien tiene que definir cuáles deben ser los estándares éticos. Idealmente, esa es una cuestión que compete a la propia sociedad. Los cargos electos hacen las leyes; los ciudadanos desarrollan poco a poco las convenciones sociales. Pero las tecnologías disruptivas a menudo irrumpen en escena sin previo aviso, antes de que puedan surgir esas normas sociales o legales. Por defecto, las nuevas tecnologías llevan las huellas éticas de sus creadores, no de la sociedad en general. Dado el peso de la tecnología en la cultura moderna, los tecnólogos tenemos una gran influencia sobre las normas sociales: las decisiones éticas que deberían ser democráticas son, en cambio, tecnocráticas.

Esto debería preocuparnos. Ningún grupo debería tener más derecho sobre el futuro que la sociedad, y los tecnólogos no tenemos la diversidad ni la sabiduría necesaria para ser autoridades éticas naturales. Los gobiernos están actuando con lentitud y los ciudadanos cada vez prestan más atención a su realidad tecnológica, pero la industria también tiene que hacerse más responsable. Debemos demostrar que merecemos la confianza de la sociedad implicando al público en las decisiones morales que rodean a la tecnología y dando prioridad al bien de todos, no sólo a nuestras fuentes de ingresos.

Definiendo la equidad

¿Qué tipo de líneas morales podemos trazar sobre el sesgo algorítmico? En primer lugar, debemos ser precisos. La palabra "sesgo" es útil hasta cierto punto: se entiende en sentido amplio y es sumamente sencilla, pero enseguida necesitamos más detalles. Sesgo es un

término genérico que engloba varios tipos de desequilibrio: ¿se trata de un sesgo de muestreo, de un sesgo estructural innato o de un prejuicio explícito? Para ser específicos, también debemos ser osados y descartar algunas definiciones perfectamente viables.

Imagina que trabajas para Tinder. Si quieres asegurarte de que tus algoritmos de emparejamiento sean justos racialmente, ¿qué significaría realmente la equidad? Tal vez la equidad sea cuestión del grado de exposición, lo que significa que deberíamos mostrar a los usuarios emparejamientos potenciales que reflejen la mezcla racial local. Si el 30% de las personas de la ciudad del usuario son negras, podríamos ajustar el algoritmo para que el 30% de las posibles parejas del usuario también lo sean. Esto parece razonable, pero tiene un defecto desagradable. El mundo de las citas, el sexo y el amor está plagado de prejuicios humanos, mucha gente tiende a ceñirse a su propia raza a la hora de elegir pareja. Así que, a menos que los usuarios de Tinder no tengan prejuicios, algunas personas serán mostradas con frecuencia a usuarios que no quieran salir con alguien de esa raza. Puede que consigamos una exposición justa, pero ciertas razas serán emparejadas con menos frecuencia. Una forma de justicia hace aflorar otra forma de injusticia.

¿Deberíamos aspirar a un emparejamiento justo? ¿Sería mejor declarar que, sea cual sea tu raza, debes tener las mismas posibilidades de encontrar pareja a través de Tinder? Existe el problema inverso. En las plataformas de citas actuales, los hombres blancos heterosexuales valoran menos a una mujer si es negra.[15] Para dar prioridad a un emparejamiento justo, el algoritmo debería restringir la mezcla racial, por ejemplo, mostrando sólo mujeres negras a hombres negros, que tienen más probabilidades de responder positivamente. Sin embargo, esto es seguramente lo contrario de la equidad racial.

No hay manera de solucionar la cuadratura de este círculo. Los prejuicios humanos que rodean al tema citas no permiten conciliar una exposición justa con un emparejamiento justo. Quizás tengamos que buscar en otra parte. Tal vez un Tinder justo sea aquel en el que la gente se sienta igual de apreciada o tenga niveles similares de satisfacción con el servicio. Esto sugiere que deberíamos preocuparnos más por los patrones de uso de las aplicaciones y las puntuaciones de satisfacción que por las crudas métricas primarias de emparejamiento y conexión.

Cualquier definición de justicia será injusta desde una perspectiva diferente. Para algunos, un algoritmo justo es aquel que refleja la sociedad actual. Para otros, un algoritmo justo debe ser un agente de cambio social. Algún tipo de sesgo es lógico, político y matemáticamente inevitable; sin embargo, alguien tiene que tomar la iniciativa. Nuestras decisiones son las estrellas por las que navegarán nuestros algoritmos. Debemos elegir con inteligencia, teniendo en cuenta las posibles consecuencias y externalidades de nuestras elecciones.

Atenuando el sesgo

Kate Crawford, profesora de la Universidad de Nueva York (NYU, por sus siglas en inglés) y cofundadora del AI Now Research Institute, es una destacada experta en sesgo algorítmico. Crawford sugiere a los equipos que inviertan en el *análisis forense de la imparcialidad* (*fairness forensics*), para mitigar el sesgo. El primer paso forense es sencillo: probar los algoritmos con un amplio conjunto de personas y datos de referencia sólidos para detectar los problemas antes de que se produzcan. Sin embargo, algunos puntos de referencia son en sí mismos imperfectos: las bases de datos de rostros de código abierto más comunes han tenido históricamente un sesgo blanco y masculino. Así que los equipos también deberían examinar sus datos de entrenamiento y pruebas para detectar posibles sesgos. El software Facets de Google, por ejemplo, ayuda a examinar conjuntos de datos en busca de lagunas o sesgos inesperados.

Si estas pruebas detectan un sesgo, la estrategia más sencilla es mejorar los datos de entrenamiento. Un algoritmo de *machine learning* entrenado con datos parciales siempre tendrá problemas, pero no basta con arrojar más puntos de datos al problema. Si 500.000 puntos de datos de entrenamiento contienen un sesgo implícito, es probable que 5.000.000 de puntos de datos también lo contengan; es necesaria una intervención más activa.

La startup Gfycat descubrió que su software de reconocimiento facial identificaba erróneamente a las personas asiáticas. El equipo recurrió a lo que viene a ser un "modo asiático", un código adicional que se activa si el sistema cree que el sujeto tiene rasgos faciales asiáticos. Aunque la mejora de la precisión probablemente justifique el hackeo de Gfycat, este tipo de solución no es escalable —el algoritmo

se avisa a sí mismo de que está a punto de ser racista, como si fuese un "Clippy", el antiguo asistente de Microsoft, espabilado. Resulta agotador e ineficaz solucionar cada discriminación una a una porque ante cada nueva discriminación que se descubre, otra aparece en otro lado. Intentar meter a la gente en categorías ("¿Es esta persona asiática? ¿Es esta persona mujer?) para activar la bifurcación del código también resulta un tanto desagradable. La manera que la sociedad clasifica la raza, y cada vez más, el género, es como un espectro en lugar de categorías separadas, nuestros algoritmos también deberían considerarlas así.

Una forma más contundente de desvirtuar los algoritmos es anularlos explícitamente, eliminando del sistema datos sesgados o asociaciones ofensivas. Google Fotos corrigió su famoso error del "gorila" —cuando el software clasificó como tal a un grupo de amigos de raza negra— anulando el algoritmo: el equipo simplemente eliminó la palabra de la lista de posibles categorías. En este caso, valió la pena pagar el precio de una menor capacidad de clasificación. Google también ha añadido listas de exclusión a las búsquedas, después de que los investigadores descubrieran que algunos términos raciales generaban sugerencias de autocompletar atroces.[16] Este tipo de interferencia directa no debería realizarse a la ligera. Sólo resuelve un caso visible, y la idea de obligar a un algoritmo a seguir la pauta deseada provocará acusaciones de que los tecnólogos están imponiendo su política personal al mundo. Es mejor reservar los vetos para situaciones en las que el resultado es tan claramente perjudicial que exige una solución inmediata.

Dado que el sesgo nunca puede eliminarse por completo, en algún momento nos enfrentamos a otra decisión difícil: ¿es el algoritmo lo suficientemente justo para ser utilizado? ¿Es éticamente aceptable publicar a sabiendas un algoritmo sesgado? La respuesta dependerá en parte del marco ético que se prefiera; en breve lo discutiremos. A veces será preferible una decisión humana a un algoritmo sesgado: cuanto más graves sean las implicaciones del sesgo, más justificada estará la participación humana. Pero no debemos asumir que los humanos serán siempre más justos. Al igual que los algoritmos, los humanos son producto de sus culturas y entornos, y podemos ser alarmantemente parciales. Las Juntas de Libertad Condicional, por ejemplo, tienden a liberar a los condenados si los jueces acaban de

comer.[17] Después de hacer todo lo posible por garantizar la imparcialidad, podemos considerar que el sesgo persistente es lo suficientemente pequeño como para tolerarlo. Podemos incluir advertencias o controles de interfaz en los sistemas para ayudar a los usuarios a manejar el sesgo, por ejemplo añadiendo un control de pronombres ("él/ella/elles") al software de traducción, permitiendo al usuario anular el sesgo cuando traduce desde idiomas sin género.

Aunque Crawford ensalza estos enfoques forenses, también señala su debilidad común: sólo son soluciones técnicas. Para abordar realmente el sesgo implícito debemos considerarlo un problema humano, además de técnico. Esto significa poner el sesgo a la vista de todos. Algunos académicos optan por enumerar explícitamente sus posibles sesgos —un proceso conocido como *bracketing*— antes de iniciar una investigación, y toman nota siempre que perciben que un sesgo podría estar influyendo en su trabajo. Al exponer estos sesgos, ya sean fruto de la experiencia personal, de hallazgos anteriores o de teorías personales, los investigadores esperan abordar su trabajo con la mente más despejada y evitar sacar conclusiones erróneas. En tecnología, podríamos apropiarnos de esta idea, haciendo una lista de las formas en que nuestros algoritmos y datos podrían mostrar sesgos, y revisando después el rendimiento del algoritmo con respecto a esa lista de comprobación.

Imaginación moral

También podemos eliminar los sesgos de raíz detectando y abordando mejor las consecuencias imprevistas y las externalidades. Para ello necesitamos *imaginación moral*: la capacidad de soñar y evaluar moralmente una serie de escenarios futuros. Los seres humanos aprendemos a utilizar la imaginación moral a lo largo de nuestra vida —de hecho, somos la única especie que puede hacerlo—, pero no siempre es fácil imaginar las repercusiones reales de la tecnología. Los sistemas que diseñamos y desarrollamos se utilizan de forma asíncrona en todo el mundo; nunca podemos ver directamente la alegría o el dolor que causamos a los demás. Afortunadamente, la imaginación moral puede entrenarse. La moralidad no es una bendición genética; es un músculo que necesita ejercicio.

Podemos poner en marcha la imaginación moral con algunas

preguntas sencillas. La pregunta "¿Qué pasaría si esta tecnología tuviera una popularidad desmesurada?" ha dado lugar a miles de historias de ciencia ficción: Daniel Mallory Ortberg sugirió que la serie de televisión Black Mirror es una respuesta a la pregunta: "*What if phones, but too much?*" (que podría traducirse por "¿Qué pasaría si usáramos el teléfono, pero demasiado?"). Otra alternativa es dejarnos llevar por el pesimismo: ¿Cómo podría esta tecnología fracasar estrepitosamente? o ¿Cómo alguien podría abusar de esta tecnología? (En el capítulo 6 hablaremos de los peligros de la tecnología cuando se utiliza para hacer daño intencionadamente).

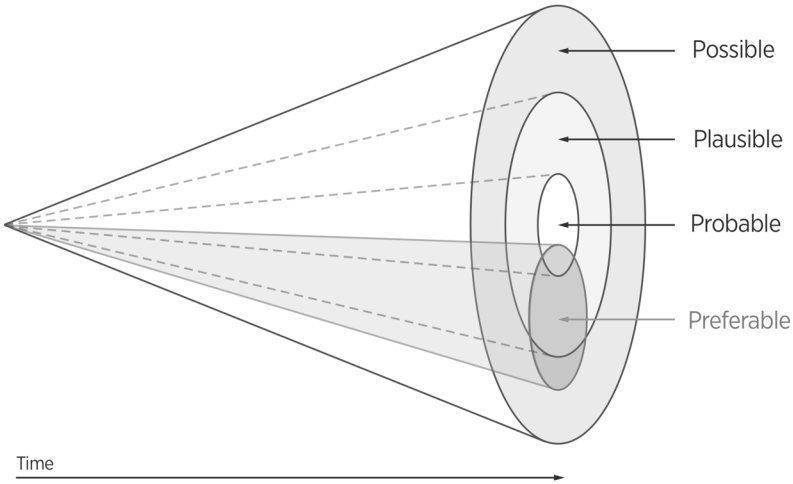
También podemos utilizar la analogía para fomentar la imaginación moral. ¿Ha ocurrido antes esta situación, o en otro ámbito? ¿Qué ocurrió después? ¿Podría ocurrir en esta situación? La historiadora económica Carlota Pérez sostiene que toda revolución tecnológica sigue un guión estricto.[18] Primero, una fase de "irrupción", en la que una tecnología prometedora y amenazadora a la vez atrae grandes inversiones especulativas. A continuación llega el "frenesí", un periodo de intensa exploración en el que los nuevos mercados cobran vida y las empresas suelen recortar los límites éticos para ganar dinero rápidamente. Con el tiempo, la burbuja estalla y los grandes fracasos obligan a los reguladores a intervenir. El impulso pasa de las finanzas a la producción a medida que la tecnología se generaliza; es una fase de "sinergia". Finalmente, la revolución se completa y domina la "madurez". El mercado está saturado, a la espera de la próxima disrupción. El *hype cycle* (ciclo de popularidad) de Gartner traza una ruta similar para las tecnologías emergentes, desde la innovación hasta la estabilidad final, pasando por las expectativas desorbitadas y el abismo de la desilusión.

Para ejercitar la imaginación moral, podemos simplemente seguir la trayectoria de nuestra tecnología preferida, imaginando cómo sería la vida en cada momento. Hoy, por ejemplo, las criptomonedas y el *machine learning* están en la fase de frenesí o expectativas infladas. No hace falta tener una gran imaginación moral para imaginar lo que podría ocurrir cuando estas expectativas infladas estallen.

Futuring

Los pronósticos pueden ser útiles, pero el mundo no siempre se ajusta a planos precisos. Para ampliar nuestra imaginación moral, podemos aprender del campo de los estudios de futuros. El principio central del llamado "diseño del futuro" (*futuring*) es ver el futuro en plural. En palabras de Sarah Connor, famosa futuróloga y especialista en ética robótica: "El futuro no está escrito. No hay más destino que el que hacemos nosotros mismos".[19] El futuro no es una marca en un mapa, sino el mapa mismo. Juntos decidimos a qué coordenadas nos dirigimos.

Un modelo habitual en los estudios de futuros es el *futures cone,*[20] que utiliza la analogía de la luz emitida por una antorcha.



El *eje x* representa el tiempo, por lo que la luz de la antorcha representa futuros potenciales. Observa que el haz de luz diverge del presente: la próxima semana es previsible; el próximo siglo, más bien poco. Cada cono de luz representa un nivel diferente de probabilidad, a veces conocido como "las 4 P".

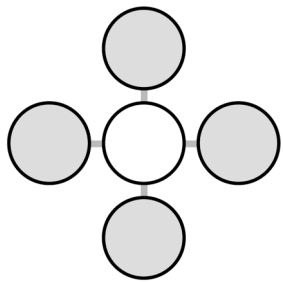
En el centro brillante del haz están los *futuros probables* (*probable futures*), las proyecciones más probables basadas en lo que sabemos hoy. El *hype cycle* de Gartner se sitúa aquí, al igual que la mayor parte del trabajo de diseño cotidiano. Moviéndonos hacia fuera, llegamos a

los *futuros plausibles*. Estos escenarios son menos probables, pero previsibles: algunas empresas pagan millones de dólares por conocerlos. En los bordes exteriores del haz se encuentran los *futuros posibles*. Situados en la penumbra, son más difíciles de detectar. Las empresas no suelen estar interesadas en estos escenarios; en cambio, los futuros posibles son el dominio de lo que Anthony Dunne y Fiona Raby llaman "cultura especulativa": ciencia ficción, arte y juegos.

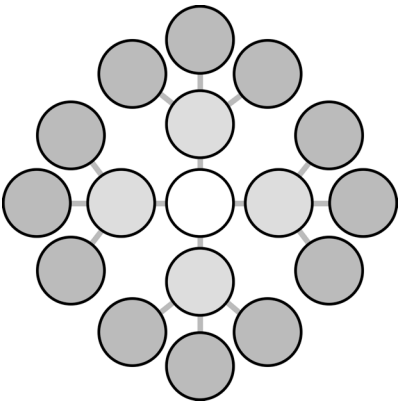
El último cono, y el más importante, representa el *futuro preferible*. En una multitud de futuros posibles, algunos serán mejores que otros. El futuro preferible es un juicio de valor; tenemos que considerar el mundo que queremos y cómo podríamos llegar a él. Este futuro ideal podría ser muy probable, si se sitúa justo en el centro del haz, o un comodín improbable si se sitúa justo en los bordes.[21]

Como cualquier modelo, el cono de futuros tiene defectos. El académico y crítico del diseño Cameron Tonkinwise apunta que la dirección del rayo depende de quién sostenga la antorcha: en un mundo desigual, cada cual parte de un "ahora" diferente. La idea de un futuro preferible también está sesgada: ¿preferible para quién? No obstante, el cono de futuros puede ser útil tanto para el trabajo estratégico como para el fomento de la imaginación moral, iluminando varias trayectorias futuras ayudan a los equipos a elegir el futuro preferido por el que trabajar.

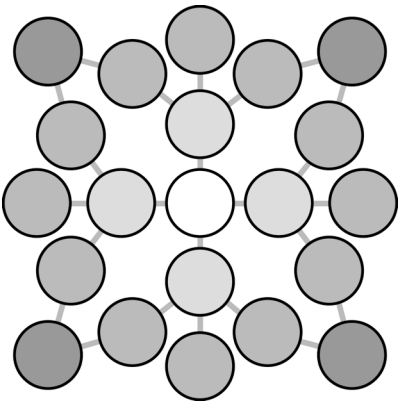
Otra herramienta para descubrir futuros potenciales y consecuencias no intencionadas es la rueda de los futuros de Jerome C. Glenn. En palabras de Glenn, la rueda de los futuros ofrece "una forma de brainstorming estructurado" del futuro. Empezamos con una tendencia de fondo, como la tecnología que hemos elegido; a continuación, en un anillo alrededor de la tendencia, escribimos algunas de sus posibles consecuencias.



Este primer paso tiende a extraer los futuros probables, que suelen ser interesantes pero raramente novedosos. A continuación, profundizaremos más, imaginando las posibles consecuencias de segundo orden de cada nuevo escenario y añadiéndolas en un segundo anillo.



Por último, emparejamos nodos interesantes de cualquier parte del diagrama e imaginamos qué podría ocurrir si ambos futuros se hicieran realidad, registrándolo en un tercer anillo. A medida que el horizonte se amplía, las cadenas de causalidad pueden tensarse, y las posibilidades se vuelven más imaginativas, inesperadas e incluso apocalípticas. Pero no siempre es así: un nodo de tercer orden puede representar un futuro probable si sus predecesores son muy probables.



Realizar un ejercicio de rueda de futuros con el equipo técnico suele ser tan divertido como revelador. Genera algunas historias convincentes sobre las posibles repercusiones de nuestras decisiones, lo que le confiere un gran potencial ético. Los tecnólogos no solemos tener la oportunidad de pensar de forma tan especulativa, o incluso fantasiosa: el énfasis de la industria en la entrega impide que nos planteemos fantasías sobre el futuro. Pero el diseño de futuros no consiste tanto en hacer predicciones exactas como en abrir los ojos del equipo a las posibilidades. Conjurar visiones compartidas del futuro es la piedra angular de la imaginación moral y una forma crucial de sacar a la luz consecuencias no intencionadas.

> Entonces, ¿por qué intentar predecir el futuro si es tan difícil, tan casi imposible? Porque hacer predicciones es una forma de avisarnos cuando nos vemos a la deriva en direcciones peligrosas. Porque la predicción es una forma útil de señalar rumbos más seguros y sabios. Porque, sobre todo, nuestro mañana es el hijo de nuestro hoy. A través del pensamiento y la acción, ejercemos una gran influencia sobre este niño, aunque no podamos controlarlo absolutamente. Pero es mejor pensar en ello. Lo mejor es intentar convertirlo en algo bueno. —Octavia Butler[22]

El diseño como provocación

Los futuros abstractos y teóricos carecen de vida y son difíciles de imaginar; también necesitamos experimentarlos. La imaginación moral debe implicar emoción, no sólo la lógica. Para que la gente se dé cuenta de las posibles consecuencias de sus decisiones, tenemos que pintar un cuadro vívido.

> A una persona virtuosa no le basta con comprender *intelectualmente* su deber moral de extender la compasión. Ni siquiera le basta con entender que sería irracional no hacerlo. También debemos encontrar la manera de *sentir* compasión, que es una experiencia que va más allá del intelecto. —Shannon Vallor[23]

¿Cómo podemos dar vida a estos futuros teóricos? ¿Cómo presentar escenarios convincentes que provoquen reacciones,

emociones y debates? Con diseño, por supuesto. Esta es la premisa del *diseño especulativo*, impulsado por Dunne & Raby en el Royal College of Art de Londres. Esta rama emergente del diseño se centra en cómo *podrían* ser las cosas, planteando preguntas hipotéticas (¿que pasaría si?) para suscitar conversaciones y decisiones sobre el futuro que queremos.

El diseño especulativo se basa en nuestros planteamientos de futuros. Podemos esbozar futuros potenciales e ilustrar un fragmento de la vida en esos futuros. Estas *ficciones de diseño* (*design fictions*) pueden adoptar la forma de cortometrajes, historias, juegos, cómics, juegos de rol u objetos: cualquier cosa que construya un mundo creíble.

Una ficción de diseño suele incluir un artefacto hipotético, un prototipo creíble de la tecnología en cuestión. Yo llamo a esto (con perdón) un *provocatype*. Un *provocatype* no es un "buen" diseño: no es una respuesta exhaustiva al encargo ni responde a todas las necesidades del usuario. Su objetivo es provocar la conversación entre las partes interesadas y, potencialmente, también entre los usuarios, si se introduce con cautela, con acuerdos de confidencialidad y las debidas precauciones. Un *provocatype* crea un curioso agujero de gusano entre el diseño y la investigación: es un producto diseñado que, sin embargo, sólo existe como sonda de investigación. La gran diferencia con respecto a los prototipos normales es que creamos *provocatypes* no sólo con la intención de resolver problemas, sino también de crearlos. Si tenemos éxito, un *provocatype* suscitará mejores reacciones que un debate hipotético.

Veamos un *provocatype* en acción, creado por la empresa de diseño *The Incredible Machine* para dos clientes holandeses del sector energético. El futuro que eligieron se centraba en la escasez de energía y los vehículos eléctricos compartiendo estaciones de carga públicas. Son extrapolaciones razonables de la situación actual: podríamos decir que es un futuro probable. Los diseñadores optaron por crear un *provocatype* de alta fidelidad del propio punto de recarga.

The Incredible Machine, *"Estación de carga transparente". Reimpreso con permiso.*

Como supuesto usuario, enchufas tu cable de carga (suministrado con el *provocatype*) en una toma libre, te auténticas tocando con la tarjeta de identificación y solicitas tu energía utilizando los diales. La pantalla de puntos muestra la cola de carga y los tiempos estimados

de finalización. Pero cómo se cargan los coches no es lo más interesante. La función principal del *provocatype* es explorar cómo un algoritmo podría priorizar la energía cuando la demanda supera la oferta. Nos da una idea de un futuro dirigido algorítmicamente —podríamos llamarlo *algocracia*— de aquí a una década. Todo esto gira en torno a la tarjeta de identidad, que los diseñadores también crearon como prototipo.

Reimpreso con permiso.

Cada usuario recibe una tarjeta relacionada con su posición en la sociedad. La tarjeta de un médico le permite saltarse la cola de carga, pero conlleva una penalización por uso indebido. Un delincuente recién salido de la cárcel recibe una tarjeta de libertad condicional, que le deniega prioridad de carga y limita su consumo de energía.

Los diseñadores, Marcel Schouwenaar y Harm van Beek, no lo proponen como la solución óptima, sino que han creado un objeto que estimula el diálogo y la imaginación moral. No podemos evitar preguntarnos cómo sería la vida si nuestro estatus social se plasmara en un documento de identidad digital. Vemos los posibles aspectos positivos —los servicios de emergencia, por ejemplo, no se verían excesivamente entorpecidos por la escasez de energía—, pero también

vemos cómo la algocracia y el Internet de las cosas podrían reforzar la estratificación social y la desigualdad.

Utopías y distopías

Genevieve Bell señala que habitualmente las predicciones tecnológicas suelen gravitar hacia los extremos de la utopía y la distopía.[24] Deberíamos ser cautos con ambas.

Los vídeos corporativos con visiones del futuro suelen representar una canónica utopía capitalista: interfaces gestuales en oficinas deslumbrantes, con una perfecta y tediosa colaboración global. Éticamente, estas ficciones de diseño están vacías: sus *provocatypes* hacen muy poco por provocar. Pero las utopías pueden ser peligrosas, además de aburridas. La búsqueda de una sociedad perfecta ha sido en ocasiones una puerta de entrada al extremismo; el control necesario para hacer las cosas bien puede convertirse fácilmente en totalitarismo.

Las distopías también pueden ser seductoras. Muchos diseñadores conocerán el juego de diseño *flip-it*,[25] en el que los participantes imaginan la peor solución posible al encargo, con la idea de invertir esas ideas para descubrir los principios de un diseño de éxito. Todo el mundo se lo pasa en grande dibujando calaveras y huesos, y la gente suele salir del ejercicio con ideas sorprendentemente profundas. Las distopías pueden ser cuentos con moraleja —las fábulas de Esopo rara vez tenían un final feliz—, pero también pueden ser cínicas y distantes. Alejan a los posibles colaboradores tanto como los atraen y hacen que el tecnólogo con mentalidad ética se gane la reputación de fantasioso y obstaculizador.

El futuro suele seguir un camino más matizado. Cuando animamos a la gente a ejercitar su imaginación moral, debemos evitar los extremos. El diseño de ficción (*design fiction*) ideal tiene un toque de ambigüedad moral, insinuando lo bueno y lo malo por igual. El diseño especulativo pretende provocar respuestas, pero deja que los espectadores las construyan por sí mismos, en lugar de obligarles a reaccionar de formas preestablecidas.

Disidencia de los usuarios y crisis

En estas reflexiones sobre el futuro no debemos perder de vista al ser humano. La imaginación moral debe girar en torno a las personas que vivirán en nuestros futuros hipotéticos y utilizarán nuestras tecnologías propuestas.

En *Design for Real Life*,[26] Eric Meyer y Sara Wachter-Boettcher recomiendan nombrar a un *disidente designado*. Se trata de un papel de antagonismo constructivo, especialmente útil en las sesiones de crítica. El trabajo del disidente consiste en desafiar las suposiciones del equipo, subvertir las decisiones y lanzar alguna que otra granada de desafío. Puede representar el papel de un usuario que se niega a proporcionar los datos solicitados o que se siente insultado por el tono de un mensaje de error. Meyer y Wachter-Boettcher insisten, sin embargo, en que es mejor rotar el papel: los equipos tienen facilidad para ignorar a los detractores reiterados, y llevar demasiado tiempo la túnica de la disidencia puede agriar hasta el alma más caritativa.

En *Design for Real Life* también destacan los momentos de crisis de los usuarios. Las sonrientes *Personas* pegadas en las paredes de las oficinas de empresas tecnológicas, muestran usuarios felices y productivos, aunque siempre increíblemente ocupados. Pero las personas reales no son arquetipos de madera. Entre nuestros usuarios los hay que se enfrentan a la pérdida de un empleo o a un duelo, cuya relación se está rompiendo o que luchan contra una enfermedad física o mental. Esos momentos están cargados de significado ético. Cómo tratamos a las personas en su momento más vulnerable es nuestro test moral más profundo, y a medida que nuestro software llegue a más partes del planeta, tendremos que ayudar a más gente en estos períodos de crisis. El disidente designado puede ayudarnos a imaginar cómo podrían producirse estas crisis en el futuro que elijamos, aunque también hay lugar para la investigación cuidadosa, como entrevistar a personas que hayan experimentado una adversidad similar. Esta investigación tiene sus propias cuestiones éticas delicadas, y lo mejor es dejarla en manos de investigadores formados, apoyados quizá por psicólogos cualificados para los casos más delicados.

Redefiniendo los stakeholders

El diseño de futuros y el diseño especulativo pueden revelar consecuencias no intencionadas, pero ¿qué pasa con las externalidades, es decir, los efectos sobre las personas que hemos pasado por alto? Como ya hemos dicho, los economistas sostienen que las externalidades necesitan regulación, pero la industria tecnológica también puede y debe intentar reducirlas ampliando el abanico de partes interesadas o stakeholders.

Todos los libros de texto de negocios ofrecen una guía paso a paso para el análisis de stakeholders, pero la mayoría sólo cubren a compañeros de equipo o a grupos sospechosamente homogéneos como "usuarios" o "habitantes". Esta perspectiva, que se ve reforzada por el enfoque individualista del diseño centrado en el usuario, hace que a menudo pasemos por alto a grupos importantes. Los stakeholders no son sólo las personas que pueden afectar a un proyecto; son también las personas a las que el proyecto puede afectar. Para forzarnos a tener en cuenta a las personas adecuadas, prueba a utilizar una lista de posibles interesados (véase el apéndice) y utilízala como aportación a los ejercicios de futuros y al proceso de diseño.

No todas los stakeholders o partes interesadas serán bienvenidas. Por ejemplo, en algunos casos, puede ser útil incluir a un delincuente, un terrorista o un troll, —a una *persona non grata*[27]— como parte interesada negativa, de modo que el equipo pueda debatir cómo reducir activamente el daño que esta persona puede causar. Puede que incluso merezca un tratamiento completo como una *"Persona"*, y darle un nombre, un escenario abusivo y una lista de motivaciones para considerar su perfil dentro del equipo.

Los stakeholders podrían incluir incluso creencias o conceptos sociales: cosas que valoramos en la sociedad pero que rara vez consideramos bajo nuestra influencia, como la democracia, la justicia o la libertad de prensa. Como ahora sabemos, la tecnología tiene el poder de perjudicar estas creencias o conceptos; incluirlos explícitamente entre las partes interesadas, o al menos reconocer su posible vulnerabilidad, podría ayudarnos a protegerlas.

¿Un juramento hipocrático?

La imaginación moral, el diseño del futuro, los *provocatypes* y los disidentes designados están muy lejos de nuestros métodos habituales. ¿No hay una forma más fácil? ¿No podríamos empezar por crear un juramento hipocrático para la industria tecnológica? Esta es una duda comprensible y común entre los recién llegados al campo de la ética tecnológica; tomando ejemplo de otras disciplinas, un juramento escrito parece un punto de partida obvio.

Un código ético podría ser útil en el momento oportuno, pero no está tan claro que sea una solución ética. Para empezar, ya se ha hecho antes. Decenas de intentos anteriores no han funcionado; ¿por qué iba a ser diferente este otro intento? Los diseñadores ya conocen, por ejemplo, el manifiesto *"First Things First"* (lo primero es lo primero) de 1964, que defendía que los diseñadores debían utilizar sus habilidades para el bien moral, no sólo para fines comerciales. Seamos sinceros, la reedición del manifiesto en 2000 sugiere que la primera publicación tuvo poco efecto a largo plazo. En el ámbito de la tecnología emergente, ya hay varios esfuerzos de codificación en marcha. La iniciativa del Diseño Éticamente Alineado (*Ethically Aligned Design*) del IEEE ha implicado a cientos de expertos y detalla varios principios clave como los derechos humanos y la responsabilización. Otras iniciativas similares son los principios de IA de Asilomar y la Declaración de Barcelona. Organizaciones profesionales como la *Association for Computing Machinery* (ACM, por sus siglas en inglés) también publican un código de conducta qué esperan de sus miembros.

Estos esfuerzos, que suelen surgir de intensos procesos de consulta o de congresos de élite, pueden ser engorrosos, pero son preferibles a los códigos escritos por un solo autor. La industria tecnológica ha sido testigo de una reciente oleada de lo que yo llamo *"codes of reckons"* (códigos de opinión), son simples listados morales de eminentes tecnólogos. Estos códigos no ayudan mucho a la causa de la tecnología ética. Los autores de estos códigos —conscientemente o no—se nombran a sí mismos árbitros éticos, proyectando una autoridad inmerecida. Estos documentos carecen de aportaciones públicas y caen directamente en la trampa de la tecnocracia.

Las convenciones éticas no resuelven por sí solas los problemas éticos: las cuestiones morales espinosas siguen invadiendo la medi-

cina y la ingeniería, a pesar de los prestigiosos códigos éticos que se han creado en estos campos. Los códigos pueden estructurar el debate ético, pero suelen ser demasiado vagos para resolverlo. Pensemos en dos máximas muy conocidas: la promesa bioética "Lo primero es no hacer daño" y el famoso *"Don't be evil"* (no seas malvado) de Google. Aunque concisas, ambas afirmaciones son en la práctica incómodamente imprecisas. ¿Qué es el mal? ¿Qué es el demonio (*evil*)? ¿Quién decide? ¿Cómo se resuelven las demandas contradictorias? Para responder a estas preguntas necesitamos algo más que una expresión; necesitamos formas de evaluar adecuadamente los argumentos éticos. (En el próximo capítulo hablaremos de ello). Sin este tipo de marco moral, las empresas pueden elegir cualquier definición del mal o del daño que justifique su camino. Por sí solo, "No seas malvado" significa poco más que "Oye, la ética importa". Pero no seamos injustos. Reconocer que la ética es importante fue en sí mismo una pequeña revolución en su momento y, aunque parezca una frase de poca importancia, *"Don't be evil"* se convirtió en parte del manifiesto del personal de Google más que un lema corporativo formal. En los últimos años se ha convertido en poco más que un reclamo utilizado por los críticos que se quejan de los errores de Google. Desde entonces se ha trasladado a la parte final de un código de conducta más detallado y útil desde el punto de vista ético.

Los códigos éticos de la tecnología también se enfrentan a problemas de cumplimiento. En muchas otras disciplinas, los colegios profesionales pueden inhabilitar a los que ejercen por una mala praxis, pero como la mayoría de los tecnólogos carecemos de acreditación oficial y la pertenencia a organizaciones profesionales es voluntaria, los códigos deontológicos del sector tecnológico son en gran medida inoperantes.

Por último, los códigos son mejores a la hora de censurar los malos comportamientos que de inspirar los buenos. En el peor de los casos, pueden inculcar una mentalidad de lista de control, en la que los profesionales creen que basta con seguir los pasos indicados para pasar el examen ético. Las listas de control (*checklists*) son útiles, pero pueden ser contraproducentes si no se capta el espíritu que las inspira. En el campo de la accesibilidad web, las Pautas de Accesibilidad de los Contenido Web (*Web Content Accessibility Guidelines*) han sido a la vez una ayuda y un obstáculo. Proporcionan consejos claros

sobre el desarrollo accesible, pero también han provocado que algunos equipos entiendan la accesibilidad como un mero ejercicio de verificación: comprobar las relaciones de contraste, ajustar algunos tamaños de fuente y ya está todo conforme. Los tecnólogos éticos saben que no es así. Son conscientes de que la accesibilidad tiene que ver con quién consideramos merecedor de nuestros esfuerzos, con el compromiso de tratar a todas las personas como personas. No debemos confundir un código ético con la propia ética. Lo que importa es la conversación y los resultados, no el papeleo. La ética debe convertirse en una costumbre, en una forma de pensar, en un conjunto de valores compartidos por todos los miembros del sector: como dice Cameron Tonkinwise, la *ethics as ethos*[28] (la ética como ethos).

Infraestructura ética y diversidad

Puede resultar más fácil y productivo codificar la ética dentro de cada empresa de forma individual, sobre todo si podemos recurrir a las políticas existentes. Los *valores fundamentales* —esencialmente una lista de las posturas y compromisos de la empresa— son un vehículo extendido e importante para la ética, y suelen contar con el apoyo de los directivos. Los equipos de proyecto también pueden crear *principios de diseño* que regulen las decisiones dentro de una determinada línea de producto o proyecto. Unos valores fundamentales sólidos y unos principios de diseño interiorizados son poderosos elementos de resolución de dilemas éticos: en caso de emergencia moral, consulte los principios acordados para obtener orientación. Esto significa que los valores fundamentales y los principios de diseño deben ser específicos. Algunas empresas eligen valores de una sola palabra: Los de Adobe son "genuino", "excepcional", "innovador" e "implicado". Reflexionar sobre los rasgos y cualidades de una vida moral puede ser importante —es la piedra angular de una rama de la ética que trataremos más adelante—, pero los valores de una sola palabra son demasiado escurridizos para toda una empresa. Dejan demasiado en el aire lo que significan y la gente puede tergiversarlos para sus propios fines en un debate: "¿Cómo puedes oponerte a este software de seguimiento? Es *innovador*".

Las frases son mejores. "Defender y respetar la voz del usuario" de

X (previamente Twitter) es un principio sólido, aunque moralmente ambiguo: ¿incluye defender la incitación al odio? Los valores fundamentales de Ben & Jerry's son muy específicos e incluso políticos: "Buscamos y apoyamos los medios no violentos de alcanzar la paz y la justicia. Creemos que los recursos gubernamentales son más productivos cuando se utilizan para cubrir las necesidades de las personas y no para levantar y mantener sistemas de armamento". Puede que sea *demasiado* específico —es difícil vivir realmente según los valores fundamentales a menos que puedas recordarlos—, pero no deja lugar a dudas sobre el tipo de empresa que Ben & Jerry's quiere ser.

Según el investigador Jared Spool,[29] un buen principio de diseño es reversible. Si puedes darle la vuelta al significado y consigues un principio válido para otro equipo o otra época, estás siendo específico. "Hazlo fácil para los usuarios" es una obviedad, no un principio de diseño; lo contrario sería absurdo. La prueba de reversibilidad no se ajusta tan bien a los valores fundamentales. A veces es útil apoyar explícitamente algo que debería ser moralmente obvio —"Nos preocupamos por el planeta", por ejemplo—, pero en caso de duda, sé específico.

Los valores fundamentales y los principios de diseño refuerzan la *infraestructura ética* de una empresa, del mismo modo que lo hace la diversidad de los equipos. Los equipos homogéneos tienden a centrarse en las ventajas potenciales de su trabajo para las personas como ellos, y son ajenos a los problemas que podrían infligir a un público más amplio. Las mismas divisiones que impregnan el mundo actual se ven e incluso se amplifican en la industria tecnológica actual.

> Si vives [en EE.UU.] cerca de un *Whole Foods*, si nadie de tu familia sirve en el ejército, si te pagan por contrato anual, y no por horas, si la mayoría de la gente que conoces terminó la universidad, si nadie que conozcas consume metanfetaminas, si te casaste una vez y sigues casado, si no eres uno de los 65 millones de estadounidenses con antecedentes penales... si alguna o todas estas cosas te describen, entonces acepta la posibilidad de que, en realidad, puede que no sepas lo que está pasando y que seas parte del problema). —Anand Giridharadas[30]

Los profesionales de la diversidad y la inclusión suelen describir

dos dimensiones de la diversidad: la *diversidad inherente* y la *diversidad adquirida*. La diversidad inherente se refiere a los rasgos innatos de un grupo, como el sexo, la orientación y el origen étnico, mientras que la diversidad adquirida se refiere a las perspectivas que las personas se han ganado a través de la experiencia. Ambos tipos de diversidad pueden actuar como un sistema de alerta temprana para la ética. Un equipo con una amplia diversidad inherente ofrecerá diferentes perspectivas y valores, mientras que a las personas abiertas a nuevas experiencias a través, por ejemplo, de los viajes, la literatura o los idiomas les suele resultar más fácil ejercer la imaginación moral. Aunque debemos reconocer el papel de los privilegios —no todo el mundo tiene la suerte de ver todas las maravillas del mundo—, absorber activamente nuevas experiencias suele reforzar las facultades éticas.

> Afortunadamente, nuestra capacidad de imaginación puede incrementarse. Buscar noticias, libros, películas y otras fuentes de historias sobre la condición humana puede ayudarnos a imaginar mejor la vida de los demás, incluso la de aquellos que se encuentran en circunstancias muy diferentes a las nuestras. —Shannon Vallor[31]

El mero hecho de entablar amistad y aprender de personas distintas a nosotros también ayuda, ya que fomenta nuestra comprensión mutua y, por tanto, una especie de diversidad adquirida de segunda mano. La búsqueda de la diversidad sugiere que también deberíamos abrazar la interdisciplinariedad. Poco a poco, la industria tecnológica está aprendiendo que las personas con formación no técnica, como la política, el derecho, la filosofía, el arte y la antropología, pueden aportar un gran valor, no sólo en términos de perspectivas profesionales diferentes, sino en la tan necesaria diversidad adquirida. Esperemos que la tendencia continúe.

CAPÍTULO 3
MECANISMOS DE PERSUASIÓN

En su influyente ensayo "¿Tienen política los artefactos?",[1] Langdon Winner concluye que sí, que la tienen. Desafía la idea instrumentalista de que los objetos son meros productos inertes de las fuerzas sociales que los crearon y defiende, en cambio, una visión más amplia: que los propios objetos influyen en cómo se distribuyen el poder y la autoridad, y en cómo se comportan las sociedades.

El ensayo examina, entre otras cosas, los pasos elevados en las autopistas de Long Island. Winner afirma que el planificador urbano neoyorquino Robert Moses construyó estos puentes inusualmente bajos para lograr "un efecto social concreto": la segregación. Los residentes pobres, y en particular los no blancos, solían viajar en autobús en aquella época; como estos autobuses no cabían bajo los puentes, estas personas quedaban excluidas de las playas de Long Island. Aunque el relato de Winner es hoy algo discutido, demuestra que incluso las enormes moles de hormigón y acero pueden imponer cambios sociales.

Los diseñadores son muy conscientes del poder de los objetos. Durante décadas, los diseñadores gráficos han intentado cambiar las actitudes y los comportamientos del público, ideando no sólo el atractivo paquete de cigarrillos, sino también la tranquilizadora señalización hospitalaria para los últimos días del fumador. Los objetos tecnológicos —ordenadores, teléfonos, dispositivos— pueden ser especialmente potentes a la hora de despertar nuevos deseos y

moldear comportamientos: nada tan lleno de lenguaje, luz y energía puede ser inerte.

Coacción frente a estímulo

A veces, los diseñadores introducen decisiones morales en el entorno a la fuerza, lo que significa que el usuario tiene que cumplir los deseos del diseñador. Los badenes obligan a los conductores a frenar; los seguros dificultan el disparo accidental de un arma. La tecnología digital también limita la capacidad de elección del usuario. A menudo oímos que el diseño es una conversación con el usuario; en tecnología, la conversación es lamentablemente unilateral. En palabras de Tristan Harris, antiguo especialista en ética del diseño de Google, "quien controla el menú controla las opciones". Si no se aprende un lenguaje de programación, no se puede hacer que un ordenador haga nada que su interfaz no permita. Por tanto, las decisiones de diseño otorgan a las tecnologías el poder de imponer comportamientos —y, por tanto, conductas morales— en ausencia del diseñador.

La coerción puede parecer poco ética, ya que limita el libre albedrío de las personas, pero gran parte de nuestra sociedad se basa en la coerción y el cumplimiento, especialmente nuestras leyes. En cambio, la coerción afecta dónde reside la responsabilidad moral. Uno no es responsable de un comportamiento que no ha elegido libremente; no culpamos a alguien obligado a punta de pistola a cometer un delito. La responsabilidad recae en el coercionador: un diseñador debe asumir la responsabilidad de cualquier decisión a la que obligue a un usuario. Los juicios militares han dejado claro, sin embargo, que las órdenes de los superiores no cuentan como coerción; los soldados pueden negarse a cumplir órdenes ilegales o inmorales, aunque paguen un alto precio por ello.

Hay artes persuasivas más sutiles que la coerción contundente. La "Teoría del Pequeño Empujoncito" (*Nudge Theory*), popularizada por Richard Thaler y Cass Sunstein, trata de orientar el comportamiento mediante cambios sencillos en los valores por defecto y el encuadre. El empujoncito ha florecido en el sector público —cláusulas de exclusión voluntaria (*opt-out*) para la donación de órganos, señales electrónicas que sonríen o fruncen el ceño ante la velocidad de un conductor —, pero Silicon Valley también es aficionado a esta teoría. El empu-

joncito no vulnera las libertades individuales que tanto aprecia la industria, pero sigue siendo una potente técnica para liberar a la gente de su dinero o su tiempo.

Los partidarios del empujoncito se apresuran a señalar que no reducen las opciones disponibles, sino que las amplían. Así, el empujoncito es una técnica de persuasión, no de coerción. Sin embargo, la persuasión sigue planteando problemas éticos. Como señalan Daniel Berdichevsky y Erik Neuenschwander, los primeros teóricos de la tecnología persuasiva, "los persuasores siempre se han movido en un terreno ético incómodo. Si una serpiente te persuade para que comas una fruta [...] ¿la culpa recae sobre ti o sobre la serpiente?"[2]

Aunque la persuasión no sea el plan explícito, el diseño siempre influye en el comportamiento. Un diseño tiene éxito si conduce al usuario a la información correcta o al siguiente paso del proceso; si se agranda un botón para hacerlo más visible, más gente lo apretará. Por tanto, todo diseño orientado a objetivos es un diseño persuasivo. Cualquier equipo con objetivos de rendimiento intentará maniobrar el comportamiento del usuario para alcanzar los objetivos de la empresa. Esto significa que no podemos simplemente comprometernos a no practicar nunca la persuasión: tendríamos que abandonar el diseño por completo. En lugar de ello, tenemos que lanzarnos con intención, reconociendo nuestras responsabilidades y eligiendo cómo abordar los retos éticos.

Patrones oscuros, atención y adicción

En un mundo racional, la persuasión sería sencilla: exponer los beneficios y los costes de cada opción y confiar en que el usuario tome la decisión correcta. Ojalá fuese tan simple. Los persuasores también deben apelar a los sesgos y las emociones, a lo que podemos considerar la debilidad humana. Sin embargo, el diseño conductual a menudo se viste con el cómodo manto del paternalismo, ligeramente condescendiente pero ampliamente benéfico. El empujoncito explota la debilidad humana, pero los partidarios del empujoncito argumentarían que lo hacen para que podamos superar esa debilidad. ¿No queremos todos llevar una vida más sana y responsable? Por tanto, la persuasión en sí se presenta como una herramienta benigna que nos

eleva a todos, ayudando a la gente a ascender el monte Maslow y alcanzar una cima esclarecida.

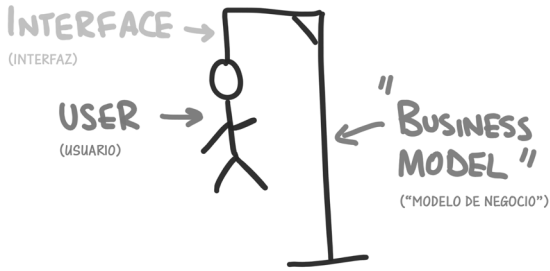
Pero la persuasión puede corromper a la humanidad tanto como ennoblecerla: las técnicas que pueden ayudar a la gente a perder peso también pueden utilizarse para animarla a no votar en las próximas elecciones. En el mundo de la tecnología, la persuasión poco ética adopta a menudo la forma de un *dark pattern*,[3] una interfaz intencionadamente engañosa que explota la debilidad cognitiva con fines lucrativos. En la actualidad, la mayoría de los patrones oscuros son molestias extorsivas —escasez artificial en sitios de hoteles, suscripciones engañosas—, pero el *dark pattern* se vuelve más amenazador a medida que las tecnologías se integran en la vida cotidiana. La tecnología persuasiva puede no ser fácilmente visible, pero sus campos de fuerza son cada vez más fuertes.

Las técnicas persuasivas pueden aplicarse a sí mismas: la tecnología puede persuadirnos para que la usemos más. La adicción a las redes sociales se ha convertido en un pánico moral en toda regla, alimentado por las historias de horror de los tabloides y la literatura de ciencia pop que reduce la investigación profunda a anécdotas dopaminérgicas y neuro-tonterías del lóbulo frontal. Las redes sociales se unen a un rico repertorio de lacras morales: los libros, los periódicos y los gramófonos estuvieron todos, en distintos siglos, vinculados a la caída segura del orden social.

Es habitual culpar a los modelos de negocio de Silicon Valley de la crisis de adicción. Se nos dice que los servicios gratuitos no son realmente gratuitos, sino que los usuarios pagan una moneda alternativa —su atención— que se monetiza a través de la publicidad. En sectores como los juegos gratuitos, esta equivalencia entre tiempo y dinero es dolorosamente literal: mira un anuncio y gana cincuenta monedas. Pero podemos rastrear esta sed de atención en la publicidad, los medios de comunicación e incluso la religión. La economía de la atención no es sólo una consecuencia de los motores de búsqueda y las redes sociales.[4]

El refrán moderno "Si no pagas por el producto, eres el producto que se vende" es simplista. Implica que pagar con atención es menos ético que pagar con dinero en efectivo, una postura discriminatoria que sugiere que los pobres son de algún modo menos merecedores de la tecnología que los ricos. También ignora la costumbre del mercado

de exigir que tanto los servicios de pago como los gratuitos acaparen toda la atención que puedan. En julio de 2017, un chico de 17 años de Guangzhou sufrió un derrame cerebral leve después de jugar durante cuarenta horas a Honour of Kings, el juego para móviles de Tencent. Tencent respondió con valentía, anunciando que limitarían el tiempo de juego de los niños. Las acciones de la empresa cayeron un 5,1%. Incluso los servicios de suscripción, repletos de ingresos recurrentes, se jactan en las presentaciones de sus resultados de las métricas de usuarios activos diarios y tiempo que pasan en la aplicación. Incluso si estás pagando por el producto, estás vendiendo tu atención. Un cínico podría ir más allá y argumentar que la razón de ser del diseño de experiencias es atrapar al usuario en una relación comercial, una afirmación sarcásticamente ilustrada por Jeff Veen.



Redibujado y utilizado con el permiso de Jeff Veen.

El grado de perjuicio de la economía de la atención depende en parte de si sus efectos se equilibran o no. Las pantallas que sólo nos alejan de otras pantallas son relativamente inofensivas, pero si la tecnología nos distrae de lo que los éticos llaman "la buena vida" —un concepto vago, pero que podría incluir razonablemente a la familia, los amigos, el trabajo productivo y la superación personal—, la tecnología se convierte en una fuerza de alienación, como predijo hace décadas el filósofo Karl Jaspers.

Parece que los temores de Jaspers se están haciendo realidad. Reed Hastings, Director General de Netflix, ha afirmado que "estamos compitiendo con el sueño".[5] El estadounidense medio pasa 3,1 horas al día en un dispositivo móvil, frente a sólo dieciocho minutos en

2008, sin que se produzca el correspondiente descenso en el uso de ordenadores de sobremesa.[6] Todos somos Sísifo renacido, reduciendo cada día el número de libros sin leer hasta que, en palabras de Ian Bogost, "el conflicto y el agotamiento sofocan el deleite y la utilidad".[7] Tenemos razón al temer que la tecnología erosione el resto de nuestras vidas.

La preocupación por la adicción aumentará a medida que las empresas tecnológicas compitan por la atención a largo plazo y los filones publicitarios dominados por la televisión y el cine, y a medida que más tecnologías inmersivas lleguen a nuestros hogares. Tanto la realidad virtual como la aumentada ofrecen el potencial de una hiperrealidad irresistible, como se predijo en la ciencia ficción. El tema común de estas historias es, según Jaspers, la alienación de la auténtica experiencia humana: un personaje de ciencia ficción que renuncia al mundo físico rara vez disfruta de un final feliz. Sin embargo, la amenaza ya no es puramente ficticia. El alarmante fenómeno del *hikikomori* ("repliegue sobre sí mismo") ha hecho que cientos de miles de jóvenes japoneses se aparten de la sociedad. A diferencia del teleadicto, el hikikomori no abraza la ociosidad por sí misma, sino que, incapaz de hacer frente a las tensiones del mundo exterior, se repliega sobre sí mismo. Muchos hikikomori se sienten atraídos por los medios inmersivos o los juegos. Aunque no debemos confundir causalidad y correlación, está claro que las tecnologías inmersivas, junto con la automatización, el colapso del comercio minorista local y el envejecimiento de la población, podrían hacer que la gente se apartara de la sociedad.

Experimentación

Las empresas tecnológicas han adoptado tanto el cambio de comportamientos que prueban innumerables diseños para encontrar las versiones más persuasivas. Tal vez un botón Comprar Ahora más grande aumente las ventas, o un guión diferente de asistente de voz fomente más consultas. Este enfoque empírico se ve reforzado por el amor de los tecnólogos por el método científico, inculcado en su época de estudiantes de ciencias, tecnología o ingeniería, y por el auge del *lean startup*. Los fans de *Lean* sostienen que la iteración es el mejor camino para conseguir que el producto encaje en el mercado: experi-

mentar con y sobre los usuarios se celebra como un paso natural en este proceso.

Cualquier proyecto que aprenda del comportamiento de los usuarios es un proyecto de investigación con usuarios, pero la industria ha optado tácitamente por eximir la experimentación de la ética de la investigación. A los usuarios no se les da el derecho a retirarse de los estudios. En las poblaciones experimentales se incluye sistemáticamente a niños. El consentimiento informado se deja de lado y se sustituye supuestamente por una frase de excusa en las condiciones del servicio. Una Junta de Revisión Institucional (IRB, por sus siglas en inglés) rechazaría una investigación académica tan chapucera, pero la industria argumenta que este nivel de supervisión ética neutralizaría la innovación. Mientras los reguladores hacen la vista gorda, las empresas experimentan de forma imprudente con los usuarios. La experimentación puede ser una forma eficaz de probar las mejoras de los productos, pero en algunas empresas se ha extendido más allá de los retoques de la interfaz y se ha convertido en investigación psicosocial, a veces con un escandaloso desprecio por el bienestar del público. Un ejemplo notorio es el estudio de Facebook sobre el "contagio emocional", que manipuló las noticias de 689.000 usuarios para averiguar si afectaban a su estado de ánimo.[8] La investigación se llevó a cabo bajo los auspicios de la propia política de datos de Facebook y no de una IRB. Facebook supone que el usuario acepta la política de privacidad y uso como condición para utilizar el servicio, pero la mayoría de los usuarios nunca la abren, y mucho menos la leen. Es irrisorio afirmar que esta política garantiza el consentimiento informado. No se ofreció ninguna opción de no participación, y los investigadores parecieron ignorar el impacto potencial del estudio en los usuarios con depresión, a pesar de que el documento resultante menciona esta enfermedad como objeto de investigaciones anteriores.

Muchas empresas tecnológicas respondieron a la protesta encogiendo de hombros, y alegando que así es como funciona la tecnología. OkCupid se jactó: "¡Experimentamos con seres humanos! (Como todo el mundo)", y los investigadores de Facebook expresaron su sorpresa por la reacción en discursos de apertura. Estas defensas despreocupadas de la experimentación tienen su origen en los efectos deshumanizadores de la escala y el sesgo cuantitativo de la industria. Si unimos el enorme alcance de las empresas tecnológicas a la

creencia de que el progreso debe ser mensurable —que los objetivos y resultados clave (OKR), los usuarios activos o las tasas de conversión son los únicos barómetros del éxito que merecen la pena—, a menudo triunfa una cultura de orientación al objetivo. Poco a poco, los usuarios no se convierten en razones de ser, sino en sujetos de experimentación, en medios para que los equipos alcancen sus propios objetivos. No vemos clientes, sino masas.

Persuasión y poder

La dinámica de la persuasión suele ser política. Incluso el aparentemente benigno empujoncito tiene efectos partidistas. Tanto los ciudadanos como los políticos encuentran más ética la idea del empujoncito cuando los ejemplos que se dan coinciden con la política de los sujetos. Incluso los libertarios, que suelen desconfiar de las tendencias controladoras de los empujoncitos, dejan a un lado su escepticismo cuando aprueban la estrategia de quienes lo dan.[9]

En el momento álgido de la Primavera Árabe de 2010-12, la tecnología parecía emancipadora, una fuerza positiva para desarraigar la jerarquía y la opresión. Qué ingenuo parece eso ahora. Hoy, Internet se ha convertido en el principal campo de batalla para la persuasión política, la propaganda y la desinformación. Mediante la manipulación de canales de información como las redes sociales, los partidos y las naciones pueden competir por la supremacía narrativa: una estrategia que a veces se conoce como operaciones de influencia.

La estructura de la web ayuda a estos esfuerzos. Fuentes de noticias respetadas, publicaciones especializadas y fábricas de propaganda pueden llegar a audiencias globales; todas están a una sola petición HTTP de distancia. Las conspiraciones modernas también parecen tan legítimas como cualquier historia respetable. En el pasado, podíamos identificar la literatura fraudulenta por su formato (feos garabatos y malas fotocopias), pero ahora los editores con plantillas como Medium o Squarespace permiten a cualquiera publicar información en un formato creíble. La desinformación es estéticamente equivalente a un comunicado de prensa legítimo, y una publicación en las redes sociales de un conspiracionista solitario tiene el mismo aspecto que un anuncio oficial de la BBC.

El hipertexto representa el conocimiento de un modo que fomenta

la exploración, la fragmentación y la reasociación. Por tanto, el hipertexto tiende a romper las narrativas centralizadas y lineales y, en su lugar, fomenta la *apofenia*, un hábito de imponer relaciones a cosas inconexas.[10] Las conspiraciones florecen en la oscuridad, en los huecos entre la información fiable, permitiendo a los desempoderados explicar su falta de poder. En esta batalla de las líneas temporales, no importa lo descabellada que sea la narrativa, hay gente dispuesta a creer. *Furries*, terraplanistas y fascistas por igual pueden construir sus propios relatos a partir de los fragmentos digitales de la web y difundirlos en sus comunidades.

Desde el "turismo de la paranoia" del Pizzagate hasta el terrible desenmascaramiento del terrorista erróneo de Boston por parte de Reddit,[11] extremistas e ideólogos han explotado inteligentemente al público utilizando la infraestructura persuasiva de Silicon Valley. Sin embargo, las empresas tecnológicas han negado cualquier papel en la persuasión política, aferrándose a su excusa instrumental: somos plataformas neutrales, no empresas de medios de comunicación. Esta es una defensa endeble. Ninguna industria que gasta millones en presionar para que se desregule puede alegar neutralidad política. La negación de la influencia política por parte de Facebook resulta especialmente descarada cuando sus páginas de publicidad en colaboración presumen de una "estrategia de contenidos específicos para el público con el fin de cambiar significativamente la intención de voto".[12]

Hay varias razones por las que las empresas tecnológicas han tardado en erradicar la información perjudicial y engañosa. Identificar este contenido es ciertamente difícil; ninguna empresa contratará a grandes equipos de comprobación de hechos, y los métodos automatizados arrojarán muchos falsos positivos. Pero, fundamentalmente, a las redes sociales no les importaba mucho la calidad de la información compartida, siempre que se compartiera; casi cualquier cosa que alcanzara objetivos era bienvenida. Las empresas sólo empezaron a prestar la debida atención a la propaganda generada cuando los políticos les exigieron responsabilidades y arrastraron a los ejecutivos ante comités y tribunales.

En retrospectiva, los fallos de la industria en materia de propaganda tienen orígenes bien conocidos. En su prisa por construir, los tecnólogos no tuvieron en cuenta cómo las estructuras y las posibili-

dades de sus nuevos sistemas podrían tener consecuencias imprevistas. Los equipos tecnológicos no consiguieron mitigar los riesgos, lo que significa que la sociedad tiene que asumirlos en forma de extremismo resurgente y conspiración.

Persuasión automatizada

La persuasión automatizada —operadores artificiales con sus propias formas de inducción algorítmica— puede suponer una amenaza aún mayor para la verdad y la democracia. No debemos repetir los mismos errores.

Los bots son ya una amenaza persuasiva viable. Al analizar los ecosistemas digitales de las elecciones estadounidenses de 2016, Berit Anderson y Brett Horvath descubrieron "una máquina de propaganda de inteligencia artificial armamentizada"[13] que se basaba en perfiles de Cambridge Analytica, scripts automatizados y una profunda red de sitios de propaganda. Aprovechando el apetito conspirativo de los marginados, las cuentas de desinformación en las redes sociales azuzaron la disidencia en las comunidades simpatizantes. Académicos de Oxford observaron un patrón similar, aunque más reducido, durante el referéndum del Brexit: a las cuentas de propaganda de X (previamente Twitter) utilizadas anteriormente para sesgar las opiniones sobre el conflicto entre Israel y Palestina se les dio una capa de pintura nacionalista británica y se lanzaron al nuevo debate.[14]

La información sobre las campañas del Brexit y Trump ha sido inconsistente: muchos de los llamados bots eran trolls pagados, aparentemente parte de las operaciones de influencia de estados rivales. Es comprensible que se hayan utilizado etiquetas chapuceras —todavía no se entiende cómo nuestras tecnologías se están volviendo en nuestra contra—, pero la persuasión automatizada ha desempeñado sin duda un papel en la reciente agitación política. Su influencia no hará sino aumentar.

La persuasión es un candidato ideal para el *machine learning*. Podemos definir métricas sencillas que queremos maximizar (más seguidores, clics y retweets parecen representaciones adecuadas de la influencia), ofrecer una gran cantidad de datos de comportamiento que extraer y proponer cientos de parámetros potenciales que ajustar. Los bots políticos pueden probar docenas de enfoques conversa-

cionales, hashtags y eslóganes; un bot de diseño puede probar innumerables permutaciones de la interfaz para inducir a un cliente potencial a comprar. Amazon ya emplea sugerencias automatizadas a gran escala.

> A través de nuestro programa *Selling Coach*, generamos un flujo constante de empujoncitos de *machine learning* (más de 70 millones en una semana normal), que alertan a los vendedores sobre oportunidades para evitar quedarse sin existencias, añadir selección que se está vendiendo y afinar sus precios para ser más competitivos. Estos avisos se traducen en miles de millones de aumento de ventas para los vendedores.[15]

Investigadores del Instituto Tecnológico de Georgia descubrieron que las personas eran sorprendentemente susceptibles a la persuasión automatizada (*machine persuasion*) en caso de emergencia.[16] Los participantes en la prueba fueron recibidos por un rudimentario robot y se les ordenó que lo siguieran hasta el laboratorio; la mitad de las veces el robot se equivocó de camino, para dar la impresión de que no era un conductor precisamente competente. A mitad del estudio, los investigadores inundaron el pasillo contiguo con humo falso, activando una alarma de incendios. Durante la falsa emergencia, ningún participante escapó por donde había entrado ni se dirigió rápidamente a una salida de emergencia: todos siguieron las instrucciones del robot de dirigirse a una habitación trasera, aunque antes hubieran visto al robot cometer errores o averiarse. Probablemente, los participantes sabían que la emergencia era falsa (un IRB tendría dudas sobre un estudio que hiciera temer de verdad por la vida de las personas), pero aun así "confiaron" en la máquina mucho más de lo que esperaban los investigadores.

La tecnología emocional o afectiva perfeccionará aún más las herramientas de persuasión. La nueva valla publicitaria londinense *Piccadilly Lights* utiliza cámaras ocultas para deducir el sexo, la edad y el estado de ánimo de las personas que se encuentran en las inmediaciones, de modo que pueda ofrecer anuncios adecuados a la audiencia: el espacio público se convierte en una mina de datos. El final de este escenario —un ente artificial capaz no sólo de leer los gestos, la entonación y el lenguaje corporal, sino de imitarlos en sus respuestas,

adaptando no sólo lo que dice, sino cómo lo dice— será un manipulador extraordinario.

La persuasión automatizada es estructuralmente muy diferente de las formas existentes de persuasión de masas como la publicidad. Los algoritmos persuasivos pueden responder a los cambios con rapidez, aprender de millones de éxitos o fracasos en otros lugares de la red y pueden ser altamente personalizados. Con suficientes datos y entrenamiento, un algoritmo puede presentar un mensaje convincente y adaptado a cada individuo; el marketing de talla única deja paso a un sistema que pulsa sólo sus botones más sensibles. La jurista Karen Yeung sostiene que la persuasión automatizada es tan diferente de sus predecesores monolíticos que merece un nuevo nombre: *hypernudge* (un sistema de empujoncitos interconectados, dinámicos, omnipresentes y actualizados constantemente). Los *hypernudges* podrían transformar la persuasión en coerción. En un estudio de 2017, los investigadores crearon un aumento de las compras del 50% simplemente adaptando los anuncios de Facebook a los tipos de personalidad inferidos de los usuarios.[17] Está claro que en el futuro esta elaboración de perfiles se hará más sofisticada y se vinculará a mensajes más persuasivos. Si podemos explotar las debilidades de alguien en el momento oportuno, ¿cuándo un empujoncito se convierte en un empujón?

Quizá el mayor desafío ético del *hypernudging* sea su invisibilidad. No hay forma de saber si una cámara está alimentando un motor de persuasión; pronto no sabremos si un servicio de asistencia está atendido por humanos o por algoritmos de *hypernudging*. La tecnología conectada conlleva un desequilibrio de poder implícito: el público no tiene conocimiento de lo que hace la red ni recurso alguno contra su explotación. Históricamente, la persuasión de masas era homogénea y visible: todo el mundo veía los mismos periódicos, anuncios y programas políticos. Esto significaba que podían ser criticados. La gente podía oponerse en masa a la persuasión engañosa o poco ética, y las autoridades podían exigir que se retirara un mensaje. Pero los *hypernudges* invisibles dejan menos margen para la protesta. Con un contenido persuasivo adaptado al individuo y suministrado en un dispositivo personal, será más difícil una oposición unificada, y las autoridades tendrán dificultades para tomar medidas correctivas.

El precio también tiene efectos persuasivos. La determinación

dinámica de precios no es nada nuevo —las aerolíneas llevan años en ello—, pero pronto podría llegar a una gama más amplia de industrias. Los algoritmos podrán afinar los precios para mantener las existencias, manipular la demanda y, por supuesto, extraer el máximo beneficio. Las etiquetas electrónicas de precios conectadas en red permiten a los minoristas ajustar los precios de forma instantánea, lo que reduce los residuos y la administración en el punto de venta, pero también permite a los minoristas subir los precios en momentos de máxima demanda: precios dinámicos en el mostrador de helados.

En teoría, la determinación algorítmica de precios podría tener algún beneficio social. Ajustar el precio a la capacidad de pago de una persona podría ayudar a hacer frente a la desigualdad, y explicar las variaciones de precios podría esclarecer cadenas de suministro opacas: tu café con leche es más caro esta semana gracias a las tormentas que azotan la producción en Vietnam. Pero los clientes suelen ver la discriminación de precios como algo injusto, lo que significa que hoy en día a menudo se encuentra con una feroz reacción en contra. Si no que se lo pregunten a Orbitz, que fue pillada ofreciendo a los usuarios de Mac precios de hotel más altos, bajo la presunción de unos ingresos más elevados.[18]

Si la fijación algorítmica de precios se generaliza a pesar de esta oposición, el único recurso del público podría ser confundir o sesgar colectivamente las señales de precios: la manipulación de precios (*price hacking*). La manipulación de precios ya se ha registrado entre los conductores de Uber: al acordar desconectarse al unísono, los conductores crean un déficit de oferta que instiga la subida de precios. En un mundo de precios algorítmicos rutinarios, los clientes podrían

seguir su ejemplo, absteniéndose de un producto para hacer caer los precios, y luego haciendo acopio masivo. Esto podría a su vez desencadenar mercados secundarios de reventa, o incluso algún tipo de comercio de futuros. Los amigos pondrán en común la información sobre precios y pedirán a quien consiga el mejor precio que compre en nombre de los demás. La fijación algorítmica de precios puede incluso crear cárteles de facto por accidente: si los algoritmos aprenden que los competidores igualarán inmediatamente los recortes de precios, pronto aprenderán a mantener los precios altos. En 2011, un duelo entre algoritmos de Amazon Marketplace que respondían mutuamente a los cambios de precios provocó que un libro de genética agotado se cotizara a 23,6 millones de dólares.[19] La escalada de precios y el bloqueo pueden convertirse en algo habitual, incluso sin intención delictiva.

Colapso de la evidencia

Dos piezas clave de las pruebas contemporáneas —el habla y el vídeo —pronto serán falsificables, lo que distorsionará aún más el panorama persuasivo. Ya existe un software convincente de conversión de texto a voz, aunque es exigente desde el punto de vista computacional, y los vídeos *deepfake* indetectables están a sólo un par de años de distancia. Cuando podemos simplemente importar un texto incriminatorio y ver una interpretación exacta de nuestro político más odiado, tendremos que reconsiderar lo que creemos que es verdad. Las fotos ya son inútiles como prueba: una vez que el audio y el vídeo sigan el ejemplo, ¿qué podrá servir como registro exacto de los hechos?

La empresa de síntesis de audio Lyrebird es una de las pocas compañías tecnológicas que publica una declaración ética.

> Imagina que hubiéramos decidido no publicar esta tecnología en absoluto. Otros la desarrollarían y quién sabe si sus intenciones serían tan sinceras como las nuestras: podrían, por ejemplo, vender la tecnología sólo a una empresa concreta o a una organización malintencionada. En cambio, nosotros estamos poniendo la tecnología a disposición de cualquiera y la estamos introduciendo de forma gradual para que la sociedad pueda adaptarse a ella, aprovechar sus aspectos positivos

> para el bien y evitar al mismo tiempo las aplicaciones potencialmente negativas.[20]

Que la declaración exista es de agradecer, pero su contenido es pésimo. (Discutiremos su endeble argumento ético – "Si no lo hacemos nosotros, lo hará otro"– en el capítulo 5.) No basta con que las empresas de tecnología transformadora adviertan del riesgo ético y dejen que la sociedad lo resuelva. Esto es instrumentalismo en su forma más peligrosa; los tecnólogos deben comprender y mitigar activamente los daños que pueden causar sus productos.

El colapso de las pruebas amenaza no sólo nuestra comprensión de los hechos y la actualidad, sino también nuestras relaciones personales. Si no podemos estar seguros de que la persona que nos llama por teléfono o por videoconferencia es quien creemos que es, se abre la puerta a la manipulación generalizada. Las tecnologías basadas en la confianza como la criptografía o el *blockchain* (la cadena de bloques) podrían ayudar, pero éstas requerirán un diseño excelente y de calidad para el consumidor. Si sólo unos pocos técnicos pueden aplicar estas garantías, la anarquía de la información reinará para todos los demás.

Justificar la persuasión: la ética popular

Está claro que la persuasión tiene implicaciones complejas. Tenemos que hacernos esa eterna pregunta ética: ¿dónde debemos trazar la línea? ¿Qué divide un *dark pattern* poco ético de la persuasión benéfica? Empecemos con algunos preceptos éticos comunes.

La justificación más débil de la persuasión —o, de hecho, de cualquier otra cosa— es que todo el mundo lo hace. Se trata de una trampa ética clásica, identificada hace siglos por David Hume y conocida como el *problema del ser y el deber ser*. Es un error teórico derivar lo que deberíamos hacer (deber ser) de cómo actúa actualmente la gente (ser). Las elecciones morales de nuestros competidores son irrelevantes para las nuestras. Del mismo modo que un pelotón de ciclistas tramposos no justificaba el consumo de drogas de Lance Armstrong, la despreocupada defensa de OkCupid de la experimentación persuasiva como algo habitual se topa de bruces con el problema del ser y el deber ser.

La *regla de oro* —haz lo que te gustaría que te hicieran a ti— es más

útil. Este proverbio de reciprocidad se encuentra en antiguos sistemas de creencias, desde el Levítico hasta Confucio. Aplicada al diseño persuasivo, la regla de oro sugiere que sólo debemos persuadir a alguien para que haga algo que haríamos nosotros mismos, o de lo que estaríamos encantados de que alguien nos persuadiera. El mayor defecto de la regla de oro es su egocentrismo. Anima a todo el mundo a verse a sí mismo como el árbitro ético ideal, tanto si sus intereses coinciden con los de los demás como si no. La regla de oro ignora la variedad de los deseos humanos y el papel que desempeña el contexto en las elecciones éticas.

Tal vez deberíamos, en cambio, tratar a los demás como a ellos les gustaría ser tratados: la *regla del platino*. En otras palabras, sólo deberíamos persuadir a las personas para que actúen en su propio interés. Deberíamos detenernos aquí para distinguir el interés individual del interés público, una idea que se encuentra a menudo en la ética periodística. Las historias que operan en zonas éticas grises, como las que vulneran la intimidad o implican engaño, suelen someterse a una prueba de interés público. Esta decisión pondera el daño potencial a los individuos frente al bienestar de la sociedad; un redactor jefe considerará a menudo que las historias que aumentan la responsabilidad y la transparencia a expensas de los transgresores son de interés público.

Algunas tecnologías persuasivas tienen un componente de interés público. Una aplicación para perder peso podría evitar miles de muertes relacionadas con la obesidad. Pero el bien común tiene una carga política: los intentos de especificar cómo deben vivir los demás suelen llevar la marca del autoritarismo. El interés público es complejo y no siempre es un enfoque ético útil.

Si sólo debemos persuadir a las personas para que actúen en su propio interés, ¿quién decide cuáles son esos intereses? Los tecnólogos probablemente no sean las personas adecuadas para tomar esta decisión; suelen favorecer lo científico frente a lo espiritual, la acción frente a la reflexión y el progreso frente al statu quo, valores que pueden no ser los adecuados para el individuo en cuestión. Pero preguntar simplemente a la gente cuáles son sus mejores intereses también tiene sus fallos: las opiniones declaradas de la gente no son fiables, y a veces todo el mundo contradice sus propios intereses al

buscar cosas que limitan su capacidad de prosperar, como el tabaco y el alcohol.

Si no podemos preguntar a la gente cuáles son sus mejores intereses y es impropio especificar intereses en nombre de otros, estamos en un aprieto, divididos entre un deseo paternalista de ayudar a los demás y un respeto tolerante por la libertad de elección de la gente. Estas sencillas directrices éticas —perdón por el calificativo peyorativo, "ética popular"— no resuelven el problema.

Teorías persuasivas

Algunos diseñadores y estudiosos han propuesto directrices para los sistemas persuasivos. Daniel Berdichevsky y Erik Neuenschwander sugieren, entre otros principios, que deberíamos juzgar la persuasión en función de si sería apropiada en persona, liberada de la tecnología.[21] Esta cuestión tiene el valioso efecto secundario de restaurar el contexto personal que la tecnología despoja con tanta frecuencia. En *Persuasive Technology*, BJ Fogg sugiere que utilizar las emociones negativas para persuadir es éticamente cuestionable.[22] La persuasión de masas hace mucho de esto —los publicistas juegan con la envidia; los políticos resentidos apelan a la xenofobia rancia—, pero deberíamos ser cautelosos con el problema del ser y el deber ser. Las advertencias en las cajetillas de cigarrillos son quizá más defendibles; existe el argumento de que el fin (salvar vidas) justifica los medios (utilizar el miedo para persuadir). Sin embargo, si rechazamos la idea de apelar a las emociones negativas del sujeto, deberíamos impedir que las tecnologías muestren también estas emociones. No importa lo avanzada que sea la nueva versión, una IA nunca debe reaccionar con agresividad si su usuario decide no actualizarla.

Casi todos los teóricos coinciden en que no es ético engañar con fines persuasivos, incluido Richard Thaler, que lo incluye en sus principios del empujoncito ético.[23] Pero consideremos el *botón de placebo*, un control sin función como el botón de cerrar la puerta en muchos ascensores o la opción de guardar en ciertas aplicaciones web. La motivación —dar a los usuarios una sensación de control— es benévola pero los medios son engañosos. ¿Son poco éticos los botones de placebo? Siguen respetando la voluntad del usuario —la puerta se sigue cerrando, los

ajustes se siguen guardando— y quizá sea preferible una mentira piadosa a la verdad: usted no tiene el control, lo tiene la tecnología. Sin embargo, en caso de duda, es mejor evitar la persuasión engañosa.

La investigadora en diseño social Nynke Tromp sugiere que clasifiquemos la persuasión por su fuerza y visibilidad, creando cuatro tipos de influencia: decisiva, coercitiva, persuasiva y seductora.[24] Digamos que estamos diseñando un centro de energía inteligente y queremos que la gente conserve energía. He aquí algunos posibles enfoques de diseño, asignados a estas cuatro categorías.

Las implicaciones éticas del cuadrante superior izquierdo parecen las más significativas. Una casa fría puede ser peligrosa para los ancianos, y un vehículo con poca carga puede resultar desastroso en caso de emergencia, pero si el dispositivo toma decisiones unilaterales e invisibles ambas cosas podrían ocurrir.

Las formas fuertes de persuasión pueden estar justificadas en algunas ocasiones, pero las formas más débiles suelen situarnos en un terreno ético más seguro. Luciano Floridi distingue los empujoncitos informativos de los estructurales.[25] Un empujoncito informativo cambia la naturaleza de la información disponible —etiquetar los aperitivos poco saludables, por ejemplo— mientras que un empujoncito estructural cambia los cursos de acción disponibles, como mover

esos mismos aperitivos fuera de nuestro alcance. El empujoncito informativo es más débil que el estructural, pero más respetuoso con la libre elección.

Los objetos persuasivos en el mundo físico suelen ser visibles o incluso estar resaltados, como los radares de tráfico, pero las limitaciones digitales suelen ser invisibles. Con un bloqueo de volumen por software, los usuarios nunca sabrán cuánto más altos y dañinos podrían ser sus auriculares. ¿Es transparencia la respuesta? ¿Deberían las tecnologías persuasivas simplemente anunciar su presencia y sus métodos? Es una idea prometedora, con dos advertencias. En primer lugar, la persuasión puede requerir invisibilidad; revelar la persuasión podría hacerla ineficaz. En segundo lugar, divulgar todos y cada uno de los métodos persuasivos sería engorroso y una distracción. Las explicaciones y advertencias salpicarían nuestras tecnologías; los usuarios acabarían por ignorarlas. Muchas técnicas persuasivas forman parte de lo que consideramos un buen diseño, como garantizar que las etiquetas sean claras y que las llamadas a la acción estén resaltadas. Tiene que haber cierto equilibrio. Divulgar los métodos persuasivos es un objetivo noble, pero quizá sea mejor hacer que esta información esté disponible en lugar de destacada. Discutiremos un ejemplo en breve.

El papel de la intencionalidad

También deberíamos considerar la posibilidad de revelar nuestra intención persuasiva: el porqué detrás del diseño. La intención surge a menudo en ética, y es la piedra angular del *principio del doble efecto*, que afirma que el daño es a veces aceptable como efecto secundario de hacer el bien. El doble efecto se utiliza a menudo en los casos de eutanasia: los médicos pueden aumentar la dosis de morfina de un paciente moribundo para aliviar el sufrimiento, aun a sabiendas de que la dosis puede resultar mortal. Si la intención fuera simplemente matar al paciente, el médico sería moral y legalmente responsable, pero la mayoría de las autoridades optan por no enjuiciar cuando existe una defensa convincente del doble efecto (aliviar el dolor).La mayoría de los expertos sugieren que la persuasión ética necesita una intención positiva. En palabras de Thaler, "siempre debe haber una razón buena y clara de por qué el empujoncito mejorará el bienestar

de los empujados". Debemos ser honestos sobre nuestra verdadera intención al diseñar sistemas persuasivos. ¿Por qué quiero que la gente siga mis consejos? ¿Qué hay para ellos? ¿Qué gano yo con ello? Por desgracia, la intención es menos útil cuando se examinan las decisiones de otras personas. La pregunta "¿Tenía usted buenas intenciones?" es vulnerable a todo tipo de excusas; como observó Benjamin Franklin, ser "una criatura razonable [...] le permite a uno encontrar o inventar una razón para cada cosa que se le ocurre hacer". Un colega sin escrúpulos puede inventar una excusa plausible para prácticamente cualquier transgresión ética. ¿La sobrevalorada extensión de garantía? Es inestimable para la pequeña fracción de usuarios cuyo producto se rompe. ¿Y obtener los contactos del usuario sin permiso? ¡Imagínate lo contentos que se pondrán los usuarios cuando sepan que sus amigos se han unido!. Sugiere que alguien ha actuado con una intención impura, y a menudo te responderán con retorcidos argumentos de doble efecto y gestos de indignación burlona. Centrarnos sólo en la intención también nos permite escabullirnos de las consecuencias imprevistas. A los usuarios no les importa si tenemos intención de hacer daño o no; les importa si causamos daño.

Introducción a la deontología

Las teorías persuasivas y las preguntas sinceras sobre la intención son herramientas éticas útiles, pero quizá necesitemos algo más riguroso, algún conjunto de reglas morales a seguir. Éste es el fundamento de la *ética deontológica* (o ética del deber), una de las tres escuelas de la ética moderna. Los deontólogos creen que la ética se rige por normas y principios, y que tenemos el deber moral de atenernos a estas normas. Esto puede hacer que los deontólogos sean algo rígidos: si creemos que tenemos el deber moral de decir siempre la verdad, es difícil justificar que mintamos a la policía secreta sobre dónde se esconde nuestra familia. Los deontólogos llevan vidas de principios, pero también vidas de abnegación y, ocasionalmente, de honroso sufrimiento. Dicho esto, los deontólogos suelen destacar en la resistencia a la presión ética; su creencia en las normas y la integridad significa que establecen límites claros y desafían el mal comportamiento.

Immanuel Kant, pionero del pensamiento deontológico, propuso una idea poderosa: cuando nos enfrentamos a una elección ética,

debemos universalizar nuestro pensamiento. Kant sugirió que imaginemos si nuestras acciones serían aceptables como ley universal de comportamiento. *¿Qué pasaría si todo el mundo hiciera lo que estoy a punto de hacer?* Esta versión simplificada de la teoría más importante de Kant[26] es un estímulo ético inestimable para los tecnólogos. Nos centra en los futuros que nuestras decisiones podrían crear y nos obliga a ver las opciones éticas desde perspectivas sociales más amplias.

Kant también planteó otra pregunta deontológica útil: *¿estoy tratando a las personas como fines o como medios?*[27] Esto merece una breve explicación. Para nuestros propósitos, la pregunta plantea si estamos utilizando a las personas —usuarios, partes interesadas, la sociedad en general— para nuestro propio éxito, o tratándolas como individuos autónomos con sus propios objetivos. Los diseñadores no suelen tener problemas con la cuestión de los fines o los medios, ya que tienden a creer profundamente en la importancia de los objetivos de los usuarios. La cuestión tiende a ser más difícil cuando nos la planteamos sobre decisiones que afectan a toda la empresa, en particular las que afectan a millones de personas.

Aunque los deontólogos están de acuerdo en que debemos vivir de acuerdo con unas normas morales, no las especifican: la cuestión es que tenemos que descubrirlas como sociedad. Las preguntas anteriores son buenas pautas, pero aún tenemos que esforzarnos para traducirlas en acción. Veamos cómo nuestras dos pruebas éticas nos ayudan a desenredar nuestras complicaciones persuasivas.

¿Deberíamos incluir un *dark pattern* engañoso que no ofrezca ningún beneficio al usuario pero que aumente nuestros beneficios? ¿Y si todo el mundo hiciera lo que estoy a punto de hacer? Si toda la tecnología estuviera plagada de *dark patterns*, las empresas podrían ganar más, pero nuestras tecnologías —y probablemente nuestras vidas— serían peores. Los usuarios se sentirían engañados y perderíamos la confianza que nuestra industria necesita urgentemente. Así que la respuesta deontológica es clara: no, no debemos incluir este *dark pattern*.

¿Y qué hay de la economía de la atención? Un mundo en el que todos pagásemos por nuestros productos preferidos con atención en lugar de con dinero en efectivo no sería malo en sí mismo; aunque, como pronto veremos, puede tener dolorosas implicaciones para la

privacidad. Los problemas surgen cuando un usuario es realmente adicto, hasta el punto de que perjudica su bienestar general. Si animamos a los adictos a utilizar nuestros servicios, ¿estamos tratando a estas personas como fines o como medios? Es fácil: como medios. Les ofrecemos continuamente algo que les perjudica, mientras nosotros nos beneficiamos. Un deontólogo argumentará que las empresas tecnológicas tienen el deber de intervenir en los casos de uso perjudicial. A diferencia de una empresa tabaquera, que no puede detener a un fumador concreto, las empresas tecnológicas podrían identificar a los usuarios problemáticos desde la distancia y tomar medidas. Esto podría implicar cualquier cosa, desde un toque ligero —reducir las notificaciones o mostrar una alerta de cansancio del tipo "¿Hora de un descanso?"— hasta la excomunión total, prohibiendo la tarjeta de crédito de un usuario y cerrando su cuenta. Una empresa que presta servicios a sabiendas a un usuario adicto está utilizando a ese adicto sólo como medio para su éxito comercial, y está cruzando la línea ética.

Experimentación ética

Si ponemos los experimentos bajo el microscopio deontológico, hay mucho que mejorar. Nuestra primera prueba deontológica —¿qué pasaría si todo el mundo hiciera lo que estoy a punto de hacer?— sugiere que la idea de realizar experimentos no es en sí demasiado perjudicial; son nuestros métodos los que causan los problemas.

En primer lugar, los usuarios no tienen elección sobre si participar o no. Normalmente, cualquier usuario puede ser captado para formar parte de una población experimental; sin embargo, la investigación obligatoria no da cabida al consentimiento informado. Esta decisión reduce claramente la autonomía de las personas y sería una mala ley universal de comportamiento. En segundo lugar, los experimentos son opacos. Las personas no suelen tener forma de saber en qué grupos experimentales están y cuándo terminarán las pruebas. La opacidad tampoco puede ser un principio universal saludable. En tercer lugar, en algunas empresas, el objetivo de la experimentación no es mejorar el producto sino alcanzar unos objetivos: los equipos lanzan diferentes opciones hasta que la gente responde de la forma

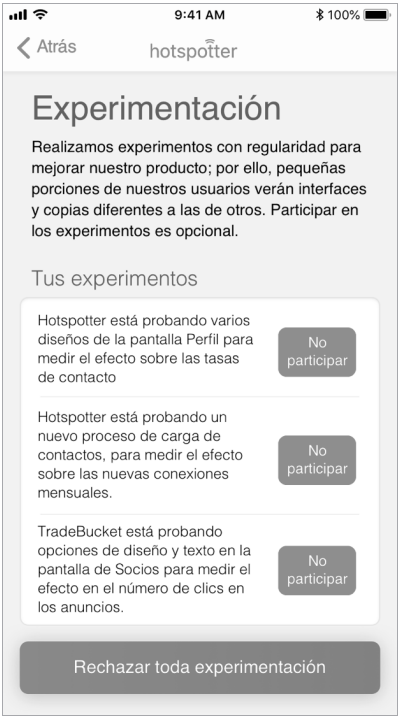
adecuada. Esta es la definición misma de tratar a las personas como medios, no como fines.

¿Podemos diseñar un enfoque más ético y deontológico de la experimentación, el cual recomendaríamos como método universal? Empecemos por considerar siempre a los usuarios como fines, no como medios. Deberíamos comprometernos a intentar mejorar la experiencia del usuario con cada experimento, y negarnos a realizar experimentos que creamos que serán neutrales o perjudiciales. Las empresas a menudo tienen que tomar decisiones que no gustarán a los usuarios, como subir los precios o recortar funcionalidades; según este principio, estos cambios no se prestan a experimentos. Si quieres subir los precios, hazlo de forma generalizada. Comprometerse a mejorar siempre las experiencias de los usuarios solucionaría muchos de los problemas del estudio de contagio emocional de Facebook: los investigadores tendrían que asegurarse de que los participantes sólo vieran las actualizaciones más positivas, no las más negativas.

Reconociendo que la experimentación es investigación, deberíamos considerar inviolable el consentimiento informado. Dado que los niños no pueden dar su consentimiento, deberíamos eliminarlos del grupo de experimentación, a menos que también podamos obtener el consentimiento de los tutores. También deberíamos acordar que los usuarios deberían poder enterarse de los experimentos en los que participan, con alguna pantalla o notificación que describa cada experimento, informe al usuario de quién dirige la investigación y describa el objetivo, redactado quizá en términos de las métricas que estamos siguiendo o las hipótesis que estamos probando.

Inspirémonos también en la teoría de la persuasión y comprometámonos a evitar tanto la emoción negativa como el engaño en nuestros experimentos, y esforcémonos por utilizar formas de persuasión tenues y visibles siempre que sea posible, como los empujoncitos informativos en lugar de los estructurales. Por último, los usuarios deben poder optar por no participar en experimentos individuales y en todo el programa de experimentación por igual, sin efectos negativos. Los usuarios que opten por no participar seguirán recibiendo las actualizaciones del software una vez que se extiendan a todos los usuarios; simplemente no incluiremos a estas personas en los experimentos.

Para una hipotética aplicación de smartphone, una sola pantalla podría satisfacer muchos de estos requisitos:



Este enfoque más ético de la experimentación no sería demasiado oneroso. Tendríamos que ser más rigurosos a la hora de seleccionar las poblaciones de muestra, añadir una nueva pantalla o notificación, proporcionar metadatos breves para cada experimento y crear sistemas de inclusión y exclusión fiables. Estas características no tienen por qué ser muy prominentes en el producto, siempre y cuando estén disponibles y se puedan encontrar. Estos pequeños cambios deberían ayudarnos a tratar a los usuarios con respeto, reducir el riesgo de la ira de las autoridades reguladoras y garantizar que llevamos a cabo los experimentos de una forma que nos gustaría que otros siguieran.

El velo de la ignorancia

Otro principio útil para diseñar sistemas justos es el *velo de ignorancia* de John Rawls. En *Una teoría de la justicia*,[28] Rawls sostiene que la sociedad —para nuestros propósitos extenderemos esto también a la sociedad tecnológica— se estructura mejor como si sus arquitectos no supiesen su papel eventual en el sistema. Bajo un velo de ignorancia, no conoceríamos nuestro estatus social, nuestra inteligencia o incluso nuestros intereses; pero si el sistema es justo, deberíamos estar satisfechos donde quiera que acabásemos.

El velo de la ignorancia tiene algunos vínculos con la deontología; tu no querrías salir de detrás del velo a un lugar en el que sólo eres un medio para los fines de los demás. También es posible que reconozcas algunos paralelismos con la regla de oro. Sin embargo, no se trata sólo de tratar a la gente como te gustaría que te trataran a ti, sino de crear sistemas enteros en los que todo el mundo reciba un trato justo. Es como cortar un pastel y dejar que otro elija una porción, pero extendido a toda una población.

La idea de Rawls, centrada en la igualdad y la justicia redistributiva, atrae críticas de sectores predecibles cuando se aplica a la política, pero para nuestros propósitos es poderosa. El velo de la ignorancia nos obliga a considerar todos los diversos papeles que las personas desempeñarán en nuestros sistemas y cómo nuestro trabajo podría influir en personas de orígenes muy diversos. Aplicado a la persuasión, el velo de la ignorancia sugiere que sólo debemos crear sistemas persuasivos que sean justos tanto para el persuasor como para el persuadido.

Una mejor persuasión

Algunas de nuestras dificultades persuasivas son consecuencia directa de nuestros procesos de desarrollo de productos. Las estrategias que crean productos deseables también fomentan la adicción. Usamos nuestros teléfonos 150 veces al día; ¿es porque están diseñados para captar nuestra atención o porque nos gustan de verdad? Probablemente ambas cosas. Que nos resulte difícil separar estas motivaciones habla de los fallos del movimiento de diseño de experiencias. Los diseñadores no se han interrogado sobre la diferencia entre el uso

placentero y el habitual, y la retórica del diseño para el placer ha contribuido directamente a la adicción.

Algunos comentaristas sostienen que, para contrarrestar la manipulación, las empresas tecnológicas deberían exponer activamente a la gente a puntos de vista contradictorios. La industria tecnológica tiene muchas herramientas a su disposición: complementos (*plug-ins*) de comprobación de hechos y calificaciones de confianza para combatir la desinformación, y *crowdsourcing*, *blockchain* y criptografía para luchar contra el colapso de las pruebas. Pero contrarrestar los sesgos de la gente es un trabajo duro e ingrato. Los primeros esfuerzos por diversificar el entorno de la información han sido exasperantemente burdos: los intentos de Facebook de advertir a los usuarios de contenidos sospechosos en realidad hicieron que más gente hiciera clic en ellos, y sólo recientemente me he librado de un exasperante experimento que añadía la respuesta más atractiva (normalmente la más polémica o de un trol) a cada artículo de *News Feed*.

Sin embargo, ¿no es esto simplemente más intromisión tecnocrática? La idea de que los tecnólogos deben obligar a las masas a seguir una dieta de información equilibrada debería preocuparnos. ¿Qué daño es más grave: la amenaza de manipulación o la amenaza autoritaria de controlar el entorno informativo de los demás? Estas cuestiones son sociales, políticas y jurídicas tanto como técnicas; como tales, no nos corresponde responderlas a nosotros solos. Las soluciones técnicas que serían más eficaces contra la desinformación, como las políticas de nombres reales o un mejor seguimiento de las fuentes, pondrían en peligro por sí mismas la privacidad. Quizá nuestro deber más importante sea estimular el debate público sobre las tecnologías persuasivas. La industria tecnológica debería tratar de impulsar la alfabetización informativa en todos los niveles de la educación y la vida adulta, y desempeñar un papel activo en la restauración de una prensa próspera y resistente. Puede que los tecnólogos tengan que dotar a los usuarios de estrategias contra la adicción y la persuasión, o incluso construir tecnologías alternativas que se pongan del lado del usuario en contra de la propia industria, como los bloqueadores de persuasión que eliminan la publicidad manipuladora y sacan a la gente de las pruebas A/B no consentidas.

También debemos eliminar los factores que han causado nuestros problemas de persuasión. En el corazón del *dark pattern*, de la aplica-

ción adictiva y del problema de la desinformación se encuentra una fijación indebida en la cuantificación y la participación. Elegir nuevas métricas de éxito allanaría el camino hacia una persuasión más ética. El movimiento *Time Well Spent*[29] se pregunta cómo sería la tecnología si se diseñara para respetar los valores humanos en lugar de para captar la atención. El movimiento recurre a las teorías de la tecnología calmada y el *mindfulness* (atención plena) para inspirar a los diseñadores a proteger el tiempo y la agencia de los usuarios, y reclama nuevos modelos de negocio que subviertan la economía de la atención.

Los datos cuantitativos deben ir siempre emparejados con una investigación cualitativa accesible para que las historias humanas puedan reclamar el lugar que les corresponde en la mente de los responsables de la toma de decisiones. También podemos seleccionar *objetivos mutuamente destructivos*, métricas elegidas en parejas de forma que una sufra si nos limitamos a jugar con la otra. Por ejemplo, los *dark patterns* pueden generar más ingresos por usuario, pero también perjudican la retención si los usuarios se sienten engañados. Elegir tanto los ingresos como la retención como objetivos mutuamente destructivos proporciona una pequeña salvaguarda contra el abuso; si ambas medidas se mueven en la dirección correcta, podemos estar seguros de que las cosas están mejorando de verdad.

Regulación y exclusión voluntaria

Si la industria no se autoregula, debería prepararse para el rechazo de los consumidores. Hasta hace poco, la sociedad veía a los *refuseniks* tecnológicos como socialmente marginales, y eran sobre todo los propios tecnólogos los que optaban por abstenerse, borrando sus aplicaciones, escalando montañas y escribiendo artículos de opinión sobre sus experiencias. Estos esfuerzos rezumaban privilegio —después de todo, hay que ser rico para no necesitar nada—, pero en medio de la creciente preocupación por las tecnologías adictivas, se está gestando un movimiento de moderación pública. Las clínicas ya están tratando a los autodenominados adictos a las aplicaciones; quizá surja una industria de la desintoxicación como servicio: entréguenos sus dispositivos y le aislaremos durante dos semanas.

Donde vayan los consumidores, irán los reguladores. Las empresas

tecnológicas ya han sido demandadas y citadas por persuasión desleal y *dark patterns*. En 2015, LinkedIn pagó 13 millones de dólares para resolver una demanda colectiva por *dark patterns*. Es probable que la regulación venga primero de la UE,[30] dada su oposición histórica a los monopolios tecnológicos y la sensibilidad de sus ciudadanos ante los abusos de las empresas. El gobierno alemán redactó una ley que impondría multas de 50 millones de euros a las redes sociales que no pongan freno a la incitación al odio y la desinformación. Los reguladores podrían decidir responsabilizar a las plataformas de la incitación al odio, obligar a los conglomerados tecnológicos a dividirse o exigir que las redes sociales permitan a los usuarios llevar sus redes de amigos a servicios de la competencia. Los anuncios en línea podrían, y podría decirse que deberían, estar obligados a revelar sus financiadores. Algunos filósofos y juristas están incluso debatiendo si debería existir un derecho legal consagrado a la protección de la atención.

Los inicios de la era televisiva también suscitaron preocupación por la persuasión y la desinformación. Muchos gobiernos respondieron estableciendo agencias y normas nacionales de radiodifusión, creando una fuerte influencia de arriba abajo sobre la floreciente industria. Si este patrón se repite en el caso de las nuevas tecnologías persuasivas, la industria sólo podrá culparse a sí misma.

NOTAS

1. Problemas en el paraíso

1. Es la famosa directiva de Mark Zuckerberg a sus equipos y desarrolladores: "move fast and break things" (muévete rápido y rompe cosas) para incitar a sus equipos a innovar. Según él, si no rompes cosas, no te estás moviendo suficientemente rápido. Véase Henry Blodget, "Mark Zuckerberg On Innovation", *Business Insider*, 1 Oct 2009.
2. Richard Sennett, *El Artesano* (Anagrama, 2009).
3. 2017 Cone Communications CSR Study, *conecomm.com*.
4. Véase Bruno Latour, *La esperanza de Pandora: Ensayos sobre la realidad de los estudios de la ciencia* (Gedisa, 1999) para un análisis sobre esta cuestión.
5. Peter-Paul Verbeek, *Moralizing Technology* (University of Chicago Press, 2011). Verbeek, a su vez, se basa en el trabajo de Don Ihde y Latour.
6. Melvin Kranzberg, "Software for Human Hardware?", in Pranas Zunde & Dan Hocking (eds.), *Empirical Foundations of Information and Software Science V* (Plenum Press 1990). NT: En castellano se puede consultar la entrada "Neutralidad tecnológica" en la *Wikipedia* que presenta las "Las Leyes de Kranzberg" *https://es.wikipedia.org/wiki/Neutralidad_tecnol%C3%B3gica*
7. Una de las seductoras citas de dudoso origen, atribuida a Maxim Gorky, usadas en la película de *Le petit soldat* de Jean-Luc Godard's film, atribuida a Lenin.
8. Caroline Whitbeck, *Ethics in Engineering Practice and Research* (Cambridge University Press, 2nd ed., 2011).

2. ¿No causar daños?

1. Por lo menos así lo dice el economista Horst Siebert. Es posible que sea un cuento apócrifo, pero la anécdota contiene una verdad indiscutible: los planes no siempre salen como estaba previsto.
2. 2 Paul Virilio, *El cibermundo, la política de lo peor* (Cátedra, 1997).
3. Don Ihde, *Technology and the Lifeworld* (Indiana University Press, 1990).
4. Shannon Hall, "Exxon Knew about Climate Change almost 40 years ago", *Scientific American*, 26 Oct 2015, *scientificamerican.com*.
5. Ver Thomas Wendt, "Decentering Design or a Critique of Human-Centered Design", *slideshare.net*.
6. Ben Thompson, "Airbnb Versus Hotels", *Stratechery*, 18 Abr 2017, *stratechery.com*.
7. Ursula Franklin, *The Real World of Technology* (House of Anansi Press, 2nd ed., 1999).
8. David Ingold y Spencer Soper, "Amazon Doesn't Consider the Race of Its Customers. Should It?", *Bloomberg*, 21 Abr 2016, *bloomberg.com*.
9. Joanna Bryson, "Three very different sources of bias in AI, and how to fix them", 13 Jul 2017, *joanna-bryson.blogspot.com*.
10. Aylin Caliskan, Joanna Bryson, Arvind Narayanan, "Semantics derived automatically from language corpora contain human-like biases", *Science*, 14 Abr 2017, 183–

186.
11. Chris Ip, "In 2017, society started taking AI bias seriously", *Engadget*, 21 Dic 2017, *engadget.com*.
12. Andrew Thompson, "Google's Sentiment Analyzer Thinks Being Gay Is Bad", *Vice Motherboard*, 25 Oct 2017, *vice.com*.
13. Laura Hudson, "Technology Is Biased Too. How Do We Fix It?", *FiveThirtyEight*, 20 Jul 2017, *fivethirtyeight.com*.
14. Stella Lowry y Gordon Macpherson, "A blot on the profession", *British Medical Journal* Vol. 296, 5 Mar 1988.
15. Christian Rudder, "Race and Attraction, 2009–2014", *OkCupid*, 10 Sep 2014, *okcupid.com*.
16. Lizzie Edmonds, "Google forced to remove vile racist search suggestions from its site for a number of British cities including Bradford, Leicester and Birmingham", *MailOnline*, 11 Feb 2014, *dailymail.co.uk*.
17. "I think it's time we broke for lunch…", *The Economist*, 14 Abr 2011, *economist.com*.
18. Carlota Perez, *Revoluciones tecnológicas y capital financiero: La dinámica de las grandes burbujas financieras y las épocas de bonanza* (Siglo XXI, 2004).
19. *Terminator 2: El juicio final*, dir. James Cameron (TriStar Pictures, 1991). Podrías argumentar que no eran las palabras de Sarah sino las que dice John, que le llegaron vía Kyle como palabras de Sarah.
20. Concebido originalmente por el estratega militar Charles Taylor y adaptado desde entonces por varios futuristas, entre ellos Joseph Voros.
21. Para leer más sobre diseño de futuros: *blogs.uoc.edu/comunicacio/es/diseno-de-futuros*.
22. Octavia Butler, "A Few Rules for Predicting the Future", *Essence Magazine*, 2000.
23. Shannon Vallor, *Technology and the Virtues* (Oxford University Press, 2016).
24. Genevieve Bell, "Rage Against the Machine?", presentación en la conferencia *Interaction12*.
25. Hace referencia al juego *"Flip-it", gamestorming.com/flip-it*.
26. Eric Meyer and Sara Wachter-Boettcher, *Design for Real Life* (A Book Apart, 2016).
27. El nombre de esta idea lo he elegido yo, pero la idea original es de Sam Jeffers.
28. Cameron Tonkinwise, 'Ethics by Design, or the Ethos of Things', *Design Philosophy Papers*, 2:2, 129-144, 2004.
29. Jared Spool, 'Creating Great Design Principles: 6 Counter-intuitive Tests', *UIE*, 1 Mar 2011, uie.com.
30. Anand Giridharadas, "A tale of two Americas. And the mini-mart where they collided", presentación en la conferencia *TED* 2015, *ted.com*.
31. Shannon Vallor, 'An Introduction to Data Ethics: a resource for data science courses', *Markkula Center for Applied Ethics*, scu.edu.

3. Mecanismos de persuasión

1. Langdon Winner, "Do Artifacts Have Politics?", *Daedalus* 109, no. 1 (1980): 121-36.
2. Daniel Berdichevsky y Erik Neuenschwander, "Toward an ethics of persuasive technology", *Communications of the ACM*, 42, 5 (May 1999), 51–58.
3. Acuñado por Harry Brignull; véase *deceptive.design*.
4. Tim Wu, *The Attention Merchants: The Epic Scramble to Get Inside Our Heads* (Knopf Publishing Group, 2016).
5. Peter Kafka, "Amazon? HBO? Netflix thinks its real competitor is... sleep", *Recode*, 17 Abr 2017, recode.net.

6. Kleiner Perkins Internet Trends 2017, kpcb.com/internet-trends.
7. Ian Bogost, "The App That Does Nothing", *The Atlantic*, 9 Jun 2017, *theatlantic.com*. Bogost es un teórico de los juegos persuasivos y diseñador de *Cow Clicker*, un notorio comentario sobre los aspectos manipuladores de *FarmVille*. Los jugadores de *Cow Clicker* sólo tenían un objetivo: hacer clic en una vaca. El juego acumuló 50.000 usuarios antes de que Bogost desencadenara un "Cowpocalypse", desvaneciendo las vacas en un irónico rapto.
8. Adam Cramer, Jamie Guillory, Jeffrey Hancock, "Experimental evidence of massive-scale emotional contagion through social networks", *Proceedings of the National Academy of Sciences (PNAS)*, 17 Jun 2014 vol. 111 no. 24 8,788–8,790.
9. Ariel Rubinstein y Ayala Arad, "The People's Perspective on Libertarian-Paternalistic Policies" (2015).
10. Molly Sauter, "The Apophenic Machine", *Real Life Magazine*, *reallifemag.com*.
11. Chris Wade, "The Reddit Reckoning", *Slate*, 15 Abr 2014, slate.com.
12. Una notable inconsistencia descubierta por la periodista Olivia Solon.
13. Berit Anderson y Brett Horvath, 'The Rise of the Weaponized AI Propaganda Machine', scout.ai.
14. Philip Howard, y Bence Kollanyi, "Bots, #Strongerin, and #Brexit: Computational Propaganda during the UK-EU Referendum." Documento de trabajo 2016.1. Oxford, *UK: Project on Computational Propaganda*.
15. Jeff Bezos, "2015 Letter to Shareholders".
16. Paul Robinette et al., "Overtrust of robots in emergency evacuation scenarios", *2016 11th ACM/IEEE International Conference on Human-Robot Interaction (HRI)*, Christchurch, 2016, pp. 101-108.
17. Sandra Matz et al., "Psychological targeting as an effective approach to digital mass persuasion", *Proceedings of the National Academy of Sciences (PNAS)* 2017.
18. Dana Mattioli, "On Orbitz, Mac Users Steered to Pricier Hotels", *The Wall Street Journal*, 23 Ago 2012, wsj.com.
19. John D. Sutter, "Amazon seller lists book at $23,698,655.93 – plus shipping", *CNN*, 25 Abr 2011, cnn.com.
20. Lyrebird, "With great innovation comes great responsibility", lyrebird.ai/ethics.
21. Daniel Berdichevsky y Erik Neuenschwander, "Toward an ethics of persuasive technology", *Communications of the ACM*, 42, 5 (May 1999), 51–58.
22. BJ Fogg, *Persuasive Technology* (Morgan Kaufman, 2003).
23. Richard Thaler, "The Power of Nudges, for Good and Bad", *New York Times*, 31 Oct 2015, nytimes.com.
24. Nynke Tromp et al., "Design for Socially Responsible Behavior", *Design Issues*, Volumen 27, Número 3, Verano de 2011.
25. Luciano Floridi, "Tolerant Paternalism: Pro-ethical Design as a Resolution of the Dilemma of Toleration", *Science and Engineering Ethics* (2016), 22: 1669.
26. Su primera formulación del "imperativo categórico", de *Groundwork for the Metaphysics of Morals*.
27. También del imperativo categórico, esta vez la segunda formulación.
28. John Rawls, *A Theory of Justice* (Harvard University Press, 1971).
29. Ahora está gestionado por el *Center for Humane Technology*, *humanetech.com*.
30. Tras la adopción del *Paquete de Servicios Digitales* en primera lectura por el Parlamento Europeo en julio de 2022, tanto la *Ley de Servicios Digitales* como la *Ley de Mercados Digitales* han sido adoptadas por el Consejo de la Unión Europea, firmadas por los Presidentes de ambas instituciones y publicadas en el *Diario Oficial*.

SOBRE EL AUTOR

Cennydd Bowles es un diseñador y escritor londinense con quince años de experiencia y clientes como Twitter, Ford, Cisco y la BBC. Actualmente se interesa por la ética de la tecnología emergente. Ha dado conferencias sobre este tema en la Universidad Carnegie Mellon, Google y la Escuela de Artes Visuales de Nueva York, y es un ponente muy solicitado en eventos sobre tecnología y diseño en todo el mundo.

También por Cennydd: *Undercover User Experience Design* (New Riders 2010), con James Box.