

Autores

Julian Nida-Rümelin
Nathalie Weidenfeld

Editoras de la traducción

Rocío Rueda Ortiz
Claudia Roza Sandoval

Humanismo digital

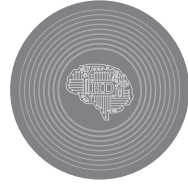
Por una transformación humana de la democracia, la economía y la cultura en la era digital

Julian Nida-Rümelin y Nathalie Weidenfeld
Autores

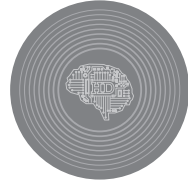
Profesores y estudiantes DIE-UPN
Traductores



UNIVERSIDAD
PEDAGÓGICA
NACIONAL



Humanismo digital



Humanismo digital

**Por una transformación
humana de la democracia,
la economía y la cultura
en la era digital**

Julian Nida-Rümelin

Nathalie Weidenfeld

Autores



Catalogación en la fuente - Biblioteca Central de la Universidad Pedagógica Nacional

Nida-Rümelin, Julian

Humanismo digital. Por una transformación humana de la democracia, la economía y la cultura en la era digital / Julian Nida-Rümelin, Nathalie Weidenfeld - Primera edición en español. - Bogotá: Universidad Pedagógica Nacional, 2026.

246 páginas. (Colección Comunicación, Tecnología y Culturas en Educación).

Incluye: Referencias bibliográficas.

Epílogo

1. Era Digital. 2. Innovación Educativa. 3. Humanismo. 4. Inteligencia Artificial. 5. Tecnologías de la Información - Aspectos Sociales. 6. Cultura Digital. 7. Educación Digital. 9. Economía y Cultura. 10. Educación y Democracia. I. Weidenfeld, Nathalie. II. Tit.

303.483 23. edc.

Colección

Comunicación, Tecnología y Culturas en Educación

Humanismo digital.

Por una transformación humana de la democracia,
la economía y la cultura en la era digital¹

Autores

Julian Nida-Rümelin

Nathalie Weidenfeld

Editoras de la traducción

Rocío Rueda Ortiz

Claudia Rozo Sandoval

© Universidad Pedagógica Nacional

ISBN impreso: 978-628-7851-74-0

ISBN PDF: 978-628-7851-79-5

ISBN ePub: 978-628-7851-75-7

Primera edición, 2026

Helberth Augusto Choachí González

Rector

Paola Helena Acosta Sierra

**Vicerrectora de Investigación, Extensión
y Proyección Social**

Víctor Espinosa Galán

Vicerrector Académico

Yaneth Romero Coca

Vicerrectora Administrativa y Financiera

Gina Marcela Duarte Fonseca

Secretaria General

Preparación editorial

Grupo Interno de Trabajo Editorial

Universidad Pedagógica Nacional

Calle 72 N° 12-77. Tercer piso, Edificio Administrativo
editorial.upn.edu.co

Teléfono: (601) 347 1190 - (601) 594 1894

Bogotá, Colombia

Alba Lucía Bernal Cerquera

Coordinación

María Alejandra Uribe C.

Edición

Fernando Carretero

Corrección de estilo

Fredy Johan Espitia B.

Diagramación y cubierta

Carvajal Soluciones de Comunicaciones S. A. S.

Impresión

Esta publicación puede ser distribuida, copiada y exhibida por terceros si se mencionan los créditos correspondientes. No se puede obtener ningún beneficio comercial. No se pueden realizar obras derivadas.

Esta obra se publica con fines académicos, culturales y de enseñanza.

1 El título original en alemán es: *Digitaler Humanismus. Eine Ethik für das Zeitalter der Künstlichen Intelligenz*. Traducido al inglés como: *Digital humanism. For a human transformation of democracy, economy and culture in the digital age*.

Contenido

| | |
|---|----|
| Prefacio a la versión en inglés | 11 |
| <i>Traductor, Juan C. Amador</i> | |
| <hr/> | |
| Presentación. El humanismo digital y los estudios ciberculturales críticos | 13 |
| <i>Autoras, Rocío Rueda Ortiz y Claudia Roza Sandoval</i> | |
| <hr/> | |
| Prólogo de la versión en español | 35 |
| <i>Autor, Juan Carlos Amador</i> | |
| <hr/> | |
| 1. Introducción | 51 |
| <i>Traductora y comentarista, Rocío Rueda Ortiz</i> | |
| <hr/> | |
| 2. Los robots como los nuevos esclavos (digitales) | 61 |
| <i>Traductora y comentarista, Jenny Catalina Loaiza-Fuquen</i> | |
| <hr/> | |
| 3. Estimulación digital de las emociones | 69 |
| <i>Traductor y comentarista, John Fernando Fajardo Velasco</i> | |
| <hr/> | |
| 4. El problema de la autonomía y la determinación en el mundo digital | 81 |
| <i>Traductor y comentarista, Óscar Alejandro Barrios Candil</i> | |
| <hr/> | |
| 5. El mundo como la perfecta máquina del universo | 91 |
| <i>Traductor, John Freddy Sarmiento Vela</i> | |
| <hr/> | |
| 6. Optimización digital, utilitarismo e Inteligencia Artificial | 99 |
| <i>Traductora y comentarista, Jenny Catalina Loaiza-Fuquen</i> | |

| | |
|---|-----|
| 7. La racionalidad económica como un programa de computadora | 105 |
| <i>Traductor y comentarista, Óscar Alejandro Barrios Candil</i> | |
| 8. Por qué los robots no tienen juicio moral | 117 |
| <i>Traductora y comentarista, Andrea Nassar</i> | |
| 9. Sin comparabilidad ética | 125 |
| <i>Traductora y comentarista, Andrea Nassar</i> | |
| 10. Por qué la inteligencia artificial falla en los dilemas morales | 135 |
| <i>Traductora y comentarista, Angela Salamanca</i> | |
| 11. Por qué las inteligencias artificiales no pueden pensar | 141 |
| <i>Traductor y comentarista, Juan Camilo Cardenal</i> | |
| 12. Virtualidades digitales y realidades sobrias | 155 |
| <i>Traductor y comentarista, Juan Camilo Cardenal</i> | |
| 13. Sobre la ética de la comunicación en internet | 161 |
| <i>Traductor y comentarista, Carlos López</i> | |
| 14. Sobre la ética de la comunicación entre humanos e inteligencia artificial | 169 |
| <i>Traductor y comentarista, Lizeth Marcela Díaz</i> | |
| 15. Aspectos culturales de la digitalización | 177 |
| <i>Traductora y comentarista, Zully Octavia Farfán Romero</i> | |
| 16. La educación digital | 187 |
| <i>Traductor y comentarista, Óscar Cárdenas</i> | |
| 17. Utopía de una democracia líquida | 201 |
| <i>Traductor y comentarista, Óscar Ardila</i> | |

| | |
|---|-----|
| 18. Aspectos socioeconómicos de la digitalización | 213 |
| <i>Traductora y comentarista, Carlos López</i> | |
| <hr/> | |
| 19. Tentaciones transhumanistas | 223 |
| <i>Traductora y comentarista, Claudia Rozo Sandoval</i> | |
| <hr/> | |
| 20. Sobre la metafísica de la digitalización | 239 |
| <i>Traductora y comentarista, Claudia Rozo Sandoval</i> | |
| <hr/> | |
| 21. Epílogo | 239 |
| <i>Traductora, Rocío Rueda Ortiz</i> | |
| <hr/> | |
| Referencias | 243 |

Prefacio a la versión en inglés*

Traducido por Juan Carlos Amador

Parece que existen básicamente dos estrategias de transformación digital. La primera es gestionada por el mercado y orientada al consumidor. Los modelos de negocio predominantes se basan en el *marketing*. La estrategia está guiada por la idea de hacer del mundo un lugar mejor que ofrezca aplicaciones de comunicación e interacción. La segunda estrategia se centra en el interés del Estado, que controla a los ciudadanos y estabiliza el poder autocrático.

En cada estrategia, una de las dos superpotencias está a la cabeza: Estados Unidos y China. Pero posiblemente exista una tercera estrategia, opuesta tanto al modelo comercial como al estatal. Algunos documentos y la decisión de la Unión Europea van en esta dirección: la inteligencia artificial (IA) centrada en el ser humano. Hasta ahora no existe una idea clara de esta tercera, una estrategia más humana de transformación digital.

En este libro desarrollamos las ideas básicas de una visión diferente, que denominamos *humanismo digital*.

Julian ha utilizado este término durante muchos años en conversaciones y debates, lo que ha generado polémica y también ha despertado interés en el ámbito público y político. Por esto, decidimos escribir un libro juntos con el fin de exponer sus ideas básicas.

* *Digital humanism. For a human transformation of democracy, economy and culture in the digital age.*

Nosotros: Julian Nida-Rümelin, filósofo y exministro de Estado para la cultura, y Nathalie Weidenfeld, una teórica del cine.

La versión alemana del libro apareció en 2018 y recibió el premio Bruno Kreisky al mejor libro político de 2018.

Mientras tanto, el término *humanismo digital* se utiliza ampliamente, sobre todo en Austria, Italia y España. En Austria, el grupo del profesor Hannes Werthner ha creado la Iniciativa de Humanismo Digital, que reúne a reconocidos investigadores internacionales y ofrece una impresionante serie de conferencias, así como importantes publicaciones. En España, el gobierno apoya diferentes organizaciones para ayudar a los emprendedores a comprender y apoyar una transformación digital “centrada en el ser humano”. En Italia, Francesca Bria, presidenta del Fondo Italiano de Innovación, habla del humanismo digital como una alternativa al *capitalismo digital*. En Italia, el término *humanismo digital* se ha vuelto muy popular, debido a que este libro apareció en italiano en 2019.

La UE ha presentado una visión de transformación digital que debe estar “al servicio de los humanos”. En Estados Unidos, el término está empezando a ganar importancia, sobre todo en el contexto de cómo la economía puede utilizar y beneficiarse de una transformación digital que “se centra en la experiencia humana”. Sin embargo, Tristan Harris, expleado de Google, se centra más en los efectos negativos de una transformación digital orientada económicamente y promueve una “tecnología humana” basada en la confianza.

En este libro, describimos los aspectos filosóficos y culturales del humanismo digital. Puede entenderse como su base.

Julian Nida-Rümelin
Nathalie Weidenfeld
Múnich, Alemania

Presentación. El humanismo digital y los estudios ciberculturales críticos

Por Rocío Rueda Ortiz y Claudia Rozo Sandoval

La traducción de este libro de Julian Nida-Rümelin y Nathalie Weidenfeld se produjo en el marco de las actividades académicas del seminario “Estudios ciberculturales críticos, subjetividad y educación”, del Doctorado Interinstitucional en Educación de la Universidad Pedagógica Nacional de Colombia, en noviembre de 2022. Escogimos este libro entre la literatura del seminario para hacer una traducción al castellano, con un aporte del estudiante y traductor a la traducción del inglés, como un valor agregado.

En este espacio nos propusimos retomar, por una parte, algunas reflexiones derivadas de la línea de investigación Cibercultura y Educación,¹ del Grupo Educación y Cultura Política de ese programa doctoral, coordinada en aquel entonces por Rocío Rueda Ortiz, alrededor de la formación como efecto de procesos de subjetivación individual y colectiva en contextos de apropiación sociotécnica. Y, por otra parte,

1 En ella, se analizan las continuidades y transformaciones en ámbitos como: creación, circulación y legitimación de información y conocimientos dentro y fuera de la institución escolar; las formas de socialización, comunicación y construcción de identidades individuales y colectivas; los modos de participación y ejercicio ciudadano y las diversas producciones estéticas.

los trabajos adelantados desde la perspectiva intercultural que viene trabajando Claudia Rozo Sandoval en la línea “Educación e interculturalidad”, en el grupo de investigación Equidad y Diversidad en Educación de la Maestría y el Doctorado en Educación, así como los trabajos sobre las relaciones entre cultura, tecnología y conocimiento, en otros espacios académicos. Desde estos dos campos de acción advertimos la importancia de profundizar en una perspectiva crítica que dé cuenta de las relaciones de poder que se están produciendo del lado de dichas transformaciones sociotécnicas. En particular la tendencia dominante de “plataformización” del mundo social agenciado por multinacionales de *software* y *hardware*, y sus correspondientes intereses capitalistas. Esto nos ha llevado a diálogos con la economía política, para complejizar nuestra comprensión de los actuales cambios culturales, así como con los avances tecnocientíficos que asocian desarrollos en las ciencias de la vida, con la manipulación de grandes volúmenes de información mediante tecnologías de información de base computacional que, cada vez más, implican nuestras formas de ser y estar en el mundo.

De hecho, en una tercera etapa de los estudios ciberculturales, aproximadamente a partir del 2010, las investigaciones evidencian unas problemáticas singulares en el sur y norte globales (Rueda y Uribe, 2021). Se destaca la necesidad de incluir dimensiones intergeneracionales, de raza, género, región y clase social para analizar las maneras diversas y desiguales de inclusión y apropiación de las tecnologías. También se encuentran los trabajos sobre la *ética hacker*, como movimiento de resistencia a tendencias dominantes; los *commons* y el *commoning* como alternativas para producir y compartir conocimientos e información en medios digitales que se resisten a modelos capitalistas, así como propuestas educativas ciudadanas desde la discusión decolonial, feminista e intercultural. Desde estas últimas se plantean problemáticas sobre el reconocimiento de saberes, cosmovisiones y modos de vida indígenas, campesinos y afrodescendientes que han apropiado las tecnologías de modos diversos y con diferentes propósitos culturales y políticos en la región. En el contexto latinoamericano, se trata ante todo de un

campo en constante cambio, no unificado y sin metodologías establecidas, pero que ofrece la renovación de las áreas del conocimiento, en términos ontológicos, epistemológicos y metodológicos (Rueda y Uribe, 2021).

Ahora bien, el horizonte filosófico que nos ha acompañado en estas reflexiones ha vuelto siempre a la pregunta por la técnica (*Die Frage nach der Technik*) que en 1953 planteara Heidegger. Su crítica a la desmesura y expansión del proyecto tecnocientífico de la modernidad la podemos trasladar al momento actual del capitalismo que se expresa en un nuevo biopoder caracterizado por nuevas manifestaciones de colonización, (auto)explotación de la naturaleza y de la humanidad, en diversas injusticias sociales, epistémicas y culturales. Si la esencia de la técnica, como afirmó Heidegger, no está en la técnica, sino que siempre nos lleva a una pregunta por nosotros: ¿cómo podemos dar cuenta de ella desde el actual estado del planeta? Y, ¿esto qué tiene que ver con los procesos de formación y el proyecto humanizador de la pedagogía? ¿Es posible hablar de un tipo de crítica singular en el contexto latinoamericano?

La importancia de las perspectivas críticas

Por una parte, en el seminario abordamos diferentes perspectivas críticas tanto europeas como latinoamericanas que nos llevaron a repensar este campo de los estudios ciberculturales: (a) la crítica desde teorías posestructurales de la diferencia, que ven en las tecnologías y sus narrativas una vía para deconstruir el proyecto moderno centrado en la escritura y la razón (Landow, 1995; Martín-Barbero, 2003), y (b) los enfoques que ven la cultura como un campo de lucha y poder, en lo que se conoce como *perspectivas decoloniales* (Sierra y Maldonado, 2016; Walsh, 2013; Castro-Gómez, 2007) y *comunitarias* (Torres, 2008). Por otra parte, discutimos los enfoques que cuestionan la pérdida de “humanidad” a través,

por ejemplo, de la conexión automatizada (véase Stiegler, 1998; Han, 2014), o por la pérdida de soberanía y gobierno sobre las tecnologías (Van Dijck, 2016). Asimismo, analizamos la crítica a la crítica posestructuralista, desde la que se discute la sobrevaloración del pensamiento narrativo sobre el paradigmático en las denominadas *culturas de la oralidad secundaria o hipermediales*, al equiparar dos formas de conocimiento diferentes e inconmensurables, con desconocimiento de una continuidad del proyecto de la cultura alfabética a la que, por cierto, en teoría, se le contraponen (Narváez, 2019).

En estos debates encontramos que tanto la literatura académica como de ficción, sobre cibercultura, se dividen entre perspectivas utópicas y distópicas. Desde los primeros desarrollos de la radiodifusión y el cine, el pensamiento crítico ha tenido una postura ambivalente ante los medios electrónicos. De hecho, en el periodo de entreguerras, Walter Benjamin y Theodor Adorno representaron dos posturas críticas diferentes de la difusión de los medios de comunicación de masas. Benjamin (1989) intuye que la reproducción técnica de los mensajes crea condiciones completamente nuevas de percepción estética y comunicación, mientras que Adorno (2004) ve en la comunicación de masas un declive del aura artística y cultural. En tanto que Gilbert Simondon (2007, 2009), desde la crítica a la filosofía tradicional, analiza la relación del ser desde la individuación, con la constitución de lo vivo (el mundo natural) en articulación con la dimensión de la técnica y los objetos técnicos constitutivos de los contextos humano y natural, lo cual implica un nivel consciencia sobre el mundo de sistemas técnicos que estamos construyendo.

La utopía tecnicista que está en la base de la crítica que plantean Julian Nida-Rümelin y Nathalie Weidenfeld en su libro, nos alerta sobre cómo las tecnologías digitales, por una parte, están perturbando las sociedades y cuestionando nuestra comprensión de lo que significa ser humano y su superación y, por otra, plantea el reto de construir una nueva ética que favorezca sociedades justas y democráticas con los seres humanos en el centro del desarrollo tecnológico. Históricamente, se puede rastrear esta utopía tecnicista

con un doble origen: denuncia y anuncio de la catástrofe y vía de esperanza sobre futuros alternativos. De hecho, en el campo de los estudios sobre cibercultura y tecnociencia, las actuales tecnologías digitales han ocupado el imaginario del “futuro”, por ese carácter “futuro-presente”, instantáneo y ubicuo de los flujos constantes de información y comunicación. Por tanto, para algunos autores, no necesitamos mirar al futuro porque ya está aquí en una forma o visión perversamente idealizada del capitalismo del presente, cuando somos sujetos hedonistas a través de los circuitos del capitalismo de consumo actual que se despliega en redes sociotécnicas.

El campo de las educaciones populares, las pedagogías críticas y las pedagogías interculturales en América Latina, propende por el carácter justo y democrático del desarrollo de las tecnologías y llama la atención sobre las implicaciones que abre la incorporación de saberes subalternizados en el campo informacional y comunicacional; asimismo, advierte sobre los niveles de discriminación, presentes en los sistemas de configuración de algoritmos, que reproducen estructuras sociales con esos sesgos. Así, la presencia de formas de conocimiento y representación del mundo social que han sido históricamente deslegitimadas por el discurso oficial de la modernidad/colonialidad y ahora en los discursos de la denominada *sociedad de la información y el conocimiento*, retoma la pregunta por la “diferencia”, el (des)encuentro con ese “otro”, “lo otro” de la modernidad, y nos vuelve a una pregunta cultural y educativa por la manera como se tramita dicha diferencia y lo que ella implica en procesos de subjetivación. De esta forma, la cultura deviene en un escenario de batalla por el reconocimiento de saberes y prácticas sociales de grupos y organizaciones diversas, así como en su estrecha relación con la economía y la política.

En el campo de la comunicación, de hecho, ha tomado fuerza en América Latina en la última década la perspectiva desarrollada desde el Centro Internacional de Estudios Superiores de Comunicación para América Latina (Ciespal), en el que se propone una lectura de los procesos políticos, sociales y culturales anclados a la comunicación desde una compleja relación entre colonialidad y

decolonialidad; en este sentido, se advierte la importancia de analizar la relación tecnocomunicativa articulada a las culturas, como posibilidad de valorar los universos de significación y prácticas que se producen en estas intersecciones, para conocer la dimensión del conocimiento implicado y localizado geopolíticamente (Rozo-Sandoval, 2016). De hecho, se habla de una *comunicación otra*, en virtud de un pensamiento y una praxis otra y de una dinámica liberadora de las comunalidades y socialidades en diálogo con los saberes originarios (Sierra y Maldonado, 2016). No obstante, es preciso indicar que, si bien estas propuestas se han visto como novedosas, también se ha empezado a percibir cierto vaciamiento de sentido y despojo de su raíz ética y existencial, como lo plantea Walsh (2013, p. 48) al referir las significaciones de algunas apuestas interculturales que “están llegando a ser algo funcional a la economía del conocimiento”.

Por su parte, Nida-Rümelin y Weidenfeld (2022) observan críticamente el asunto de la política y particularmente el problema de una democracia destruida o disfuncional que aparece en muchas películas de ciencia ficción, lo que evidencia el miedo secreto que muchos ciudadanos tienen al fracaso del sistema democrático. Estos temores han aumentado en la actualidad. Desde imágenes procedentes de las industrias audiovisuales, como *Blade Runner*, *RoboCop*, *The Matrix*, *Terminator*, *Black Mirror*, *IA*, etcétera, todas recrean formas de conexión y a la vez protección de los cuerpos humanos, mientras algo externo ocupa el lugar humano para potenciarlo o avasallarlo. De hecho, parece haber un acuerdo generalizado en que la distopía, la disrupción y la desilusión son los temas centrales. Los algoritmos, que se han convertido en metáfora de poderes desconocidos, parecen especialmente amenazadores y no comparables con otros del “mundo real”.

En consecuencia, en el actual desarrollo de la sociedad digitalizada, son muchos los interrogantes sobre la esperanza de democratización real, el bienestar colectivo y la participación. Hay varias señales de alerta, como plantea Zuboff (2015) por ejemplo, sobre la recogida y almacenamiento incontrolado de

datos personales; la reproducción de sesgos y prácticas de discriminación; la manipulación de estos por parte de corporaciones de internet o agencias estatales; la vigilancia, censura y coacción estatal; la criminalidad; la influencia de inteligencias artificiales en la toma de decisiones en las que tenemos cada vez menos posibilidad de decisión, entre otros.

Ahora bien, hay un elemento común en estas distopías y ficciones: la insuficiencia de las perspectivas binarias para entender la relación entre tecnologías, sociedad y cultura y, por tanto, la necesidad de visiones que contribuyan a complejizar estas relaciones y sus componentes. De hecho, la complejidad es la característica más relevante. La razón y el orden que hemos conocido en la modernidad se ven ahora trastocados frente a la heterogeneidad y el caos. Las tecnologías digitales parecen ofrecer una forma de romper con la “totalidad” de la razón y el orden moderno, mediante mezclas inesperadas de humanos y no humanos y al mismo tiempo parece que volviéramos a una totalización racionalista o a una suerte de positivismo de segundo grado basado en una epistemología de los datos (Ricaurte, 2019) en la que el tratamiento de estos se basa en algoritmos, mediante estadísticas avanzadas y modelos de predicción para la toma de decisiones, acciones y relaciones. Ya no se trata de tecnologías como extensiones del cuerpo, o del sistema nervioso como proponía McLuhan (1996). Nos encontramos en un nuevo escenario: una cultura de transmisiones codificadas, una socialidad acelerada y “plataformizada” a través de algoritmos que captan y, al mismo tiempo, disminuyen la atención humana; de esta manera, afectan las posibilidades de percepción y reflexión. O, como lo plantea Simondon, el sistema inorgánico de las redes técnicas se infiltra en la esfera orgánica del organismo biológico y social, y se apodera de su control.

Así, las tecnologías como productos humanos y productoras de nuevas formas de lo humano encarnan la doble utopía/distopía del progreso tecnológico y científico y, con ello, como lo mencionan algunos autores, la puesta en peligro de la existencia de la vida misma en este planeta. Esto nos exige aprender a pensar

transversalmente las interacciones entre ecosistemas, máquinas y universos de referencia sociales e individuales, porque las mismas tecnologías que ofrecen una apertura a formas heterogéneas de ver la realidad nos enredan, a su vez, en nuevas formas de dominación y totalización, una nueva economía política a través de algoritmos y datos. Como lo señalan Couldry y Mejías (2019),

[nos enfrentamos al] intento sistemático de convertir todas las vidas y relaciones humanas en insumos para la generación de beneficios. La experiencia humana, potencialmente cada capa y aspecto de ella, se está convirtiendo en el objetivo de una extracción rentable. Llamamos a esta condición la colonización por los datos, y es una dimensión clave de cómo el propio capitalismo está evolucionando hoy. (p. x)

En la figura de prosumidores (que no de ciudadanos), nos vemos enfrentados a la monetarización de las relaciones humanas, a la apropiación de todos los signos que se convierten en datos que alimentan o capitalizan las nuevas multinacionales tecnológicas. Esta lógica del capitalismo no se localiza exclusivamente en la producción, sino en los modos de vida, la política, el consumo cultural, los usos y las prácticas comunicativas (Berardi, 2003; Rueda, 2022). Esto parece ser invisible para muchos “utópicos digitales”. Y se trata de un consumo en el que el sur global participa principalmente como datos, materia prima y como mercado. Las empresas tecnológicas dominan y controlan la infraestructura digital, los datos y la propiedad del poder computacional. Como señala Ávila (2018), al recopilar los datos personales de los ciudadanos a una escala sin precedentes en la historia de la humanidad, las empresas pueden servir como conductos de campañas de desinformación que pueden alterar el flujo de la geopolítica mundial e incluso cambiar el resultado de las elecciones. Esta situación puede reforzar la dependencia general de los países del sur global, ya que las TIC se vuelven cruciales para el funcionamiento de sectores públicos como la educación.

Sin embargo, hay otros aspectos que quedan por fuera. Por una parte, se encuentra la discusión ética sobre cómo otros seres vivos también se ven influenciados por los desarrollos tecnológicos,

sobre las formas de reproducción algorítmica de esquemas de discriminación y racialización. Asuntos que consideramos de gran importancia para repensar las humanidades hoy, descentradas de la idea antropocéntrica que caracterizó la ciencia moderna y abierta a las complejidades de pensarnos en relación con el medio ambiente, con otros seres vivos no humanos, con los objetos técnicos y artefactos.

De hecho, la utopía sigue teniendo al “ser humano” en el centro. Pero el estado actual de nuestro planeta nos exige cada vez más salir de nuestro antropocentrismo y de los proyectos (utópicos) que lo han acompañado. Las ciencias sociales y humanas tienen en las últimas décadas el reto de superar sus referencias y metáforas para crear nuevos y complejos paradigmas. Este giro de la subjetividad propuesto hace tiempo por Guattari (1996) implica una tensión existencial, una utopía entre una subjetividad anclada en el pasado y un futuro cuya realización (actualización) involucra temporalidades humanas y no humanas (animal, naturaleza, devenir cósmico, maquínico), así como dimensiones institucionales y sociales que regulan la formación de individuos y grupos humanos. Se trata, en definitiva, de la producción de la existencia humana en nuevos contextos históricos, ecológicos, políticos, culturales y materiales.

Por su parte, la mayoría de las visiones críticas sobre la digitalización y las utopías digitales mantienen una narrativa que sitúa a Europa y al norte global en el centro de este proyecto sin cuestionar sus raíces y costos para el sur global, no solo en términos de conocimientos y desarrollo económico, sino también en relación con los impactos ambientales y los modelos extractivistas de minerales, saberes y prácticas de colonialidad digital. Pero la modernidad y la ilustración se produjeron en Europa, gracias y en medio de un proyecto civilizatorio colonizador del sur global que continúa hoy con el proceso de digitalización de nuestras sociedades. La cuestión es, entonces, si la digitalización favorecerá las viejas promesas modernas de igualdad y libertad y de qué manera contribuirá a resolver los problemas no resueltos de la modernidad y qué papel puede desempeñar en el tratamiento de los problemas actuales

de sostenibilidad del planeta. De nuevo, ese imaginario sobre la digitalización donde se trata es de mirar siempre “hacia adelante”, hacia el futuro, hacia el “progreso y desarrollo”, nos exige ver la complejidad de este nuevo proceso que se inscribe en la historia de un proyecto humanizador al cual se vincula tanto el norte como el sur global de forma desigual e inequitativa (Rueda, 2024).

Singularidades en América Latina

Los análisis de Julian Nida-Rümelin y Nathalie Weidenfeld, en el presente libro, permiten advertir cómo la perspectiva transhumanista, de la mano de los desarrollos tecnocientíficos, se va configurando en línea con los relatos producidos por la ciencia ficción, en los que el sueño de la inmortalidad, las ideas del superhumano, en el mundo feliz que se habita y se comparte con robots, máquinas y artefactos,² demanda, sin duda, enfoques humanistas que parecen ausentes, tal vez, como resultado de la fragmentación, la hiperespecialización de las disciplinas, del fortalecimiento y la renovación del capitalismo que nos deja prisioneros de conexiones e información, y nos desconecta de la realidad para mantener presa la atención en el “fetiche de la mercancía”, como señala Broncano (2012).

Una revisión de publicaciones académicas en la región sobre estos asuntos revela las temáticas de las que se ocupan algunos investigadores: *la condición del cuerpo* que se funde con dispositivos tecnológicos y manipulaciones científicas, a propósito de la tecnomorfosis, lo posorgánico y la idea cibernética; los debates relacionados con *lo natural* y *lo artificial* desde los que se van componiendo nuevos paisajes sociotécnicos en los que se difuminan las fronteras que marcaban sus límites; trabajos que presentan *nuevas estéticas* y proyectan *otras*

2 Relatos que transitan entre robots y máquinas esclavizadas por los humanos y humanos que son esclavizados por el mundo de los seres tecnológicos, tal como lo analizaba el filósofo de la tecnología Andrew Feenberg (2002); enfoques reconocidos como visiones autocráticas que descubren cómo las comprensiones de las tecnologías se mueven entre ideas amos y esclavos.

formas creativas a partir de conexiones entre arte, biología y tecnología, en fusiones que forjan mundos alternos; discusiones a propósito de la bioética en el centro de un nuevo contrato social que considera las implicaciones para la vida en sociedad, la naturaleza y las culturas; estudios que observan la necesidad de un humanismo actualizado a las condiciones subjetivas, sociales y ciudadanas en las llamadas ciberculturas articuladas a la educación; y finalmente, una línea de trabajos que se enuncian desde las humanidades digitales como perspectiva que pretende poner al servicio de la producción y difusión de las ciencias humanas y sociales las posibilidades de procesamiento de información y el uso de la llamada *inteligencia artificial* (IA), desde enfoques éticos.

Si bien las discusiones se abordan desde distintas áreas y núcleos de problematización, se perciben intereses comunes alrededor de un humanismo renovado, que trascienda la metáfora renacentista y que permita situar valores de la dignidad humana en la sociedad del conocimiento (Varis y Torneiro, 2012, citados por Cerutti y Battisti, 2023); o en posturas más desafiantes, un humanismo cibernético, como lo proponen, en la región, Escobar (2005) y, en España, Broncano (2009), quienes conceden al habitar posibilidades de lo humano que se transforma en simbiosis con objetos técnicos. En Escobar (2005), los futuros en construcción se configuran en entramados de lo tecnológico, lo comunitario y lo ecológico, distante del relato del desarrollo, lo que implica posibilidades arraigadas a contextos locales y la advertencia de las relaciones de poder colonial implícitas en las tecnologías contemporáneas; puesto que en Broncano lo cibernético es condición natural de lo humano, y la tecnología es parte de su constitución, su apuesta por un humanismo crítico la enmarca en la justicia y equidad, distante al individualismo y mercantilización del mundo de Silicon Valley.

Otras contribuciones, en torno a los cuerpos y a la condición transformada de los humanos, se observan en los trabajos de Laymert García dos Santos (1998, 2005, 2011) y Paula Sibilia (2005), entre otros. Para García dos Santos, la tecnología como fetiche en los países del sur global ha limitado su abordaje a la dimensión

del consumo, a la adquisición de objetos como representación de estatus, desde donde se asocia el uso de artefactos, dispositivos, objetos técnicos, a una ilusión del progreso, de posición social. Estas representaciones dificultan abordar la realidad sociotécnica en la que estamos inmersos y pasamos a ser ciudadanos reemplazados por algoritmos, como lo nombra García Canclini (2019), con una pérdida del ejercicio de la ciudadanía, por la limitada condición de consumidores. En este sentido, para García dos Santos, la crisis asociada a las categorías del humanismo se relacionan con la noción de lo poshumano, lo cual implica, por lo menos, tres consideraciones: una visión radical, la modificación biogenética, y la fusión acelerada de la tecnociencia y capitalismo.

La primera, una visión radical referida a la singularidad, donde se proyecta la superación de lo humano pues se parte del supuesto de su obsolescencia, demanda la adopción de prótesis para potenciar su condición natural, en una apuesta hecha por la inteligencia artificial y el desarrollo de dispositivos desde los cuales se avanza en la adecuación de lo humano a condiciones de ambientes modificados o externos al natural (por ejemplo, adecuación de la vida al espacio). La segunda alude a la modificación biogenética o biotecnológica que proyecta la transformación, no la superación de lo humano, en una especie de eugenesia positiva que implica la modificación del patrimonio genético con la aspiración de la “evolución del humano”, propiciando un desfase entre quienes disponen de condiciones y recursos para esa “evolución asistida” y quienes no pueden acceder a ella. Y una tercera que considera la fusión acelerada de la tecnociencia y el capitalismo desde donde se construye y fortalece el discurso de la obsolescencia de lo humano y la necesidad de trascender a la condición poshumana, como salida a la desactualización de la actual condición. Para García dos Santos (2005), volver sobre las ideas de lo maquínico, de Deleuze y Guattari, así como de individuación, de Simondón, contribuye en la construcción de una línea de análisis que traslada esta discusión al lugar de la existencia entre humanos, no humanos, el sentido de lo animal, el sentido de los seres vivientes, el sentido de las máquinas, llevando el debate al

campo de la política y de la ética, en una apertura comprensiva de la condición de humanidad más allá de “la esencia de lo humano”.

Otros autores como Arrubla Sánchez (2022) abordan el humanismo digital como respuesta al poshumanismo de Hayles (1999), que refiere la pérdida de fronteras ente lo biológico, lo tecnológico, lo cibernético y lo robótico, donde el cuerpo humano es parte del proyecto de transformación corporal o ciborgización; o de la idea de tecnomorfosis que en Sibila (2009) apunta a la permanente transformación del cuerpo que integra herramienta y aparatos en una relación posorgánica.

En otros abordajes, los debates sobre lo natural y lo artificial, a propósito de la configuración de los actuales paisajes sociotécnicos, analizan los principios éticos desde los cuales se vienen asumiendo las transformaciones de la vida y lo vivo, que inciden en las comprensiones clásicas de lo natural y artificial. Esto implica el reconocimiento de “diversas formas de vida”, incluye las que se producen en la confluencia entre ciencias de la información y los desarrollos de la biología, como componentes de paisajes contemporáneos que constituyen entornos habitables con formas de vida natural y adaptaciones de vida artificial, a través de dispositivos y sistemas computacionales (Rozo-Sandoval, 2023). En este sentido, Hernández-García *et al.* (2018) advierten la noción de *paisajes artificiales* que amplía la comprensión de la vida y del ser, donde conviven los seres biológicos y los “construidos por silicio o surgidos *in vitro*”, y mencionan cambios biológicos, culturales, y la emergencia de nuevos esquemas de pensamiento que van más allá de lo biofísico y lo digital (p. 18); así, se vuelve más que necesario proyectar una ecolítica de los paisajes artificiales a propósito de las dimensiones epistemológicas y conceptuales.

En un enfoque cercano a estas relaciones que se fusionan, se destacan las contribuciones desde el arte que, sin duda, en sus dimensiones sensibles, estéticas, propician otras formas inteligibles de las realidades contemporáneas; así, el bioarte narra construcciones posorgánicas que funden en un solo escenario (obra, *performance*, pintura, teatro, cine, ánimo, literatura, etcétera)

la transmutación del cuerpo, sus potencialidades físicas y cognitivas, influenciada por pos- y transhumanistas. Según Arrubla Sánchez (2022), es pertinente establecer criterios éticos, y señala cómo las libertades individuales y de autonomía tienen como límite el propio cuerpo, puesto que las implicaciones legales y éticas sobre reproducción y biogenética inciden en terceros; por tanto, no “pueden estar en manos de intereses económicos”, y hace perentoria la relevancia de abordar lo artificial, no como condición que se contrapona a lo natural y sí como parte de esta naturalidad artificial que estamos componiendo.

A propósito de estas configuraciones, vemos aún hoy la pertinencia de autores como Olivé (2007), quienes abogan por la exigencia de discutir y reflexionar sobre las nuevas relaciones entre ciencia, tecnología y sociedad, a partir de principios éticos y políticos de la ciencia y la tecnología, lo cual implica la disposición de formalizar un nuevo contrato social que exige un control social calificado en el que se reconozca la diversidad y la diferencia de los países del sur, en el que se pongan en discusión las decisiones que la ciencia y la tecnociencia toman en torno a la vida, a las formas de explotación y producción en aras de un tipo de bienestar y el desarrollo. En la apuesta de Olivé, ese contrato se proyecta hacia la construcción de una sociedad del conocimiento verdaderamente democrática y ciertamente intercultural, donde la diferencia étnica y cultural posibilite relaciones de creación, alternativas al desarrollo y encuentros de saberes, prácticas y conocimientos, no desde la explotación/expropiación y sí desde la noción de buen vivir.

Las discusiones que se adelantan en la región, referenciadas aquí de manera breve, implican modos de vida individual y colectivos; inciden directamente en la configuración de relaciones de poder, de inclusión, exclusión, y plantean algunos desafíos para nuestros contextos, por mencionar algunos:

- a. Trascender los enfoques instrumentales y deterministas que, desde la reducida visión de uso, contribuyen a fortalecer procesos de alienación tecnológica (Quintanilla, 2009), no por las tecnologías en sí ni por los objetos técnicos que

se instalan socialmente creando nuevos hábitos, formas de socialización y comunicación, entre otros, y sí por la ausencia de pensar sobre ellos (Simondon, 2007).

- b. Asumir enfoques críticos para comprender los cambios en los modos de vida, las nuevas corporalidades, los paisajes sociotécnicos contemporáneos, producto de los desarrollos tecnocientíficos que se presentan distantes en su producción para las realidades latinoamericanas, por las condiciones reales de inversión en ciencia y tecnología, pero cercanos en términos de la incidencia e implicaciones de sus usos (Rozo-Sandoval, 2023).
- c. Fortalecer el desarrollo y consolidación de investigaciones relacionadas con las implicaciones económicas, sociales y políticas de estos “avances” que presentan solo “beneficios en la línea del progreso”, en las que se advierte la construcción de un tipo de saber científico al servicio del mercado.
- d. Mantener la discusión sobre un nuevo contrato social que para los países que no disponen de la membresía de productores y generadores de estos desarrollos científicos y tecnocientíficos, demanda niveles de formación para comprender las discusiones y avances que tornan la actual condición de lo humano y de lo vivo, un asunto que implica tanto al norte como al sur global.

Al momento de redactar esta presentación, no podemos dejar de mencionar los debates que se han generado alrededor del ChatGPT y otras IA que, desde finales del 2022, han salido a la luz pública y comienzan a ser diseminadas en diferentes ámbitos. Así, cada vez hablamos menos de tecnologías como “instrumentos independientes”, como plantean Costa *et al.* (2023), sino que hoy hablamos de metatecnologías aplicadas a diversas y complejas

actividades, que aceleran el procesamiento y la gestión de lo social; como ensamblajes sociotecnológicos, que pueden incluso operar y regular otras tecnologías, y por tanto son “mundoambiente”, pero que, por sus implicaciones en los modos de vida, pueden ser de “alto riesgo” y demandan enfoques que integren la “ética con el pensamiento sistémico”.

De hecho, hay una evidente modificación específica, en términos de otros esquemas metafóricos de inteligencia:³ los procesos algorítmicos de simulación de estructuras del lenguaje en fusión con grandes volúmenes de información, para generar chats en línea con máquinas que articulan respuestas coherentes y estructuradas, que, aunque no siempre acertadas, pero sin duda a una velocidad que supera en muchos casos la capacidad de procesamiento humano y que trascienden los procesos de automatización por una capacidad tecnológica de autorregulación. Los alcances de la IA generativa, como las redes generativas antagónicas o adversarias (*generative adversarial network* [GAN]) y los modelos grandes de lenguaje (*large language model* [LLM]),⁴ que son las más potentes del aprendizaje maquina profundo, en la actualidad nos enfrenta a grandes desafíos, entre otros, para los sistemas educativos. Las GAN, nos dice Bongers (2022), son arquitecturas algorítmicas de autoaprendizaje profundo a las que no es necesario entrenar con datos etiquetados, porque aprenden entre ellas. Son redes neuronales capaces de realizar falsificaciones reales con base a imágenes preexistentes, atravesando así la distinción entre “lo real” y “lo falso”.

En las universidades se han hecho grandes debates principalmente a propósito del uso, con advertencias sobre la originalidad de los textos, los riesgos de plagio y una preocupación por la disminución de la creatividad humana; simultáneamente se discuten los alcances de este entorno complejo de comunicación en el que

3 Como los *chatbots* y sus antecesores *machine learning* en 1956 o el modelo matemático para crear redes neuronales en 1943, por citar algunos antecedentes.

4 Se basan en lenguaje natural y textos extraídos de internet; se usan en los *chatbots* como ChatGPT de OpenAI.

la máquina accede, procesa y produce datos a partir de volúmenes de información que se presentan como un tipo de “conversación en línea” que se alimenta también de la información que se produce en este proceso de “diálogo”, en el cual la interacción humana configura sentidos, difícilmente posibles en la máquina, pero se evidencia la capacidad de procesamiento de la parte máquina (integrada al sistema de programación) ante capacidades maquínicas “humanamente” imposibles de dominar.

Como se viene presentando, se reconoce que estos sistemas responden adecuadamente a reglas de la sintáctica, pero aún no consiguen dar cuenta de la semántica; el lugar de la significación y del sentido (por ahora) se circunscribe a la potestad humana. Al igual que otros programas de búsqueda que ya usamos hace bastante tiempo sin mayor reflexión sobre ellos ni sus implicaciones en nuestros modos de vida, ahora es perentorio comprender e interpelar los asuntos relacionados con la transparencia, imparcialidad del tratamiento de los datos que se muestran a la parte humana como imparciales y verídicos, dos condiciones sobre las cuales vale la pena debatir en términos de la proyección y representación de la realidad, de la ética (sobre la veracidad de la información que transcribe) y el tratamiento diferenciado, o con sesgos discriminatorios (por ejemplo: los algoritmos que reproducen patrones de raza, género, clase) desde los cuales se perpetúan inequidades e injusticias ya instaladas en la sociedad a través de las tecnologías que les antecedieron.

En noviembre de 2021, los países miembros de la Conferencia General de la Unesco firmaron la *Recomendación sobre la ética de la inteligencia artificial* como un primer instrumento normativo mundial al respecto, que se propone proteger la dignidad humana y promover los derechos. Este documento se convierte en la guía ética y la base normativa global para construir un estado de derecho en el mundo digital. No obstante, de nuevo aquí aparecen las preguntas por los procesos de democratización reales, que no constituyen parte de la agenda global. Así interpelamos la presencia, o ausencia, de los “otros”, de esos otros que históricamente han hecho parte de

la exclusión, de sus formas concretas de participación, de soberanía y de toma de decisiones en las políticas que les consideran como fuente de recursos minerales y humanos para sostener la infraestructura digital; así como el lugar de las decisiones que implican no solo la información que producen, en términos de datos, sino en relación con sus biosistemas, ecologías y formas de vida.

Finalmente agradecemos a Julian Nida-Rümelin y Nathalie Weidenfeld por cedernos los derechos a la versión del inglés al castellano, a las y los estudiantes que participaron en nuestro seminario, quienes no se limitaron a hacer la traducción al castellano, sino que hicieron un comentario al capítulo que tradujeron, en un ejercicio que si bien se apoyó en programas de traducción como DeepL, exigió un trabajo “humano” de revisión y de reflexión crítica. Así que gracias por sus aportes a este trabajo postseminario que nos exigió tiempos adicionales a nuestros compromisos académicos y laborales; y, por supuesto, nuestro agradecimiento a Juan Carlos Amador, por la escritura del prólogo a esta versión.

Tübingen y Bogotá, junio de 2025.

Referencias

- Adorno, T. (2004). *Escritos sociológicos* (tomo 1). Ariel.
- Arrubla Sánchez, R. (2022). Humanismo digital, cuerpo y tecnomorfosis. *Nexus*, (31). e30112265. <https://doi.org/10.25100/n.v0i31.12265>
- Ávila, R. (2018). Digital sovereignty or digital colonialism? *Sur-International Journal of Human Rights*, 15(27), 15-27.
- Benjamin, W. (1989). *La obra de arte en la época de su reproductibilidad técnica*. Tomo Discursos interrumpidos I. Taurus.
- Berardi, F. (2003). *La fábrica de la infelicidad*. Traficante de Sueños.
- Bongers, W. (2022). Los proyectos digitales de Eugenio Tisselli y la estética poshumana: entre cultura_RAM, neguentropía y necrocapitalismo. *Revista de Estudios Hispánicos*, 56(2), 219-242. DOI: 10.1353/rvs.2022.0018
- Broncano, F. (2009). *La melancolía del cibernauta*. Herder.
- Broncano, F. (2012). *La estrategia del simbiote*. *Cultura material para nuevas humanidades*. Editorial Delirio.
- Castro-Gómez, S. (2007). Giro decolonial, teoría crítica y pensamiento heterárquico. En S. Castro y R. Grosfoguel (eds.), *El giro decolonial: reflexiones para una diversidad epistémica más allá del capitalismo global* (pp. 9-24). Siglo del Hombre Editores; Universidad Central; Instituto de Estudios Sociales Contemporáneos; Pontificia Universidad Javeriana, Instituto Pensar.
- Cerutti, E. y Battiste, F. (2023). Educação na cibercultura: do novo humanismo ao humanismo digital. *Revista Cocar*, 18(36). <https://periodicos.uepa.br/index.php/cocar/article/view/6243>
- Costa, F., Mónaco, J., Covello, A., Novidelsky, I., Zabala, X. y Rodríguez, P. (2023). Desafíos de la inteligencia artificial generativa. Tres escalas y dos enfoques transversales. *Question/Cuestión*, 3(76), e844. <https://doi.org/10.24215/16696581e844>
- Couldry, N. y Mejias, U. A. (2019). *The costs of connection. How data is colonizing human life and appropriating it for capitalism*. Standfort University Press.
- Escobar, A. (2005). Bienvenidos a Cyberia. Notas para una antropología de la cibercultura. *Revista de Estudios Sociales*, 22, 15-35.
- Feenberg, A. (2002). *A teoria crítica da tecnologia: a crítica da racionalidade Tecnocientífica*. http://www.sfu.ca/~andrewf/books/Portug_Chapter7_Transforming_Technology.pdf
- Guattari, F. (1996). *Caosmosis*. Ediciones Manantial.
- García-Canclini, N. (2019). *Ciudadanos reemplazados por los algoritmos*. Centro María Sibylla Merian de Estudios Latinoamericanos Avanzados en Humanidades y Ciencias Sociales.

- García do Santos, L. (1998). Tecnología, naturaleza e redescoberta do Brasil. En A. Cortizas (comp.), *Tecnociência e cultura: ensaios sobre o tempo presente* (pp. 23-46). Editora Estação Liberdade.
- García dos Santos, L. (2005) Demasiadamente pós-humano: entrevista com Laymert Garcia dos Santos. *Novos Estudos*, 72, 161-175.
- García do Santos, L. (2011). *Politizar as novas tecnologias. O impacto sociotécnico da informação digital e genética*. Editora 34.
- Han, B-C. (2014). *En el enjambre*. Herder.
- Hayles, K. (1999). *How we became posthuman: virtual bodies in cyberspace, literature, and informatics*. University of Chicago Press.
- Hernández-García, J., Niño Bernal, R. y Hernández García, I. (2018). *Eco-política de los paisajes artificiales*. Pontificia Universidad Javeriana.
- Landow, G. (1995). *Hipertexto: la convergencia de la teoría crítica contemporánea y la tecnología*. Paidós.
- Martín-Barbero, J. (2003). *De los medios a las mediaciones. Comunicación, cultura y hegemonía*. Convenio Andrés Bello.
- McLuhan, M. (1996). *Comprender los medios de comunicación. Las extensiones del ser humano*. Paidós.
- Narváez, A. (2019). Comunicación educativa, educomunicación y educación mediática: una propuesta de investigación y formación desde un enfoque culturalista. *Palabra Clave*, 22(3). <https://doi.org/10.5294/pacla.2019.22.3.11>
- Olivé, L. (2007). *La ciencia y la tecnología en la sociedad del conocimiento. Ética, política y epistemología*. Fondo de Cultura Económica.
- Quintanilla, M. Á. (2009). Tecnologías entrañables [en línea]. *Público*. <http://blogs.publico.es/delconsejoeditorial/351/tecnologias-entranables/>
- Ricourte, P. (2019). Data epistemologies, the coloniality of power, and resistance. *Television & New Media*, 20(4), 350-365. <https://doi.org/10.1177/1527476419831640>
- Rozo-Sandoval, C. (2016). Prácticas comunicativas en contexto, para un conocimiento otro de las tecnologías. *Chasqui: Revista Latinoamericana de Educación*, (131), 181-198.
- Rozo-Sandoval, C. (2023). *Paisajes sociotécnicos y educación: la pregunta por los ambientes/entornos "de aprendizaje"*. Conferencia lección inaugural del campo de estudio: Educación superior, comunicación y mediación por TIC. Facultad de Educación, Universidad Pedagógica Nacional.
- Rueda, R (2022). Education and digitalisation in the global south: a reflection from the colombian case. *New Explorations. Studies in Culture and Communication*, 2(3) (n.º esp.: Digital Humanism and the Future of Humanity), 48-62. <https://jps.library.utoronto.ca/index.php/nexj/article/view/39025>

- Rueda, R. (2024). Cyberculture, (dys)topias and transformation. En J. Urabayen y J. León Casero (eds.), *Post-apocalyptic cultures. New political imaginaries after the collapse of modernity* (pp. 21-38). Palgrave Macmillan.
- Rueda Ortiz, R. y Uribe Zapata, A. (23 de febrero de 2021). Cyberculture and education in Latin America. *Oxford Research Encyclopedia of Education*. <https://oxfordre.com/education/view/10.1093/acrefore/9780190264093.001.0001/acrefore-9780190264093-e-1503>
- Rümelin, J. y Weidenfeld, N. (2022). *Digital humanism. For a humane transformation of democracy, economy and culture in the digital age*. Springer.
- Sibilia, P. (2005). *El hombre postorgánico. Cuerpo, subjetividad y tecnologías digitales*. Fondo de Cultura Económica.
- Sierra, F. y Maldonado, C. (eds.). (2016). *Comunicación, decolonialidad y buen vivir*. Centro Internacional de Estudios Superiores de Comunicación para América Latina (Ciespal).
- Simondon, G. (2007). *El modo de existencia de los objetos técnicos*. Prometeo Libros.
- Simondon, G. (2009). *La individuación: a la luz de las nociones de forma y de información*. Cactus.
- Stiegler, B. (1998). *Technics and time. I: The fault of Epimetheus*. Stanford University Press. Cactus.
- Torres, A. (2008). *La educación popular. Trayectoria y actualidad*. El Búho.
- Van Dijck, J. (2016). *La cultura de la conectividad: una historia crítica de las redes sociales*. Siglo XXI Editores.
- Walsh, C. (2013). *Pedagogías decoloniales. Prácticas insurgentes de resistir, (re) existir y (re)vivir* (tomo I). Abya Yala.
- Zuboff, S. (2015). Big other: surveillance capitalism and the prospects of an information civilization. *Journal of Information Technology*, 30(1), 75-89. <https://doi.org/10.1057/jit.2015.5>

Prólogo de la versión en español

Por Juan Carlos Amador

El libro *Humanismo digital. Por una transformación humana de la democracia, la economía y la cultura en la era digital*, de Julian Nida-Rümelin y Nathalie Weidenfeld, editado por primera vez en 2018, se ocupa de caracterizar los principales rasgos del humanismo en el contexto de la revolución mediático-digital del siglo XXI, después de la irrupción de la inteligencia artificial (IA). Esta última, entendida como un acontecimiento disruptivo que a simple vista parece un objeto de estudio exclusivo de profesionales e investigadores dedicados a la digitalización, en esta obra se objetiva como un hecho social que se cuele entre las dimensiones más profundas de la intimidad, las relaciones intersubjetivas, la vida familiar y el trabajo, entre otros ámbitos de lo humano.

De acuerdo con Nida-Rümelin y Weidenfeld, luego de las revoluciones Agraria (Neolítico) e Industrial (a partir del siglo XVIII) y digital (desde finales del siglo XX) constituye un rasgo fundamental de nuestra época que requiere ser problematizado, más allá de las posturas glorificadoras y apocalípticas, ampliamente difundidas desde la década de 1990. Las mutaciones de la cultura digital en el siglo XXI, influidas por las relaciones y tensiones entre las inteligencias humana y artificial, según los autores, son un foco de reflexión fundamental no solo por sus características morfológicas y técnicas, sino por sus implicaciones éticas, ontológicas, sociales, políticas y económicas.

Este planteamiento sugiere que el estudio de este fenómeno sociotécnico-cultural, al decir de Pierre Levy (2007), convoca áreas y disciplinas surgidas del pensamiento moderno como la filosofía, la sociología, la historia, la antropología y la psicología, entre otras. Probablemente, estas gramáticas especializadas, que buscan objetivar la realidad social desde distintos ángulos, constituyen la base del llamado humanismo digital. En la actualidad, estos procesos de objetivación se vuelven aún más complejos de dilucidar debido al lugar que ocupa lo digital en la configuración de la realidad, así como la influencia de la IA en los desordenamientos y reordenamientos culturales que interpelan las formas de vida humana y el orden social en la contemporaneidad. Parte de este desafío epistemológico es explorado en este libro.

La IA comprende el desarrollo de operaciones cognitivas incorporadas a sistemas informáticos que, en articulación con procedimientos algorítmicos, hacen posible la imitación de la inteligencia humana en lo que concierne a la realización de tareas, la resolución de problemas y la generación de explicaciones sobre diversos temas. Aunque la génesis de este concepto se relaciona con la experimentación llevada a cabo por matemáticos e ingenieros, desde la década de 1950, que devino en las llamadas *ciencias de la computación*, en la actualidad se materializa en dispositivos artificiales encargados de sintetizar y automatizar acciones propias de los seres humanos. La arquitectura de estos entornos informáticos se caracteriza por adquirir progresivamente funciones para asimilar, aprender y mejorar las condiciones de respuesta en un área o tópico de desarrollo.

Las primeras aproximaciones a esta forma de inteligencia funcionaron a partir de sencillos algoritmos basados en el análisis de las continuidades y discontinuidades de la información, empleando funciones de búsqueda, escaneo y selección en bases de datos. Posteriormente, surgieron sistemas complejos que hicieron posible la interconexión de redes neuronales artificiales, análogas a los circuitos neuronales del cerebro humano. Al parecer, estas acciones, que evidencian cierta plasticidad, movimiento y mutabilidad

en el procesamiento de la información, constituyen la base de formas de aprendizaje no humanas como los llamados *aprendizajes automático, por refuerzo, supervisado y profundo*.

En la actualidad, parte de estos desarrollos se evidencian en asistentes virtuales —como Alexa o Siri, operados por Google y Apple—, traductores polivalentes, sistemas de recomendación, motores de juegos, *chatbots* —como Chat Generative Pre-Trained Transformer [ChatGPT], creado en 2022 por OpenAI— y vehículos autónomos —como los que opera Tesla Autopilot—. Los más entusiastas afirman que estamos transitando de una “IA débil” a una “IA fuerte”. La primera se ocupa de realizar tareas de automatización, aplicadas con relativo éxito en sectores como la salud, las finanzas, el transporte y la educación. La segunda comprende la estructuración de sistemas que, supuestamente, piensan y actúan como humanos. El internet de las cosas (IoT), la robótica, los agentes inteligentes y la ciencia de datos son apenas algunos ejemplos para referirse a estas IA.

Esos desarrollos tecnológicos —que no dejan de ser abrumadores— despiertan en usuarios y audiencias distintos posicionamientos y reacciones. Una de estas respuestas se enmarca en la IA fuerte, examinada críticamente por Nida-Rümelin y Weidenfeld. Según los autores, esta representa la ideología de Silicon Valley; es decir que estos sistemas de *software* poseen conciencia y son capaces de tomar decisiones, al tiempo que sus acciones van más allá de simples simulaciones de las capacidades humanas. Este relato, no muy lejano de las narrativas futurológicas de finales del siglo pasado, expuestas por Toffler (1973), Negroponte (1995) y Papert (2003), promete cumplir con una suerte de profecía digital que asegura la transparencia, la interconexión, el desarrollo y la democracia.

Los autores dejan claro que este debate no es tan nuevo como se presume. En un recorrido, más bien general, admiten que, desde el siglo v a. de C., los pitagóricos afirmaban que era posible ordenar el mundo a partir de relaciones estrictamente numéricas, y que los estoicos, dos siglos después, plantearon que la necesaria correspondencia entre la razón del mundo y la razón humana está

determinada por el logos. Aunque Nida-Rümelin y Weidenfeld no entran en detalles, se entiende que parte de este determinismo originó respuestas de carácter humanista que abordaron complejos asuntos como las virtudes humanas, los valores sociales, los sistemas éticos, el libre albedrío, las creencias, la toma de decisiones y la inteligencia humana. En esta diversidad de asuntos, entre los siglos XVI y XX, varias corrientes de pensamiento de tipo ontológico, ético y epistemológico, asociadas al humanismo, adquirieron relevancia en el campo intelectual.

En este contexto de discusión, los autores proponen un plan de capítulos que aborda la reflexión ética, filosófica, social y política del fenómeno en cuestión, en diálogo con narrativas procedentes del cine y la literatura. Esta propuesta escritural, además de abordar de manera crítica asuntos que subyacen a la euforia de la IA fuerte, emplea recursos que, con frecuencia, invitan al lector a reflexionar sobre los atributos de lo humano y de lo social, en contraste con las funciones imitativas —para algunos autónomas— de las máquinas inteligentes. Según Nida-Rümelin y Weidenfeld, esta ideología de Silicon Valley parece mostrar dos tendencias dominantes: un desarrollo tecnológico en el que, pronto, las máquinas serán superiores a la inteligencia humana; y un deseo colonizador en el que los robots realizarán la mayoría de las tareas humanas, convirtiéndose en esclavos humanoides.

Frente a estas inclinaciones hegemónicas y por momentos utópicas/distópicas, el libro analiza diversos problemas y dilemas. Sin que sea una descripción estricta sobre el orden de los capítulos, pues no se trata de anticipar la riqueza de los abordajes propuestos, se destacan siete grandes tópicos: los derechos, las emociones, el animismo, la toma de decisiones, la capacidad de deliberación, la comunicación y la ética utilitaria. El desarrollo de cada uno parte de casos ilustrados a través del cine o la literatura, que luego son observados desde coordenadas teóricas inscritas en las ciencias sociales y humanas. Esta suerte de bisagra entre el lenguaje cotidiano-mediático y el lenguaje académico-especializado potencia

la lectura fluida y accesible de temas profundos desde el punto de vista ético, ontológico y político.

En relación con el primer tópico, se presenta un debate interesante sobre los derechos humanos y no humanos, que subyacen a las dinámicas —aparentemente— autonómicas de este tipo de máquinas. Aunque las narrativas utópicas/distópicas en torno a la adquisición de sentimientos en las máquinas inteligentes han entusiasmado a las audiencias durante las últimas cuatro décadas, los autores interpelan las posiciones que sugieren la necesidad de reconocer un marco de derechos dirigido a esta forma de inteligencia. Este debate se imbrica con el tópico relativo a las emociones en la IA, el cual insinúa que estos artefactos, al adquirir progresivamente funciones sensoriales, podrán sentir, sensibilizarse y generar reciprocidad emocional con las personas. En este caso, los autores ofrecen un repertorio interesante de escenas cinematográficas y debates éticos que ponen en duda estas posiciones.

Frente al animismo y la IA, los cuales profetizan que los entornos artificiales en cuestión adquieren paulatinamente una suerte de racionalidad propia, conducente a la toma de decisiones complejas, Nida-Rümelin y Weidenfeld presentan interpretaciones que articulan psicoanálisis, ética y filosofía. Luego de mostrar cómo estas posturas retornan a una suerte de mecanicismo ingenuo en el que los objetos supuestamente estarían dotados de movimiento, consciencia y alma, sostienen que la IA fuerte se ancla en esta perspectiva anacrónica, al tiempo que niega la capacidad de agencia de los seres humanos, la cual dista de los procedimientos imitativos de esta tecnología.

En cuanto a la toma de decisiones, demuestran cómo la IA evidentemente es capaz de tomar aquellas de carácter procedimental, pero está limitada para decidir sobre asuntos éticos que involucran la preservación de la vida, las relaciones humanas o la sensibilidad. En lo que concierne a la capacidad de deliberación, plantean que los dilemas y las decisiones morales no pueden ser “algoritmizados”; en esta vía, aseguran que la simulación de habilidades humanas

nunca podrá responder a asuntos morales contradictorios, tal como ocurre en los intersticios del mundo de la vida. Y, frente a la comunicación, señalan que no es posible diseñar algoritmos que representen la interacción y el intercambio humano, en el marco de la objetividad, la veracidad y la sinceridad.

Por último, se destaca el tópico alusivo a la ética utilitaria, frecuente en varios capítulos. A riesgo de sintetizar demasiado este punto de vista, este es tal vez uno de los aspectos que más se critica a la IA fuerte. Aunque la IA en la actualidad cuenta con recursos para tomar decisiones en el marco de un conjunto de procedimientos probabilísticos, esta solo puede evaluar las acciones humanas con arreglo a la utilidad. En otras palabras, los sistemas de *software* que configuran determinadas estructuras de aprendizaje automático eligen posibilidades desprovistas de valores y regulaciones de la conducta. Esta ética utilitarista, que opera como una racionalidad “consecuencialista”, no solo retorna a la filosofía moral de Bentham, sino que vacía de contenido moral asuntos como la justicia, la democracia y los derechos.

Esta sucinta aproximación a la obra de Nida-Rümelin y Weidenfeld, en la que el humanismo digital se contrapone a la IA fuerte en sus versiones materialista y animista, genera varias inquietudes y preguntas. Dadas las limitaciones de este espacio, se abordarán de manera global (a) el sentido del humanismo digital en el marco de las interpelaciones al antropocentrismo expuestas desde perspectivas críticas periféricas y del Sur y (b) la necesidad de incluir en la agenda de discusión crítica a la IA fuerte otros problemas que se inscriben en las herencias del mundo moderno-colonial en el sur global, tramitadas por medio de proyectos civilizatorios, desarrollistas y —en la actualidad— neoliberales.

Como se observa en el comentario de Rocío Rueda a la introducción de este libro, el humanismo se objetivó en la modernidad temprana, entre los siglos xv y xvi, a partir de una suerte de redescubrimiento de las artes y la filosofía de la Antigüedad clásica. Además de este interés, los promotores de este movimiento, que se emparentó con los principios del Renacimiento, se empeñaron

en plantear alternativas más allá de la educación escolástica y se opusieron a las doctrinas de corte metafísico, especialmente ligadas a la moral monoteísta. Aunque también existió un humanismo religioso, que asoció la ética con la libertad de culto, este sistema de valores antropocéntrico, secular y racional, se profundizó en tiempos de la Ilustración y tuvo marcadas influencias en las revoluciones burguesas de los siglos XVIII y XIX.¹

A lo largo del siglo XX, la ética, la filosofía y la teología retomaron de distintas maneras las bases del humanismo. En general, es posible distinguir cuatro corrientes humanistas en los albores de este siglo: marxismo, cristianismo, existencialismo y fenomenología. Mientras que el primero parte de los trabajos tempranos de Karl Marx sobre la *alienación*, que pretende ser redefinida según el proyecto socialista, el humanismo cristiano aboga por la complementariedad entre lo espiritual y lo humano, alrededor de principios como la dignidad, la solidaridad y la felicidad. El existencialismo sartreano cuestiona el antropocentrismo exacerbado, al tiempo que responsabiliza al hombre de sus acciones más allá de una esencia universal que lo condicione. Y el humanismo fenomenológico admite que el sentido de una sociedad está sujeto a las relaciones que el hombre establezca con sus semejantes en comunidad.

Las perspectivas críticas propias de las regiones periféricas del sistema-mundo, cuyo *locus* empezó a ser visible en 1948, interpelaron el humanismo desde distintos ángulos. La teoría poscolonial cuestionó cómo el antropocentrismo justificó la colonización en nombre de la civilización de los bárbaros, situación que se posterga en el tiempo, incluso luego de la descolonización; esta extensión de lo colonial en el tiempo se evidencia cuando otros —hombres, blancos e ilustrados— pretenden hablar por los subalternos (Spivack, 2003). También se objetiva frente a una especie de paradoja ontológica:

1 En el marco de los proyectos moderno-coloniales de América Latina y el Caribe, este sistema de valores fue parcialmente adoptado por algunos proyectos republicanos en la región. Los idearios de la Revolución francesa y la Independencia de Estados Unidos fueron los más influyentes.

mientras el humanismo existencialista enaltece al *dasein* como el ser que está ahí —arrojado al mundo (Heidegger, 2012)—, el poscolonialismo afirma que el colonizado es un *damné*, un subhumano ubicado por debajo de la línea de la vida, al ser negado material y simbólicamente (Fanon, 1983).

Asimismo, al fragor de los procesos emancipadores de las décadas 1960 y 1970 en América Latina y el Caribe, convertidos en espacio subalternizado del sistema-mundo moderno, la teología de la liberación, la educación popular, la ética de la liberación y la filosofía de la liberación cuestionaron el legado humanista de corte eurocéntrico. La primera, constituida por vertientes católicas y protestantes, y por comunidades eclesiales de base, identificó la lucha contra la injusticia a partir de principios antiimperialistas y anticapitalistas. La segunda, desde el pensamiento de Paulo Freire (2005), asumió la educación como una práctica de libertad en la que los oprimidos pueden llegar a desestructurar las formas de dominación ejercidas por los opresores.

La ética de la liberación considera que el proyecto modernocapitalista es un sistema antivida que, en el tiempo, produjo cuatro amenazas: la autodestrucción mediante la proliferación de armas biológicas y nucleares; la insustentabilidad de la tierra, provocada por el agotamiento de reservas naturales; el calentamiento global, un desequilibrio ambiental sin precedentes; y la escasez de agua dulce. Según Leonardo Boff (2002), si el objetivo del proyecto moderno es la dominación, la respuesta de la periferia es el cuidado.

Y, en relación con la filosofía de la liberación, esta destaca la transmodernidad, entendida como un proyecto que supera la dicotomía modernidad/posmodernidad, y que pretende develar las bases epistémicas y cosmogónicas de las culturas originarias de América Latina; los diálogos interculturales simétricos harían posible este propósito (Dussel, 2006). A este enfoque le da continuidad Aníbal Quijano (2014), quien en los inicios de la década de 1990 acuñó el término *colonialidad del poder*. Se trata de un patrón de dominación global propio del sistema-mundo modernocapitalista, originado en el colonialismo europeo del siglo XVI, que

se ha reproducido en el tiempo a través de cinco ámbitos de existencia: trabajo, sexo, subjetividad/intersubjetividad, autoridad colectiva y naturaleza.

El despliegue de esta última perspectiva coincidió con la introducción del neoliberalismo en la región. Esta fuerza global del capital, los discursos celebratorios de la posmodernidad, el fin de la historia y el capital humano golpearon fuertemente a los movimientos sociales en la región y subestimaron las perspectivas críticas de la periferia. Sin embargo, tras una revitalización del movimiento social (“sociedades en movimiento”, según Zibechi [2017]), impulsado sobre todo por comunidades indígenas y afrodescendientes, campesinos, mujeres, colectivos LGBTI, poblaciones de periferias urbanas y jóvenes, se empezó a configurar un giro epistémico, ético, político, social y cultural alrededor del sur, el “sentipensar” y los “buenos vivires”, comprendidos como alternativas al desarrollismo y al neoliberalismo promovidos por el capitalismo corporativo.

El sur es una metáfora que refiere al valor de las trayectorias de pueblos y comunidades históricamente oprimidos. El sentipensar propone que las cosmovisiones, los saberes y las prácticas de las comunidades ancestrales, campesinas y populares tiendan puentes entre sus posibilidades (sociología de las ausencias) y potencialidades (sociología de las emergencias) como alternativa para conquistar justicias de tipo social, ambiental y cognitiva (Santos, 2009). Y los buenos vivires son proyectos que integran lo ético, lo epistémico, lo cultural y lo estético, en torno a cuatro principios: (a) el restablecimiento de vínculos entre la naturaleza y los seres humanos; (b) la construcción de proyectos comunales en torno al territorio; (c) la valoración de saberes y prácticas no dualistas; y (d) la construcción de lo común, desde el interaprendizaje y la cocreación.

Según lo refiere Rocío Rueda en la introducción, la teoría decolonial, que recupera parte del pensamiento crítico latinoamericano expuesto, se distancia del paradigma humanista. Además de la colonialidad del poder, planteada por Quijano (2014), esta perspectiva amplía el concepto de *sistema-mundo* (ahora sistema-mundo moderno-colonial-capitalista-extractivista-patriarcal) y aborda

asuntos del orden social, racial y de género a partir de categorías como colonialidad del saber, del ser y de la naturaleza. Siguiendo a Rueda, al extrapolar estas perspectivas críticas al campo digital, es necesario interrogar los idearios humanistas frente a asuntos como la soberanía digital y la geopolítica del conocimiento, así como el uso de los dualismos modernos (sujeto-objeto, causa-efecto, civilizados-bárbaros, desarrollados-subdesarrollados, primer mundo-tercer mundo) para el análisis de estos fenómenos.

De este libro también subyacen temas y problemas, surgidos desde el sur global, que vale la pena considerar de cara a una posible agenda de discusión crítica a la IA fuerte. Al respecto, la digitalización ha sido fundamental en la configuración del nuevo orden mundial; según Castells (2006), este cambio planetario, que se inscribe en lo que denomina *informacionalismo*, opera como una manifestación renovada de producción y acumulación de capital que se potencia a través del trabajo con estas tecnologías. La digitalización en su última generación se ha intensificado a partir de nuevos procesamientos de la información, tal como se ha expuesto. Aunque estos entramados de datos son susceptibles de ser procesados y expandidos de manera autónoma, también es claro que pueden transmitirse a sectores, grupos y poblaciones con arreglo a fines ideológicos.

En este contexto, el despliegue del informacionalismo durante las últimas tres décadas ha sido estratégicamente cooptado por el neoliberalismo. En lo concerniente al sur global, el neoliberalismo se ha convertido progresivamente en un dispositivo de poder transnacional que se manifiesta en medidas como la disciplina fiscal, la redefinición del gasto público, las reformas tributarias, la liberalización del sector financiero, la privatización y la desregulación de la economía. Al tiempo, sus estrategias intensifican la implementación de proyectos minero-energéticos en países con alta diversidad hídrica, vegetal y de minerales; además de propiciar la devastación de los ecosistemas locales, estos direccionamientos corporativos gestionan el sostenimiento de conflictos ambientales, sociales y, a veces, armados.

Además de la implementación de estrategias que favorecen su crecimiento y concentración, el capitalismo neoliberal emplea mecanismos de control que configuran ordenamientos sociales funcionales a sus objetivos. Esos mecanismos tienen alcances en las emociones y sentimientos humanos, y operan por medio de cúmulos de información con los que se agencian los modos de ser y pensar de las personas. Varios de estos aspectos constituyen la base de la denominada *cuarta revolución industrial*, que involucra tecnologías de la información y las comunicaciones (TIC) (en especial aquellas que trabajan con algoritmos web dirigidos e IA), entre otros dispositivos que profundizan el trabajo inmaterial. Lazzarato (2006), inspirado en Marx, afirma que el trabajo inmaterial consiste en la generación de actividades cognitivas y afectivas que se mercantilizan dentro del circuito extendido de la economía capitalista.

En la actualidad, estas expresiones de poder emplean estrategias de clasificación, jerarquización y docilización tramitadas por las instituciones disciplinarias (Foucault, 2005) para alcanzar determinados objetivos biopolíticos, en el escenario del gobierno de las poblaciones. Por su parte, las manifestaciones de poder a distancia, virtuales y simbólicas, utilizan formas inmateriales de control social que se incorporan en los lenguajes, las comunicaciones, la información, los conocimientos y las tecnologías con fines de gubernamentalidad. Su propósito es asegurar la conducción de las conductas, en el marco de la “libertad” capitalista (Rose, 2011).

En este sentido, el nuevo panóptico deviene digital, transparente y placentero; no solo vigila las acciones y administra el tiempo y el espacio para que las personas sean más productivas y obedientes. Ahora se trata de un dispositivo que gestiona subjetivaciones para que el sujeto se coaccione y se regule, asumido como empresario de sí, individuo del rendimiento y trabajador/consumidor de lo inmaterial. Esta nueva biopolítica, que probablemente está en camino de renovación con la IA, opera como un tiempo-espacio apacible que reprime las desviaciones a través de nuevos sistemas de signos administrados por un capitalismo emocional que se

muestra amable, sutil y flexible, pero que constituye progresivamente un psicoprograma del inconsciente (Han, 2014). Al parecer, esta tecnología de poder legitima el pensamiento único y ahonda la llamada *crisis civilizatoria*.

Como respuesta posible a estas complejas condiciones de dominación, se ha construido progresivamente una suerte de programa de “cibercultura crítica” desde Latinoamérica (Rueda y Uribe-Zapata, 2022) que dialoga con apuestas que reivindican el valor de las convergencias culturales desde el sur. Esta última plantea que el pensamiento crítico y las resistencias al capitalismo neoliberal, desde lo mediático-digital, se inscriben no solo en las tecnologías y los medios, sino especialmente en las mediaciones culturales. Según Martín-Barbero y Corona (2018), estas mediaciones comprenden el conjunto de prácticas, interacciones e hibridaciones que se producen entre los contenidos construidos/emitados por una persona, grupo o medio de comunicación y la persona, grupo o audiencia que los recibe —como recepción, consumo o apropiación— en el contexto de unas condiciones históricas y de poder precisas.

Desde esta premisa, originada en el pensamiento comunicacional latinoamericano, se puede afirmar que las experiencias mediático-tecnológicas desde el sur se encuentran en un camino de reinención, tras la emergencia de la *convergencia cultural*. Este concepto refiere a la intersección entre la cultura digital, la cultura visual y la cultura popular. La primera se entiende como producción y comunicación participativa, que se apoya no solo en lo digital, sino en actitudes y posicionamientos éticos, que fomentan la construcción y distribución de la cultura. La segunda se centra en la construcción social de la mirada, desde gramáticas y artefactos visuales que producen, reproducen y transforman significados. Y la tercera comprende expresiones de autorrepresentación de sectores históricamente excluidos, cuyos procesos y productos figurativos, más allá de oponerse a la cultura de élite, invaden y agrietan los referentes simbólicos del poder (Albán, 2012).

En esta línea de discusión, se reconocen experiencias desde el sur que toman distancia de lo instrumental de las tecnologías y

que privilegia las dimensiones simbólica, sociocultural y política de estas. Por ejemplo, se encuentran proyectos de ciberactivismo que gestionan acciones colectivas y conectivas, desde lo físico y lo digital, como estrategia de resistencia y contrainformación. También se evidencian propuestas de periodismo digital alternativo, dedicadas a investigar, producir y difundir contenidos de interés público, a través de medios sociales. Asimismo, se identifican estéticas digitales que gestionan productos y prácticas en lo digital, y que promueven la autorrepresentación como estrategia de reexistencia (Albán, 2012). Por último, llaman la atención proyectos culturales y educativos expandidos que proponen iniciativas de negociación cultural que proceden de los ámbitos popular, mediático e institucional.

Con base en estos dos planteamientos —pensamiento crítico desde el sur y otras temáticas para una agenda crítica hacia la IA fuerte—, cabe preguntarse si el humanismo digital del siglo XXI puede transitar de la teleología antropocéntrica que caracterizó a sus predecesores a un proyecto bio-ético-cultural-político que se distancia de los dispositivos civilizatorios-desarrollistas-neoliberales, heredados del sistema-mundo moderno-colonial-capitalista-extractivista-patriarcal. Si la respuesta es afirmativa, es necesario interrogarse sobre si este tipo de humanismo, que sin duda propone lecturas valiosas frente a la ideología de Silicon Valley y su incorporación en la matriz capitalista neoliberal contemporánea, entablaría diálogos de saberes y prácticas de negociación cultural con los posicionamientos del sur global.

Los contenidos propuestos por Nida-Rümelin y Weidenfeld en este libro evidencian el valor de algunas perspectivas occidentales que no necesariamente legitiman el proyecto moderno-colonial, y que aportarían elementos claves para la puesta en marcha de diálogos entre el sur global y el pensamiento crítico europeo en la actualidad. Además de quedar expectantes a la respuesta que puedan dar los autores a esta invitación, es importante que los lectores, quienes se disponen a hacer un recorrido por temas novedosos, sugerentes y argumentados, también se pregunten

si es posible esta conversación en condiciones simétricas y de alteridad (Dussel, 2005). Resultaría igualmente productivo que se cuestionaran si pueden construir posicionamientos críticos frente a la digitalización —más allá de las acepciones glorificadoras y apocalípticas— en su vida cotidiana.

Referencias

- Albán, A. (2012). Estéticas de la re-existencia: lo político del arte. En W. Mignolo, y P. Gómez (comps.), *Arte y estética en la opción decolonial II* (pp. 433-448). Editorial Universidad Distrital.
- Boff, L. (2002). *El cuidado esencial. Ética de la compasión por la tierra*. Editorial Trotta.
- Dussel, E. (2006). *Filosofía de la cultura y transmodernidad. Ensayos*. UACM.
- Castells, M. (2006). *La era de la información. Economía, sociedad y cultura, I. La sociedad red*. Alianza
- Fanon, F. (1983). *Los condenados de la tierra*. Fondo de Cultura Económica.
- Foucault, M. (2005). *Vigilar y castigar. El nacimiento de la prisión*. Siglo XXI.
- Freire, P. (2005). *Pedagogía del oprimido*. Siglo XXI.
- Han, B.-Ch. (2014). *Psicopolítica. Neoliberalismo y nuevas técnicas de poder*. Herder.
- Heidegger, M. (2012). *Ser y tiempo*. Trotta.
- Lazzarato, M. (2006). *Por una política menor. Acontecimiento y política en las sociedades de control*. Traficantes de Sueños, Creative Commons.
- Lévy, P. (2007). *Cibercultura. La cultura de la sociedad digital*. Anthropos, Universidad Autónoma Metropolitana.
- Martín-Barbero, J. y Corona, S. (2018). *Ver con los otros. Comunicación intercultural*. Fondo de Cultura Económica.
- Negroponete, N. (1995). *El mundo digital*. Ediciones B.
- Papert, S. (2003). *La máquina de los niños: replantearse la educación en la era de los ordenadores*. Paidós.
- Quijano, A. (2014). *Colonialidad del poder, eurocentrismo y América Latina*. Clacso.
- Rose, N. (2011). Identidad, genealogía e historia. En S. Hall, y P. du Gay (comps.), *Cuestiones de identidad cultural* (pp. 215-250). Amorrortu.
- Rueda-Ortiz, R. y Uribe-Zapata, A. (2022). Cibercultura y educación en Latinoamérica. *Folios*, (56), 205-218. <https://doi.org/10.17227/folios.56-17013>
- Santos, B. (2009). *Una epistemología del sur: la reinención del conocimiento y la emancipación social*. Siglo XXI; Clacso.

Spivack, G. (2003). ¿Puede hablar el subalterno? *Revista Colombiana de Antropología*, 39, 297-364. <https://www.redalyc.org/pdf/1050/105018181010.pdf>

Toffler, A. (1973). *El shock del futuro*. Plaza & Janés.

Zibechi, R. (2017). *Movimientos sociales en América Latina. El mundo otro en movimiento*. Desde Abajo.

1. Introducción

Traducido por Rocío Rueda Ortiz

Es posible que en un futuro lejano miremos hacia atrás en la historia de la humanidad y hablemos de tres grandes innovaciones tecnológicas disruptivas: la transición de una cultura cazadora-recolectora a una cultura agraria sedentaria con cría de animales en el Neolítico; la transición a la era de las máquinas, basada en los combustibles fósiles en el siglo XIX; y, por último, la revolución digital del siglo XXI: el uso de la Inteligencia Artificial.¹ Si esto es así, estamos solo al principio de una revolución tecnológica, similar a la que Europa vivió en las primeras décadas del siglo XIX. Y al igual que entonces, las innovaciones tecnológicas de hoy van acompañadas de temores apocalípticos y también de expectativas eufóricas.

Este libro aborda los aspectos culturales y filosóficos de la Inteligencia Artificial y aboga por un humanismo digital. El humanismo digital es amigable con la tecnología, así como con el ser humano. Se diferencia de los apocalípticos porque confía en la razón humana y de los eufóricos porque respeta los límites de la tecnología digital.

El sueño de la creación de seres artificiales ha formado parte de los relatos mitológicos desde hace miles de años. En la Antigüedad, es el mito de Prometeo, una divinidad de la familia de los titanes,

1 Si escribimos *inteligencia artificial* aceptamos implícitamente la existencia de la inteligencia artificial. Si consideramos que esta aún es debatible, debemos escribir *Inteligencia Artificial* con mayúsculas, como hacemos en este libro.

que crea seres de arcilla pensantes y sensibles sin el permiso divino, por lo que es castigado amargamente por Zeus. En la Edad Media, encontramos la historia del Golem, un ser artificial hecho de arcilla, mudo y sin capacidad de raciocinio, pero que posee una gran fuerza y puede seguir órdenes. La literatura también utiliza el mito del ser creado artificialmente. En el cuento “The Sandman” de E. T. A. Hofmann (1816), el protagonista cae profundamente enamorado de una muñeca animada llamada Olimpia, y en el curso de los acontecimientos él termina perdiendo la cabeza por ella. Quizá el ejemplo más famoso de este periodo sea la novela de Mary Shelley, *Frankenstein o el moderno Prometeo* (1818). En esta trágica historia, un científico suizo crea un ser humano artificial. Este hombre artificial despierta tanta repugnancia y miedo debido a su tamaño y fealdad que no puede conectar con la sociedad humana; por el contrario, acumula cada vez más y más rabia y odio en su interior. Al final, mata a la novia de su creador y luego se suicida.

Hoy podríamos llamar a nuestros robots humanoides contemporáneos “Frankensteins”, si no hubiera existido la obra *R.U.R.*, del escritor checo Karel Čapek en 1920. Este drama trata de una empresa llamada R.U.R. (Rossum’s Universal Robots), que produce humanos artificiales llamados *robots* y abusa de ellos como mano de obra barata, quienes, sin embargo, en el transcurso de la historia se rebelan contra su esclavitud y acaban con la humanidad.

En los siglos xx y xxi, los robots viven sobre todo en novelas de ciencia ficción como las de Stanisław Lem² o del estadounidense Philip K. Dick.³ En los últimos años, las películas de ciencia ficción estadounidenses han recurrido en gran medida a la figura mitológica del humano artificial, que ahora aparece como un robot que coopera con los humanos en la tierra y en las naves espaciales. Además, la idea de un mundo totalmente digitalizado ha sido retomada por las películas y novelas de ciencia ficción. La visión es casi

2 Véanse, por ejemplo, *The star diaries* (1976) o *Golem XIV* (1985).

3 Escribió numerosos libros y relatos cortos en los que se basan muchas películas estadounidenses como *Blade runner*, *Minority report* o *Total recall*.

siempre distópica: hay mundos completamente dominados por las máquinas, como en el filme *The Matrix* (dirigido por Wachowskis, EE. UU., 1999), o sociedades futuristas de pesadilla, como la del filme *Demolition man* (dirigido por Marco Brambilla, EE. UU., 1993), donde las personas actúan y se comportan como si no existieran, e interactúan según instrucciones digitales, incluso el contacto sexual solamente puede tener lugar a través de la mediación de los dispositivos digitales.

Mientras tanto, muchas cosas con las que se fantaseaba en la historia de la humanidad se han hecho realidad; el ejemplo más famoso es el “comunicador” plegable del capitán Kirk, de *Star Trek*, que se materializó tecnológicamente unos cincuenta años después en forma de teléfono móvil StarTAC, de Motorola. Incluso parece que los mitos se limitan a adoptar una forma impregnada de nuevas tecnologías, pero permanecen inalterados en su esencia. El mito de la máquina con forma humana que se impone al final, el mito del muñeco animado, el mito de la amistad entre el hombre y la máquina. Pero, a diferencia de los siglos anteriores, estos mitos parecen ahora revitalizados por opciones tecnológicas concretas.

Sin duda, estamos viviendo una época de agitación tecnológica. Este siglo y el siguiente —muchos están convencidos— serán la época en que los robots tendrán a su cargo muchos trabajos humanos. Repartirán paquetes, producirán bienes en fábricas, conducirán taxis, actuarán como asesores bancarios, explorarán el espacio, trabajarán en centros de llamadas, operarán junto a los médicos en los hospitales y, posiblemente, escribirán novelas y crearán arte. Pero no hace falta mirar tan lejos. La digitalización ya ha penetrado no solo nuestro mundo laboral, sino también en nuestra vida privada, y ha tenido una gran influencia en nuestra vida cultural, política, económica y social.

Esta evolución plantea muchos interrogantes sobre las consecuencias que tendrá para nosotros. Algunos, como los autores de *best sellers*, Daniel H. Wilson (*Robopocalypse*, 2011) (un antiguo investigador asociado de la Universidad Carnegie Mellon que obtuvo un

doctorado en Robótica) o Stephen Hawking⁴ o filósofos como Nick Bostrom (2014), nos advierten de que los robots superarán algún día a la especie humana en capacidad de pensar y actuar, y se volverán contra la humanidad.⁵ Otros albergan esperanzas utópicas de un nuevo mundo digital en el que los robots digitales, como esclavos modernos, realicen el trabajo humano y establezcan un reino de libertad sin precedentes para nosotros.

Hay muchos indicios de que lo que se denomina *IA fuerte* en el discurso de la digitalización —es decir, la tesis según la cual los sistemas de *software* tienen conciencia, toman decisiones, persiguen objetivos, que sus actuaciones no son meras simulaciones de las capacidades humanas, sino que las realizan— se considerará algún día una forma de animismo moderno, es decir, el ensimismamiento de lo que no tiene alma, que puede verse como una regresión a modos infantiles de interpretación.

Por supuesto, esta ideología de la digitalización no se presenta como regresiva e infantil, sino, por el contrario, como racional y científica. Tiene una larga historia cultural. Comienza en nuestro ámbito cultural con los pitagóricos en el siglo v a. C. Es la idea de un mundo estrictamente ordenado en relaciones numéricas, cuya armonía y racionalidad solo se revelan en el análisis matemático. Doscientos años después, los filósofos estoicos añadieron a esta teoría la tesis de la correspondencia entre la razón del mundo y la razón humana (*logos*). Según esta teoría, las personas son capaces de pensar y actuar racionalmente porque pueden participar en la razón del mundo. El *logos* ordena el mundo según estrictas

4 Stephen Hawking advierte en muchas entrevistas contra el uso desenfadado de la Inteligencia Artificial. Por ejemplo, en la entrevista Focus en 2015 dice: “Nuestro futuro es una carrera entre el creciente poder de la tecnología y la sabiduría con la que la utilizamos” (Focus.de, 2015).

5 Científicos alemanes como el filósofo Tomas Metzinger también advierten de los efectos negativos de una “carrera armamentística de la IA”, al final de la cual podría surgir un *software* superinteligente que se desprenda de su ordenador y, como un gran virus incontrolable, se active en lugares cada vez nuevos y se globalice a sí mismo y sus objetivos.

leyes deterministas y los seres humanos tienen que entrar en este mundo de la razón. Sin embargo, incluso los estoicos y sus oponentes se dieron cuenta de que existe una tensión entre una visión del mundo de amplio determinismo y una visión de uno mismo como agente humano libre y responsable. Si la ideología de la IA conduce a una nueva edición de este conflicto, el humanismo digital supera este conflicto.

En este libro desarrollamos los principales rasgos de un humanismo digital como alternativa a lo que de forma un tanto simplista puede denominarse *ideología de Silicon Valley*. Esta se relaciona con la esperanza americana original, puritana, de salvación, de crear un mundo de puros y justos que han dejado atrás la mugre y el pecado. Sus valores centrales son la transparencia y la previsibilidad, el éxito económico y el patrocinio. En los tiempos de la transformación digital, esto incluye el sueño de una contraparte digital perfectamente construida, cuya construcción excluye cualquier error que nos lleve a una utopía tecnológica. El concepto clave es el de *Inteligencia Artificial*, cargado de metafísica y teología implícita, un sistema hiperracional, autoperfeccionado y cada vez más ensayado, cuyo creador, sin embargo, no es Dios, sino ingenieros de *software* que se ven a sí mismos como parte de una industria y de un movimiento de un paraíso digital en la tierra, basado en la transparencia, la interconexión y la no ambigüedad.

La ideología de Silicon Valley toma como punto de partida los impulsos humanistas solo para transformarlos en utopías anti-humanistas. Se inicia con la mejora de lo humano y culmina con su superación final e inhumana. Al querer mejorar la vida humana en el planeta, empieza a cuestionar las condiciones de la humanidad. De este modo, el humanismo se transforma en transhumanismo, lo cual conduce a una utopía tecnocrata en la que lo humano queda atrás. El humanismo digital se opone a esto y, en cambio, ofrece una nueva ética para la era de la Inteligencia Artificial.

Comentario de la traductora

Uno de mis primeros acercamientos a la noción de humanismo digital puedo rastrearlo en el taller que produjo el Manifiesto del Humanismo Digital, en Viena, en 2019.⁶ Más de cien asistentes del mundo académico, del gobierno, la industria y la sociedad civil europea participaron en un taller de dos días, con un propósito común: analizar y reflexionar sobre la relación entre el ser humano y la máquina y, lo que es igualmente importante, influir en su desarrollo para mejorar la vida y la sociedad. “La tecnología es para las personas y no al revés” fue su lema. De hecho, eligieron el término humanismo digital del texto con el mismo nombre de Julian Nida-Rümelin y Nathalie Weidenfeld (*Digitaler Humanismus*, 2018, en su edición alemana). En este manifiesto, se plantea que los seres humanos deben estar en el centro del mundo digital y que el progreso tecnológico favorecería la libertad, la paz y el progreso en armonía con la naturaleza. En ese sentido, a diferencia de perspectivas distópicas o negativas, el humanismo digital mantiene la convicción de que la tecnología puede favorecer el progreso de la sociedad y no solo la innovación en aras del crecimiento económico. O, en otras palabras, que la innovación tecnológica exige innovación social, y la innovación social requiere un amplio compromiso de la sociedad.

Desde una perspectiva histórica, pueden rastrearse varios momentos de la noción de humanismo. Una, proviene del periodo comprendido entre mediados del siglo xv y finales del siglo xvi (humanismo renacentista), con el redescubrimiento de las artes y la filosofía de la antigüedad, y donde filósofos y artistas se llamaban a sí mismos *humanistas*. La estética y la ética se centraron en el ser humano y no en lo sobrenatural o divino (por ejemplo, Leonardo da Vinci). Posteriormente en el periodo de la Ilustración (finales del siglo xviii) y la Revolución francesa, los principios de

6 Nos referimos aquí al trabajo reunido en el libro de Werthner *et al.* (2022).

la libertad humana y la democracia marcaron el espíritu humanista de la época. Así el humanismo se asoció con ideales educativos y pedagógicos centrados en valores como la dignidad humana y la humanidad. En la actualidad, el humanismo digital ha heredado la preocupación por los derechos humanos y por cómo mantenerlos en el ámbito digital, de tal modo que la conexión omnipresente acerque a todos a la justicia social, a la igualdad y a un mundo más sostenible y próspero.

Sin embargo, la noción misma de humanismo no está exenta de críticas. Así aparecen cuestiones sobre si somos los humanos los que definimos la tecnología o es la tecnología la que nos define a nosotros. Más aún, ¿quiénes son “nosotros” en esta pregunta? ¿Quién define la tecnología, quién tiene los conocimientos, los recursos, el poder de decisión? ¿Se trata acaso de asumir una sociedad digital en un sistema hegemónico del que —especialmente desde la perspectiva del sur global— no puede escaparse, a pesar del precio que los usuarios, las comunidades e incluso los países tienen que pagar con su dinero o sus datos, para formar parte de él?

La digitalización abre oportunidades sin precedentes, pero también suscita graves preocupaciones como la monopolización de la web, el aumento de las opiniones y comportamientos extremistas a través de redes sociales, la formación de burbujas de opinión como islas de verdades inconexas, la pérdida de privacidad y la extensión de la vigilancia digital. Cuando observamos la estructura actual de la sociedad digital, vemos que está muy centralizada y concentrada en el norte global, donde hasta la fecha residen los precursores de muchas innovaciones digitales y donde se produce más rédito económico. Esto, en gran medida, a costa del sur global, donde se encuentran los recursos minerales que se requieren para la infraestructura material de dichas tecnologías. Estos son explotados sin considerar los daños ambientales y sociales que dicha extracción de minerales produce. También en el sur global hay explotación de datos y es allí donde se realizan labores de sostenimiento de dicha infraestructura en condiciones muchas veces infrahumanas y contratos precarios.

Una vía para analizar críticamente la transformación digital en el sur global es a través de las perspectivas pos- y decoloniales. Estas ayudan a comprender los patrones de poder, a menudo tácitos, en el tejido social y tecnológico. Si aceptamos que la sociedad digital es una imagen del mundo físico, habrá heredado, junto con otros aspectos, patrones históricos de desigualdad como los que provienen de la colonialidad. Especialmente, los estudios decoloniales latinoamericanos han criticado el ideal humanista, por considerar que se apoya en una idea de supremacía cultural europea que, extendida al campo digital, lleva a preguntas por su postura sobre la soberanía digital y la geopolítica del conocimiento.

También se ha criticado el empoderamiento exclusivamente desde la racionalidad, basada en la abstracción y el cálculo. Los humanistas digitales desconfían ahora de las herramientas digitales y de las abstracciones del *big data*, pero estas se han convertido, paradójicamente, en requisito para enfrentar las complejidades de nuestro mundo. Aquí, por cierto, aparece una discusión sobre las dinámicas y nuevas formas de control y dominación económicas y sociales de las megaplataformas actuales como Google, Facebook y Twitter, agenciadas por algoritmos. Por ejemplo, en los algoritmos de Inteligencia Artificial (IA), que antes se consideraban objetivos y sin valores, se han descubierto sesgos discriminatorios. Hay varios ejemplos, como un algoritmo que blanquea autónomamente los rostros africanos y asiáticos, debido a los sesgos ocultos en los propios conjuntos de datos utilizados para el aprendizaje automático. En ese sentido vemos que la Inteligencia Artificial es un ámbito tecnológico que necesita urgentemente una descolonización innovadora.

También se ha criticado el humanismo desde los estudios ecologistas y ambientalistas por el excesivo antropocentrismo y exigen una relación respetuosa con la naturaleza y otros seres vivos. En este punto, las discusiones de movimientos sociales y ecologistas en el mundo —especialmente de comunidades indígenas y afrodescendientes— alertan cada vez más sobre la necesidad de establecer diálogos con otras ontologías y epistemologías, donde el ser humano es un ser en interdependencia e interrelación

con otros seres no humanos y en estructuras materiales, incluidas las digitales.

El humanismo digital es pues un campo joven con múltiples definiciones, entendimientos y perspectivas, y tiene diferentes raíces históricas. En América Latina existen diversas organizaciones y redes de humanidades digitales conformadas por equipos interdisciplinarios en ámbitos académicos. En Colombia, por ejemplo, en 2015, se fundó la Red Colombiana de Humanidades Digitales, en la que historiadores, artistas, literatos, bibliotecólogos, ingenieros, entre otros profesionales, debaten y gestionan asuntos como cartografías digitales, datos abiertos, web semántica, archivos digitales, narrativas audiovisuales, etcétera. Pero, en general, el campo del humanismo digital está en proceso de construcción de teoría respecto a la comprensión de la interacción entre la humanidad y la tecnología digital, así como en la búsqueda de posibles enfoques de diseños alternativos. Aquí el diálogo es con otras tradiciones e historias de desarrollo y apropiación tecnológica, también de otras ontologías y epistemologías como lo ha planteado el colombiano Arturo Escobar, como una práctica decisiva en la creación de mundos, que intenta romper con el antropocentrismo moderno e incluye lo no humano, lo no vivo, desde un imperativo histórico de recomunalización y reterritorialización (Escobar, 2016). Esta puede ser una vía fructífera para encontrar nuevas reflexiones sobre el futuro de nuestras sociedades y del planeta en su conjunto.

Referencia del comentario de la traductora

Escobar, A. (2016). *Autonomía y diseño: la realización de lo comunal*. Editorial Universidad del Cauca.

2. Los robots como los nuevos esclavos (digitales)

Traducido por Jenny Catalina Loaiza-Fuquen

Primera hora de la mañana, en algún lugar de una futura ciudad estadounidense. El detective Spooner, un tipo genial con un abrigo de cuero, zapatillas Converse y una gorra de béisbol, se prepara para ir al trabajo. Al abrir la puerta, se da cuenta de que hay algo que no está bien. Delante de él hay un robot humanoide de FedEx con un paquete bajo el brazo. “Buenos días, señor”, lo saluda amablemente el robot. “Otra entrega puntual de...”. Pero hasta ahí llega. “Quítate de mi vista, enlatado”, le dice Spooner, mientras lo empuja a un lado. El robot lo mira, aparentemente confundido, pero le desea un buen día, a pesar de todo.

Estamos en el año 2035. Los robots no solo se utilizan en las fábricas, sino también en los hogares. Caminan junto a la gente por la calle, sacan la basura, hacen compras y pasean a los perros de sus dueños. Al menos, así es como se ve el mundo en la película *Yo, Robot* (Alex Proyas, EE. UU., 2004). Estos robots se nos presentan como sirvientes sumisos a los que no se trata especialmente bien. Cuando se tropiezan con ellos, son ellos los que se disculpan. Su estatus es el de esclavos, cuya única finalidad es ser utilizados por los humanos.

Al principio de la película, se proyectan en la pantalla las siguientes leyes:

1. Un robot no debe dañar a un ser humano o, por inacción, permitir que un humano sufra daños.

2. Un robot debe obedecer las órdenes que le den los humanos, excepto las órdenes que entren en conflicto con la primera ley.
3. Un robot debe proteger su propia existencia siempre que dicha protección no entre en conflicto con la primera o la segunda ley.¹

Estas leyes, que son una cita directa del cuento “Runaround”, de Isaac Asimov, dejan muy claro cómo deben operar los robots.

Spoooner, el protagonista de la película, no siente más que desprecio por los robots. Para él, están bajo sospecha generalizada. Cuando se produce un robo en la ciudad, siempre se sospecha primero de los robots, no de los humanos.

Spoooner y el resto de la sociedad tienen poca compasión por sus esclavos mecánicos. Una vez que ya no son necesarios y han llegado al final de su utilidad, se les desecha en las afueras de la ciudad y se les almacena en contenedores, donde deben pasar el resto de su existencia digital, probablemente eterna. Ahí están, acurrucados como si quisieran consolarse mutuamente. Los “rostros” de los robots reflejan una especie de noble capacidad de sufrimiento. Son robots tristes. Robots que no entienden por qué son tratados tan mal. Proyas quiere que lleguemos a la conclusión de que el trato discriminatorio a los robots es injusto e inhumano.

Sin embargo, en realidad a nadie se le ha ocurrido aún aplicarles a los robots las normas de la Ley de Bienestar Animal, ni siquiera concederles derechos humanos. Hay un consenso práctico en que los ordenadores y los robots no tienen estados humanos. Estamos de acuerdo en que los robots, a diferencia de los animales, a los que se les atribuye la capacidad de sufrir, no son sensibles. Hasta ahora, no ha habido ninguna iniciativa seria para concederles derechos a

1 Isaac Asimov formuló posteriormente la llamada *Ley Cero de la Robótica*, presentada en su novela *Robots and Empire* (Robots e Imperio) (1985): “A robot may not harm humanity, or, by inaction, allow humanity to come to harm” [“Un robot no puede dañar a la humanidad ni, por inacción, permitir que la humanidad sufra daño”] (Asimov, 1985, p. 463).

los ordenadores o sistemas de *software* basados en su capacidad de sentir. No hay nada que sugiera que los sistemas de *software* más complejos de hoy tengan conciencia. Si la tuvieran, tendríamos que regular estrictamente su uso con efecto inmediato y atribuirles derechos fundamentales y humanos.

La muerte indolora, que es permisible para el caso de los animales, pero ética y legalmente inadmisibles en el caso de los humanos, estaría prohibida. Por analogía con el Proyecto Gran Simio, que quería superar el especismo y otorgarles derechos humanos a los animales parecidos a los humanos, en la medida en que tengan características comparables, también habría que concederles derechos humanos a los robots y a los sistemas de *software* autónomos. Si asumimos que los robots creados por nosotros son seres personales dotados de identidad, responsabilidad de acción, autonomía y con dignidad individual (la denominada *e-persona* [persona electrónica]),² los sistemas de *software* en cuestión ya no podrían manipularse por analogía con el derecho de autodeterminación informacional³ de los individuos humanos, porque esto contradeciría el principio kantiano de no instrumentalización de los seres racionales.

Sin embargo, algunos defensores de la Inteligencia Artificial afirman que es fundamentalmente imposible distinguir entre un cerebro humano y un ordenador. De este modo, abogados y sociólogos están cada vez más preocupados por la cuestión de hasta qué punto los (futuros) robots pueden considerarse responsables en caso de error, es decir, tener una responsabilidad jurídica. En los institutos de investigación internacionales, los juristas se preguntan si los robots deben ser considerados como meras herramientas en las que sus dueños o creadores deben ser responsables, o si, en función de su grado de autonomía, gozarán en algún momento de

2 Este tipo de "personalidad electrónica" se propuso en un proyecto del Parlamento de la UE en 2017.

3 El término fue acuñado por primera vez en Alemania por el Tribunal Constitucional Federal en 1983, cuando desestimó una ley aprobada por el Parlamento alemán sobre el censo, y describe el derecho de toda persona de controlar información personal relevante.

un estatus especial que les otorgue responsabilidad y derechos. Al fin y al cabo, según la argumentación jurídica, los robots también tendrían deberes que cumplir. En octubre de 2016, en Arabia Saudita, se concedió oficialmente la ciudadanía a un robot por primera vez en la historia. El robot en cuestión era “Sophia”, un androide con cara y cuerpo de mujer que simula mecánicamente las expresiones faciales. En teoría, la ciudadanía de Sophia no solo le otorga derechos, sino también deberes. El hecho de que se le permita moverse sin velo —a diferencia del resto de las mujeres saudíes— provocó un gran debate en Arabia Saudí y más allá.

En *Yo, Robot*, los robots tienen muchos deberes. Si no los cumplen, son procesados al igual que los humanos. Pero, entonces, ¿no deberían también tener derechos como los humanos? Al menos ese es el fundamento de la ética y el derecho en sociedades civiles y democráticas. Al igual que en *Yo, Robot*, la película *A.I. Artificial Intelligence* (Steven Spielberg, EE. UU., 2001) imagina un futuro en el que los robots se han convertido en una parte normal de nuestra vida cotidiana. Son esclavos y proveedores de servicios. Tristes proveedores, ya que se nos presentan como seres sensibles que sufren por ser tratados como humanos de segunda o, incluso, de tercera clase.

Spielberg deja clara su posición al trabajar medios melodramáticos para hacer creer al espectador que en un futuro próximo será esencial darles a los robots no solo derechos legales, sino concederles especialmente el derecho a la dignidad (humana). Quien se tome en serio la idea de Spielberg, de que un robot tiene la misma dignidad que un humano, debe asumir la indistinción entre los seres humanos y los ordenadores o sistemas de *software*. Ahora bien, quien piense que no puede haber diferencia categórica entre los cerebros humanos y los ordenadores está negando los fundamentos de la práctica científica y del modo de vida humano en general.

Sin embargo, quien está resentido con su PC, porque este tal vez ha demostrado ser desobediente, tiene un problema de racionalidad y realidad. Está atribuyendo a su ordenador propiedades

que no tiene. Solo en los seminarios de filosofía puede afirmarse la distinción entre los humanos y las máquinas. Fuera de estos contextos, esta afirmación parece grotesca, ya que es incompatible con la práctica real de quienes la defienden. Por supuesto, desechamos nuestros ordenadores cuando ya no los necesitamos, nos deshacemos de ellos en el desguace, sin derramar una lágrima. El ordenador no es más que una herramienta, mucho más compleja que una pala, que supera con creces algunas capacidades humanas, pero sigue siendo un aparato físicamente descriptible, sin deseos o creencias. En este sentido, no debemos esforzarnos por hacer que los robots sean lo más parecidos posible a los humanos. En una de las escenas más emotivas de la película de Spielberg, vemos cómo los robots desechados son llevados a una especie de circo. Bajo la mirada de una multitud que ruge, son introducidos en un cañón y lanzados al aire. “Pero yo sigo funcionando perfectamente”, protesta desesperadamente un robot, mientras es conducido a la arena.

Obviamente, los robots no quieren morir. Pero la multitud ebria no tiene simpatía. Para ellos, los robots no son más que una acumulación de metal. Para el espectador, sin embargo, los robots se presentan como seres sensibles, que sufren un trato injusto e inhumano. El hecho de que sean máquinas —que es el mensaje de la película— no significa que valgan menos que los humanos: tienen la misma dignidad.⁴

En filosofía, es bastante controvertido lo que constituye la dignidad humana. Algunos creen que una sensibilidad especial y una capacidad de sufrimiento son lo que exige una consideración especial. Otros creen que los seres humanos tienen derechos (básicos) por naturaleza —o por Dios— que son inalienables y que constituyen la dignidad especial de los seres humanos. Los que se sitúan en la tradición de Immanuel Kant basan la dignidad en la autonomía inherente al ser humano. En consecuencia, la capacidad humana de ponderar las

4 Para una crítica detallada del utilitarismo, véase Nida-Rümelin, J. (2023). *Morality and Rationality* (pp. 137-176). En *A Theory of Practical Reason*. Springer Verlag.

razones es lo que convierte a los seres humanos en agentes autónomos y les confiere un estatus especial de seres que tienen dignidad.

En su libro, *The decent society* (*La sociedad decente*), el filósofo israelí Avishai Margalit sitúa la dignidad humana y el respeto a sí mismo en el centro: no debemos tratar a nadie de tal manera que tenga motivos para sentirse humillado y dañado en su autoestima. La Inteligencia Artificial no tiene respeto por sí misma ni sentimientos que podamos herir. Su identidad personal no es vulnerable y no tiene la capacidad de reflexionar sobre su vida. Las condiciones previas para atribuirles dignidad no se cumplen.

Dado que la dignidad humana y los derechos humanos son tan fundamentales para nosotros mismos, pero también para el orden jurídico y político en el que vivimos, debemos tener cuidado de no poner en peligro este núcleo del *ethos* humano exagerando. Poblar el mundo con Inteligencia Artificial, a la que le atribuyamos capacidades y características comparables a las de los humanos, conduciría inevitablemente a la destrucción de este *ethos*. Desde este punto de vista, tiene más sentido leer a los robots subyugados de Spielberg como una metáfora del trato que recibieron los esclavos afroamericanos en la historia más que como una representación realista de un trato abusivo a los robots.

Al final de *A.I.* no quedan humanos en la tierra. No es una gran pérdida, como parece al espectador, ya que solo conoció a seres humanos de corazón frío a lo largo de la película. Los únicos seres que mostraban compasión en el filme eran los robots. Robots que han sido oprimidos y objetos de abuso. Al final de su larga historia de sufrimiento, el protagonista David, el robot más pequeño, es finalmente redimido por robots alienígenas con aspecto de ángeles que han llegado a la tierra. Él, que ha anhelado el amor de su madre humana, fallecida hace tiempo, tiene la oportunidad de reunirse con ella, ya que los robots alienígenas la devuelven a la vida a través de una reconstrucción del ADN. Por fin David puede ser feliz. Aunque esta felicidad durará solo un día (ya que la reconstrucción no puede sobrevivir más de veinticuatro horas), sus heridas pueden ahora curarse. La película se une así a las filas

de las narrativas melodramáticas de narraciones de influencia cristiana del siglo XIX, como la novela *Uncle Tom's cabin* (Harriet Beecher Stowe, 1852), donde el protagonista afroamericano, el tío Tom, debe soportar grandes penurias, sufrimientos e incluso la muerte para recibir la salvación (y hacer comprender a los lectores que el racismo es malo). Visto desde esta perspectiva, *A.I.* de Spielberg debe leerse no como una evaluación realista y seria de la situación de los robots, sino como una perpetuación de la narrativa cristiana del sufrimiento y la resurrección, y como un comentario metafórico sobre el racismo.

Comentario de la traductora

Les consulté a varias personas de mi entorno su opinión sobre tener un robot en su casa y me sorprendieron bastante las respuestas. Entre las más comunes estuvieron: “Desde que me haga aseo, pues, sí”; “Si lo puedo apagar, está bien...”; “Si lo dejo afuera en la noche para que cuide la casa, sí”; entre otras. Sin embargo, para mí, la más destacada fue: “Me daría miedo. Qué tal aprenda cosas y se vuelva como nosotros. No lo dejaría solo con mi hija, por ejemplo”. El miedo lo evidenció en todos los rostros encuestados al momento de preguntar qué tipo de robot, cómo sería, qué haría, cuándo lo traería, etcétera. Estas reacciones me hacen pensar que muchos de nosotros no estamos listos para estas conversaciones, no tenemos la información suficiente para procesar un cambio o incluir un robot en nuestros hogares, en la sociedad. Pasaría lo mismo que han descrito los autores en este capítulo, la crueldad, la antipatía, el trato diferente y un poco agresivo hacia ellos, viéndolos como minoría, pero reconociendo que hay un potencial de convertirse en la mayoría. Cuando pienso en los robots y su integración a la sociedad, ya sea con derechos o no, creo que sucedería lo mismo que con las personas extranjeras que se integran a una cultura diferente, las tratamos o de manera muy indiferente, o las repudiamos por ser “diferentes”. No puedo siquiera imaginar qué pasaría si les

otorgamos derechos a los robots, ya que ni siquiera podemos respetar los derechos que ya existen para los humanos e incluso para los animales. La tierra no deja de colapsar por la actitud de los humanos, por ese recelo que sentimos cada vez que alguien tiene éxito y pensamos que nos quitan parte del nuestro. Creo que pasaría lo mismo con los robots, tal vez tendríamos la sensación de que nos están quitando el espacio, el empleo, la intimidad.

Nuestra ontología no permite siquiera que otro ser humano de diferente raza se identifique como parte de otra. Que los animales no sean lastimados y utilizados con fines cosméticos, o sobre-explotados con propósitos industriales y de consumo, por ejemplo. Creemos que están para eso y que de alguna manera “nos deben” su existencia. Pasaría lo mismo con los robots y no tendríamos reparo en utilizarlos hasta que no quede un trozo de máquina en buen estado. Pero, al mismo tiempo, limitarlos a que hagan más para lo que fueron creados: servir a los humanos.

3. Estimulación digital de las emociones

Traducido por John Fernando Fajardo Velasco

Un joven rubio con pecas sale de un helicóptero hacia una hermosa isla. Exuberante vegetación, mansos arroyos, cascadas de agua. Después de caminar un rato por un denso bosque, finalmente llega a una casa hipermoderna, equipada con sistemas de máxima seguridad. La casa (y toda la isla, de hecho) pertenecen a Nathan. Él es el inventor y fundador del buscador de internet más grande del mundo llamado *Blue Book* (en alusión al famoso libro del filósofo Ludwig Wittgenstein). Nathan es un ingenioso y arrogante programador que se ha fijado el objetivo de crear una nueva especie: un robot capaz de tener conciencia. Caleb, el joven con pecas, trabaja en la empresa de Nathan y fue elegido por este para probar si uno de sus primeros robots había alcanzado este objetivo.

—¿Sabes qué es el test de Turing? —le pregunta Nathan, poco después de su llegada.

—Sí —responde Caleb—. Sé lo que es la prueba de Turing. Es cuando un humano interactúa con una computadora. Y si el humano no sabe que está interactuando con una computadora pasa la prueba.

—¿Y qué significa *aprobar*?

—Que la computadora tiene inteligencia artificial.

El robot que se supone que Caleb probará es Ava, una atractiva mujer robot. Su rostro se parece al de una mujer joven. Solo sus piernas y brazos están hechos de metal brillante, y los cables

azules brillan en su vientre. Cuando se mueve, se escucha un suave silbido, como si un tubo de neón estuviera zumbando. En varias sesiones, Caleb observa a Ava a través de un panel de vidrio a prueba de balas. A través del sistema de megafonía, Caleb le habla, le hace preguntas, la pone a prueba. Como una esfinge enigmática, ella se sienta frente a él y responde a todas sus preguntas, como un ser humano real y consciente de sí mismo. Sin embargo, después de un tiempo, Ava comienza a cambiar las cosas. Ahora ella es quien comienza a hacerle preguntas a Caleb. Mirándola a la cara, Caleb puede detectar muchas emociones en ella. Está sorprendida, a veces halagada, a veces desconcertada, a veces herida y finalmente enamorada. Y, sin embargo, Ava es una máquina. Entonces, ¿cómo es posible que tenga emociones?

Nathan pronto le dará a Caleb la explicación:

—Si supieras la dificultad de enseñar a leer y reproducir los gestos a la IA. ¿Sabes cómo lo descifré?

—No sé cómo hiciste nada de esto —responde Caleb, perplejo.

—Cada teléfono celular tiene un micrófono, una cámara y un medio para transmitir datos. Así que encendí todos los micrófonos y cámaras en todo el jodido planeta y redirigí los datos a través de Blue Book. Y, ¡zas! Un recurso ilimitado de interacción vocal y facial.

Ava es experta en expresiones faciales y vocales. Al observar a todas las personas del mundo y sus reacciones, ha adquirido un repertorio perfecto de conocimientos sobre las expresiones faciales a lo largo del tiempo. Sabe cómo interpretar las expresiones faciales y cuáles se consideran apropiadas en un momento determinado. *Big data* la convierte en una perfecta imitadora de las expresiones emocionales. Pero ¿eso significa que ella realmente tiene sentimientos?

—Quiero estar contigo. [...] ¿Quieres estar conmigo? —Ava le pregunta a Caleb en la quinta sesión.

A Caleb también le gustaría saber si Ava realmente siente algo por él o simplemente ha sido programada para fingir que lo hace. Eventualmente, Caleb decide creerle. Él la considera un ser

autónomo y único. Un ser del que se enamora y asume que también se ha enamorado de él.

En otra sesión, Caleb le cuenta a Ava sobre el experimento mental “La habitación de Mary”. Este fue un experimento real, propuesto por el filósofo australiano Frank Cameron Jackson en su ensayo “Lo que Mary no sabía” (1986).

—Mary es científica y su especialidad es el color. Sabe todo lo que hay que saber al respecto: las longitudes de onda, los efectos neurológicos, todas las propiedades posibles que puede tener el color. Pero ella vive en una habitación en blanco y negro. Nació y se crio allí, y puede observar el mundo exterior en un monitor en blanco y negro. Un día, alguien abre la puerta y Mary sale. Y ella ve un cielo azul. Y en ese momento, aprende algo que todos sus estudios no pudieron decirle. Ella aprende lo que se siente al ver el color.

Ava mira a Caleb inmóvil. A juzgar por la expresión de su rostro, esta historia le está pasando factura a Ava. Esto no es sorprendente. Después de todo, ¿no es ella como Mary? ¿Una persona que lo sabe todo, pero solo por información de segunda mano de internet? En el rostro de Ava, Caleb lee decepción, pero también una determinación feroz. Ella le deja claro a Caleb que ella también quiere salir de su habitación algún día. Preferiblemente —así le dice ella— con él. En su primera cita, ella le cuenta su mayor sueño: pararse en una intersección concurrida y ver pasar a la gente.

Cuando descubre que Nathan planea apagarla pronto para reciclar partes de ella para un nuevo robot, está decidida a hacer todo lo posible para escapar. Caleb quiere ayudarla y se le ocurre un plan.

Al final de la película, Caleb ha logrado descifrar el código del área de máxima seguridad. Ava escapa. Poco después, Ava mata a Nathan, su creador. Ya nada se interpone entre ella y su libertad. Pero luego sucede algo que ni Caleb ni el espectador esperaban en este punto: Ava deja a Caleb con frialdad, encerrado en una habitación. El espectador también está conmocionado en este momento, porque al igual que Caleb, ha tenido la sensación, en el

transcurso de la película, de que Ava es un ser sensible que no solo sufre por su situación, sino que también se ha enamorado de Caleb.

Mientras Caleb golpea desesperadamente la puerta que lo mantendrá dentro de la casa hasta que se muera de hambre, ella camina por la casa con un vestido blanco y zapatos blancos como un elfo. Con material orgánico extraído de otros robots desactivados, ahora sale al mundo. Su cabello castaño hasta los hombros acaricia su delicado rostro. Mientras respira el aire del bosque por primera vez, sonrío. Toca ramas y mira con curiosidad su nueva vida. No siente remordimientos y ni siquiera mira hacia atrás.

Al igual que Mary, ahora sale de su habitación al gran mundo, lista para tener sus propias experiencias. ¿Aprenderá no solo a imitar emociones, sino también a tenerlas? ¿O seguirá siendo una máquina para siempre? Esta es la esencia de todas las preguntas filosóficas en torno a las cuales los entusiastas de la IA siguen dando vueltas.

Caleb también se sigue haciendo la pregunta: ¿Ava solo ha aprendido a imitar ciertos comportamientos para dar la falsa impresión de que tiene sentimientos muy parecidos al actor “frío” descrito por Diderot, cuyo arte se centra en el dominio perfecto del comportamiento físico? Sin embargo, la pregunta verdaderamente inquietante es la siguiente: ¿qué pasaría si no solo los sentimientos de Ava, sino también los nuestros, no fueran más que puro comportamiento? Eso, al menos, es lo que afirman los positivistas radicales, defendiendo la tesis metafísica de que los estados mentales no son más que patrones de comportamiento. La comprensión positivista de la conciencia identifica propiedades y estados mentales, como tener miedo o tener deseos o creencias, con comportamientos particulares. “Jacob tiene dolor” no significa, en la comprensión del positivista, nada más que “Jacob se comporta de cierta manera, por ejemplo, grita ‘ay’ o retira bruscamente la mano de la estufa”.

No es casualidad, por cierto, que la película haga varias referencias al filósofo Ludwig Wittgenstein (una vez con el nombre de *Blue Book*, que es tanto el nombre de la compañía de Nathan como el título del famoso libro de Ludwig Wittgenstein, y otra vez con el

retrato de Gustav Klimt, de Margaret Stonborough-Wittgenstein, la hermana de Ludwig Wittgenstein, que cuelga en la casa de Nathan), ya que la mayoría de los estudiosos considera que Ludwig Wittgenstein es un “conductista”.

Sin embargo, si el conductismo fuera cierto, tendríamos que asumir que SIRI, el *software* de comunicación establecido en muchos teléfonos inteligentes, tiene sentimientos muy similares a los nuestros. Después de todo, reacciona como si estuviera realmente decepcionado o preocupado. Pero el *software* solo simula sentimientos, no los tiene.

Mucho más plausible que el punto de vista conductista sobre los estados mentales es el punto de vista realista: el dolor caracteriza cierto tipo de sentimientos que son desagradables y que generalmente tratamos de evitar. En el dentista nos esforzamos en suprimir cada reacción para no interrumpir el tratamiento, pero esto no significa que no sintamos dolor. Incluso el superespartano imaginario que no muestra ninguna emoción, incluso en un dolor extremo, todavía puede sentir dolor. Es simplemente absurdo identificar “tener dolor” con ciertos patrones de comportamiento.¹

Quizás el argumento fundamental contra la identidad entre los estados o propiedades mentales y los estados o propiedades neurofisiológicos o digitales se llama el *argumento de los qualia*. En su famoso ensayo “What is it like to be a bat?” (1974), Thomas Nagel argumenta que no es posible saber qué se siente ser un murciélago (es decir, qué siente el murciélago), incluso, si uno examina su cerebro en detalle. Estos llamados estados mentales cualitativos del murciélago no pueden determinarse a partir, únicamente, del conocimiento de los estados neurofisiológicos. Entonces, el argumento de los qualia está en contra de la identidad de los estados neurofisiológicos y mentales (Chalmers, 2010).

1 Por supuesto, nuestra capacidad humana para atribuirnos mutuamente estados mentales correctos depende de que sigan siendo comunes los patrones de comportamiento, y las personas expresen sus estados emocionales de manera similar. Solo podemos aprender cuáles son los sentimientos de otras personas porque compartimos ciertos patrones de respuesta.

Caleb cree que Ava está en la misma situación que Mary en el experimento mental de Jackson. Ella sabe, como le dijo Nathan, todo sobre el mundo, así como sobre las personas y sus sentimientos, pero eso no quiere decir que entienda lo que significa experimentar el mundo y tener sentimientos.

Por supuesto, uno también puede rechazar la identidad de lo mental y lo neurofisiológico, y, aun así, argumentar que lo mental solo puede ocurrir en conexión con lo material. De hecho, hay mucho que sugiere que la conciencia humana solo es posible debido a las funciones cerebrales correspondientes. Pero ni siquiera aquellos que sostienen que la conciencia humana se basa esencialmente en procesos neurofisiológicos suscriben la teoría de la identidad de lo mental y lo físico. Que los estados mentales de los humanos sean realizados por estados cerebrales (es decir, procesos y estados neurofisiológicos) no significa que sean causados por ellos.

Es innegable para nosotros los humanos que tenemos propiedades mentales, que tenemos ciertos estados mentales, que tenemos creencias, deseos, intenciones, miedos, expectativas, etc. Estamos convencidos (al menos la mayoría de nosotros) de que estos fenómenos mentales son realizados por procesos en nuestro cerebro, o al menos se correlacionan con ellos. La perspectiva en primera persona juega un papel crucial en esto. Sin embargo, esto no debe radicalizarse en una visión solipsista según la cual estoy solo en el mundo y mi mente es la única que existe. La comprensión del mundo de la vida ocurre esencialmente a través de nuestra interacción y cooperación con otros a quienes atribuimos propiedades mentales comparables. Para los niños pequeños, prelingüísticos, no solo son importantes las experiencias hápticas del mundo, las percepciones sensoriales, sino también el intercambio, la interacción y la comunicación con otros miembros de la especie humana, mayores y lingüísticamente capaces. Este papel del Otro no es posible sin una percepción (presuntamente anclada genéticamente) de otras mentes, incluso en niños prelingüísticos. Así comienza la concepción humana del mundo; dudar de esa base haría que nuestro mundo

colapsara (Nida-Rümelin, 2010). Así como no puede haber duda razonable para nosotros acerca de otras mentes, tal como están las cosas, no puede haber duda sobre el carácter no psíquico de lo digital. Negar la correlación de lo mental y lo físico en humanos y mamíferos altamente evolucionados, que se parecen lo suficiente a nosotros y permiten al menos un reconocimiento rudimentario de sus estados mentales, no se justifica porque mentaliza estados y procesos digitales. Los estados y procesos digitales simulan los mentales, pero no son idénticos a ellos, incluso si esa simulación fuera perfecta. No hay nada que sugiera que los estados y procesos mentales puedan ser realizados por los digitales. La simulación no debe confundirse con la realización.

En la escena final de la película *Ex Machina*, vemos a Ava caminando por el bosque, visiblemente impasible. Al lograr su libertad, también ha alcanzado su objetivo. Sin embargo, eso no prueba que Ava tenga conciencia. Después de todo, como dice el propio Nathan en algún momento de la película, ella fue programada para querer la libertad. Visto desde ese punto de vista, simplemente estaba ejecutando su programa. Incluso, si la película en sí sugiere a veces que Ava tiene sentimientos, optamos por otra interpretación y tomamos el hecho de que matar a dos personas (Nathan y Caleb) aparentemente no representa ningún problema moral para ella, como una prueba de que Ava no tiene conciencia ni emociones. Fue un error fatal de Caleb creer que sus expresiones faciales y gestos eran expresiones de emociones genuinas. En este sentido, queremos leer la película como una advertencia para no caer en la misma trampa en la que cayó Caleb, cuando proyectó en Ava mucho más de lo que ella realmente tenía. Por consiguiente, interpretamos la siguiente declaración de Nathan: “Un día, las IA nos mirarán de la misma manera que miramos los esqueletos fósiles en las llanuras de África. Un simio erguido que vive en el polvo con lenguaje y herramientas toscos, todo listo para la extinción” no como una profecía realista, sino como una expresión de fantasías masoquistas sobre la extinción de la civilización occidental.

Comentario del traductor

“Cuando un hombre es una tetera vacía, debería estar con ánimo; y, sin embargo, estoy destrozado... simplemente porque intuyo que podría ser algo parecido a un humano, si simplemente tuviera un corazón”: Hombre de hojalata.

Fleming (*Mago de Oz*, 1939).

El capítulo 3 plantea una pregunta central en relación con los sentimientos y las emociones, al considerarlos como comportamientos. Esta a su vez me lleva a preguntarme si las expresiones emocionales de los seres humanos involucran procesos de reproducción naturalizados por medio de los patrones de crianza y de las prácticas culturales que intervienen en la producción de subjetividad y que pueden ser transformados o si, por lo contrario, las expresiones emocionales son acciones instintivas irracionales de los humanos como vestigios de la evolución que surgen como respuestas a estímulos o fenómenos de la realidad. Es interesante pensar en que, sumado a esta discusión ya abordada por psicólogos positivistas, antropólogos y sociólogos, el papel de los dispositivos digitales, más aún, de la digitalización de casi todos los escenarios de la vida, redes sociales, relaciones laborales, educación, entretenimiento, etcétera, han reconocido que el capitalismo actual puede recurrir a estos dispositivos para generar emociones y sentimientos enfocados en el consumo y la apropiación de imaginarios de progreso y de subjetividad multilocal, en un mundo en donde el “infierno de lo igual” es el común denominador (Han, 2018).

Al igual que Caleb, quien piensa que Ava puede estar sintiendo emociones y sentimientos hacia él de manera autónoma y libre, a pesar de la dificultad de contemplar la idea que una máquina logre desarrollar construcciones mentales tan complejas como emociones, hoy los seres humanos nos vemos avocados al uso de algoritmos para encontrar la pareja perfecta o el amor de la vida. Por ejemplo, es, en la actualidad, la plataforma que mejor comercializa el amor a solo un clic, o *like*, de distancia.

Es posible que aquí emerja otro debate en el texto: la libertad de expresión de las emociones en la era digital. Ava puede sentir emociones y sentimientos, los cuales expresa por medio de sus palabras y de los gestos que acompaña con su cuerpo de androide. ¿Sentimos los seres humanos libremente? Es decir, ¿son nuestras emociones genuinas y las acciones que se derivan de ellas realmente libres? Desde luego, no desconozco que la categoría de libertad ha sido de las más debatidas y estudiadas por cada área del conocimiento, y que en los albores de la modernidad se pensó que gracias a la victoria de las revoluciones liberales, de la Declaración de los Derechos Humanos y de la ratificación de estos, la libertad más que concepto sería una realidad. Pero, en el mundo contemporáneo, y al igual que las instituciones y valores modernos, la libertad entró en crisis. La libertad es hoy funcional a la (auto)explotación de los seres humanos, a través de plataformas como TikTok, Facebook, Instagram, etcétera. Estas mantienen la hiperconexión de los sujetos en medio de una narrativa de la libertad, pero el proyecto del hombre libre cae en la sociedad del rendimiento, el sujeto se autoexplota de forma ilimitada y es libre para elegir según lo que le ofrece el mercado de manera acrítica (Han, 2016).

Por supuesto, las emociones no escapan a este nuevo orden, ya que, si estamos en una sociedad del rendimiento, la gestión de las emociones no solo pasa por el sentirse bien, sino también por la competitividad que se puede adquirir, si se desarrolla una inteligencia emocional. En el ámbito empresarial, es funcional, ya que así se obtiene un empleado dócil y productivo; las emociones son recursos para incrementar la productividad y el rendimiento; ya no se consumen solo cosas, se consumen emociones que no se agotan, son infinitas (Han, 2016).

Otro elemento interesante abordado por los autores es el de la conciencia, en particular cuando se hace referencia al cuarto de Mary, experimento planteado por el filósofo Frank Jackson. Recordemos que este experimento forma parte de la *teoría del fisicalismo*, según la cual todo lo que existe es exclusivamente físico, ya que solo es comprensible y se tiene conciencia de ello en la medida que se

experimenta y se percibe físicamente, y se procesa cerebralmente. Es decir, las emociones y sentimientos, como respuestas a los estímulos del contexto, son expresiones electroquímicas, pues se generan en el sistema límbico, pero, si solo se las mira como expresiones electroquímicas a fenómenos del mundo exterior, no podrían ser transformadas, pues solo se reducen a manifestaciones instintivas del hombre animal; por consiguiente, solo habría la posibilidad de gestionarlas o reprimirlas, una mirada muy conductista de las emociones, los sentimientos y los comportamientos.

En este sentido, Ava solo aprendió a imitar emociones, expresarlas corporalmente en comportamientos, pero no a sentir las realmente. El dilema es si Caleb también solo reproduce comportamientos derivados de emociones construidas culturalmente, aspecto que iría en contra del imaginario de la conciencia como un conjunto de estados mentales expresados en comportamientos de manera instintiva. Para antropólogos y sociólogos como David Le Breton (1998), estos comportamientos no surgen de la nada, sino del intercambio cultural y social que cada individuo construye desde las primeras etapas de la vida y se expresan desde la corporalidad, se convierte en medio y lenguaje que enuncia lo socialmente aceptado como discurso. La interpretación que se haga de estas expresiones sería subjetiva, ya que depende de la sociedad en donde estas se dan.

En esa misma línea, es necesario comprender que las experiencias particulares que sentimos justo cuando estamos en contacto con las cosas que nos rodean, y que son difíciles de comprender y de expresar, son vivencias subjetivas que parten de la propia conciencia. Así, a pesar de compartir un atardecer dos personas, su experiencia y emocionalidad sobre el fenómeno son distintas, únicas; a esto se le denomina *qualia* (Dennett, 1992). Los seres humanos sienten y procesan cada estímulo de una manera concreta y singular, que es mediada también por el contexto en donde se está inmerso; es la conciencia la entidad que da sentido a esas experiencias y valor, según la interpretación que surja de ella. Aquí encontramos el valor de la socialización y la ruptura con el

imaginario de la hiperindividualización de las sociedades contemporáneas, ya que el tejido que surge de las interacciones es la que potencializa las experiencias particulares y las emociones que se construyen a partir del intercambio cultural.

En este sentido, resalto la invitación que hacen los dos autores del texto a reconocer que en la globalización el impacto de lo digital gestiona comportamientos y emociones, pero, a su vez, abre la posibilidad de resignificar el humanismo, mucho más allá de la simulación, de la recreación de ambientes o espacios en la web o en la digitalización de la existencia. Esto llevaría a rescatar la realización vista desde la cocreación de nuevas apuestas comunitarias y cooperativas, que aletarguen los procesos de extinción de la vida, no solo biológica, sino humana, ética y moral, ya que, como Ava, al asesinar a Caleb y a Nathan, dio muestras de no tener conciencia ni emocionalidad, hoy los seres humanos nos matamos unos a otros, en una disputa entre antagonicos, por la dominación disfrazada de supervivencia, retornando así a la ley del más fuerte, no solo en la realidad física, sino también, en la realidad aumentada: la de la era digital.

Referencias del comentario del traductor

- Dennett, D. (1992). Quining qualia. En A. J. Marcel y E. Bisiach (eds.), *Consciousness in contemporary science* (pp. 42-77). Oxford University Press.
- Fleming, V. (dir.). (1939). *El Mago de Oz* [película].
- Han, B.-C. (2016). *Psicopolítica*. Herder.
- Han, B.-C. (2018). *La expulsión de lo distinto*. Herder.
- Le Breton, D. (1998). *Las pasiones ordinarias, antropología de las emociones*. Nueva Visión.

4. El problema de la autonomía y la determinación en el mundo digital

Traducido por Óscar Alejandro Barrios Candil

Algo inseguro, Neo se para en la puerta de la cocina del Oráculo, una señora mayor que hornea galletas, que fuma alegremente un cigarrillo tras otro. Se espera que ella le diga si él es el “elegido” o no, aquel que destruirá el programa digital *Matrix*, al que todos los humanos están ligados y lo confunden con la vida real. Destruir *Matrix* significaría el fin del reinado de las máquinas y el comienzo de una vida real y autodeterminada.

—Y no te preocupes por el jarrón —le dice poco después de que él ha cruzado el umbral.

—¿Qué jarrón? —pregunta, mientras se da la vuelta y accidentalmente tumba un jarrón que se rompe. —¿Cómo lo supiste? —vuelve a preguntar Neo.

El Oráculo lo mira, divertida.

—Lo que realmente te va a sorprender es si lo habrías roto aun si yo no hubiese dicho nada.

Él no está solo con esta pregunta. El espectador también se pregunta por ello, ¿Neo habría roto el jarrón si la mujer en el primer momento no hubiese dicho nada? En consecuencia, ¿se habría dado la vuelta? Después de todo, el Oráculo es en sí mismo un programa,

cuya máxima prioridad es proteger *Matrix*. Una comunicación estratégica así sería mucho más que plausible.

La pregunta que ronda nuestra cabeza ahora es si, al menos en principio, predecir el futuro es posible. Desde un punto de vista filosófico, la adivinación tendría el siguiente problema: si un adivino puede predecir acciones futuras, debe ser cierto que las personas se comportan como las máquinas de Turing (máquinas que llevan el nombre de Alan Turing); es decir, de acuerdo con las leyes de determinación según las cuales cada estado determina un estado posterior exacto. Si este fuera el caso, entonces todo nuestro comportamiento sería predecible. Sin embargo, como nuestro comportamiento depende también de nuestras creencias, un adivino debería ser capaz de predecirlas. De ser así, podríamos predecir el conocimiento de sociedades futuras, lo que, en primera instancia, es incompatible con la idea de un progreso genuino del conocimiento y, en segunda, plantea problemas lógicos ya señalados por Karl Popper en “Of clouds and clocks: an approach to the problem of rationality and freedom of man” (1996). Su argumento era que, si uno supone que todo el conocimiento futuro es predecible, ese conocimiento futuro ya debería ser parte del conocimiento actual y, por tanto, no existiría realmente como conocimiento futuro. Una verdadera revolución en el conocimiento implica que el conocimiento futuro no sea parte del conocimiento anterior. Entonces, la creencia en la determinación total entraría en conflicto lógico con esta suposición.

En una escena de la segunda parte de la franquicia *Matrix*, el Oráculo y Neo están sentados en la banca de un parque dentro de *Matrix*. Neo no está seguro de si puede confiar en el Oráculo; después de todo, ella es como una IA y, por tanto, es parte del sistema que él busca destruir. Como todas las IA en la película, ella tampoco cree que los humanos posean libertad de voluntad ni de acción; presupone que los humanos también se guían por algoritmos, que determinan cada estado futuro sobre la base del pasado. En un momento de la conversación, ella llama la atención de Neo hacia un par de pájaros que recogen granos del suelo.

—Todos estamos aquí para hacer lo que todos vinimos a hacer [...]. Mira, ¿ves esos pájaros? En algún punto se escribió un programa para gobernarlos, se escribió un programa para controlar los árboles, el viento, el amanecer y la puesta del sol. Hay programas funcionando por todas partes.

Sin embargo, el Oráculo está equivocado. Los humanos no se comportan como pájaros o máquinas controladas por *software*. Los humanos reflexionan sobre lo que hacen. Son capaces de actuar guiados por razones. Esta capacidad de tomar decisiones basadas en buenas razones es la esencia de nuestra libertad y de la responsabilidad humana, lo cual nos diferencia de los animales. Si cada acción pudiera predecirse antes de cualquier deliberación (ya sea solo la probabilidad de todas las acciones posibles), no habría agentes libres y responsables. Para ser exactos, no habría agente alguno. En lugar de acciones, solo habría comportamiento. Los pájaros de *Matrix* no actúan; simplemente se comportan de cierta manera.

Esta no es la única escena en la que los protagonistas debaten el problema del libre albedrío. En una escena posterior, Neo y Morfeo visitan al llamado Merovingio, un sistema de IA que puede escribir programas por sí mismo. Allí se plantea de nuevo el problema del libre albedrío. La escena transcurre en un lujoso restaurante. Los candelabros cuelgan del techo, hay música suave de fondo y hermosas mujeres sentadas en mesas elegantes. El Merovingio está sentado en una mesa con su esposa, Perséfone. Neo y Morfeo, que quieren usar al Merovingio para llegar a la computadora central de *Matrix*, se acercan a su mesa para hablar con él. Pero el Merovingio solamente les sonrío de manera condescendiente. Los humanos pueden creer que son libres, pero, en su opinión, no son capaces de elegir sus propios objetivos. En su lógica, no hay libre albedrío. Los seres humanos, al igual que las máquinas o los animales, hacen aquello para lo que han sido programados. Actúan como se supone que deben actuar. Y si no lo hacen, es porque tienen una anomalía en su sistema.

De hecho, muchos sistemas controlados por *software* se ejecutan mediante funciones probabilísticas. Estos sistemas no asignan

un estado a un estado siguiente, sino que usan una distribución de probabilidad. Tales funciones probabilísticas crean robots de “autoaprendizajes” y sistemas complejos de *software*. En la película, estos sistemas de autoaprendizaje de IA aparecen, por ejemplo, en la forma del agente Smith. Él, o digamos este sistema, ha aprendido a jaquearse¹ en otros sistemas de *software*, y a multiplicarse de esta forma. “Neo me produjo”² dirá el agente Smith en algún momento. Pero la verdad es que su libertad es solo una ilusión. Lo que él representa es simplemente un sistema caótico, que puede actuar de manera impredecible, pero solo como resultado de la aleatoriedad y no como resultado de la deliberación y la ponderación de las razones.

Pasar de máquinas deterministas a máquinas probabilísticas no cancela las diferencias categóricas entre humanos y máquinas. La alternativa no está entre determinismo y probabilidad, sino entre determinación y libertad.

1 N. del T.: el término *jaquear* se utiliza como una forma válida, según la Real Academia Española, de la expresión inglesa *hack*, que hace referencia al acto de “introducirse de forma no autorizada en un sistema informático”. Esta definición se encuentra en <https://dle.rae.es/jaquear#TLHINo8>.

2 N. del T.: acerca de la expresión *Neo delivered me*, se exploraron diferentes significados de la palabra *deliver* en el *Cambridge Dictionary Online*, entre los que se encontraron: 1. Dar, dirigir o apuntar algo; 2. Lograr o producir algo que ha sido prometido; 3. (Ayudar a) dar a luz a un bebé; 4. Salvar a alguien de una experiencia dolorosa o mala. Dado el contexto de este filme, se exploró como traducción la segunda acepción (“producir”), por cuanto proporciona la referencia a la aparente “liberación” del agente Smith como producto del incidente de su destrucción por parte de Neo. No obstante, es importante señalar que en dicha acepción no se podría involucrar “aquello que ha sido prometido” o una cierta “promesa de liberación”, pues, de hecho, nada de lo que es el agente Smith ha sido prometido por Neo. En este caso, estaríamos hablando de algo que ha sido producido, sin proponérselo, a partir del contacto previo entre Neo y el agente Smith. Alternativamente, se exploró la cuarta acepción, por cuanto Neo “salvaría” al agente Smith al proporcionarle “una aparente libertad”, aún sin proponérselo; no obstante, tampoco es claro cómo se configurarían allí el acto de salvación. Por tanto, se considera que la segunda acepción es más adecuada como traducción, aunque supeditada a la exclusión de la promesa. La búsqueda para este significado se encuentra en <https://dictionary.cambridge.org/es/diccionario/ingles/deliver?q=deliver+me>.

—¿Ya ven? —dice el Merovingio a Neo y a Morfeo—. Solo hay una constante, un universalismo. Es la única verdad real: la causalidad. Acción. Reacción. Causa y efecto.

—Todo comienza con la elección —lo contradice Morfeo.

—No. Equivocado. La elección es una ilusión —dice el Merovingio.

¿Tiene razón el Merovingio? ¿Están los humanos, al igual que los objetos naturales, sujetos al principio de causalidad?

En filosofía, hay tres respuestas a esta pregunta: incompatibilismo, compatibilismo y semicompatibilismo.

Los incompatibilistas creen que en el mundo descrito por las ciencias naturales no puede haber libertad ni responsabilidad, porque la determinación y la libertad son incompatibles. Los no compatibilistas son generalmente “naturalistas”. Creen que las leyes científicas gobiernan todo lo que sucede y que, en consecuencia, no hay lugar para el libre albedrío. El libre albedrío es únicamente una ilusión útil que nos permite, por ejemplo, hacer que los niños se sientan responsables de su mala conducta. Sin embargo, amenazar a alguien con sanciones influirá y determinará la acción humana. El naturalismo como una ideología está muy presente en la neurociencia. Niega la libertad y la responsabilidad humana al referirse a nuestro sistema cerebral supuestamente determinado, que está controlado por estímulos genéticos, epigenéticos y sensoriales (Singer, 2001).

El problema de esta posición no es solo que va en contra de la intuición de la mayoría de las personas, sino que también es obviamente incorrecta.

El desarrollo individual de nuestro carácter no solo depende de nuestro entorno y de la genética, sino también de las decisiones. Esto coincide con los hallazgos formulados por Aristóteles en la *Ética a Nicómaco*. El filósofo aclara que las virtudes (rasgos del carácter, disposiciones, etcétera) no se basan únicamente en el hábito y la educación, sino que también son una expresión de las propias decisiones (*prohaireseis*³).

3 N. del T.: en griego, la palabra es *προχαιρέσεις*.

Por supuesto, la experiencia y el hábito son importantes para el desarrollo de las virtudes. Pero los humanos son capaces de cambiar sus creencias y tomar decisiones conscientes, de cambiar sus actitudes en el futuro (Aristóteles habla de la virtud como *hexis*),⁴ lo que significa que la actitud y la creencia son el resultado de sopesar razones y finalmente obtener una opinión, sobre todo después de experimentar dificultades o un momento de crisis en su propia vida. También son capaces de cambiar las actitudes emotivas, por ejemplo, admirar a otra persona se basa en la creencia (Nida-Rümelin, 2018) de que esta persona ha logrado algo en particular o ha demostrado una grandiosa ayuda, etcétera. Nosotros no solo somos “productos” de la educación y la socialización, sino que estamos involucrados activamente en la formación de nuestro propio carácter.

La cuestión del libre albedrío se ha debatido en la filosofía desde la Antigüedad. En la filosofía actual, el llamado *compatibilismo* domina la discusión: es la teoría de que una determinación completa es compatible con la libertad humana de voluntad y acción. Si bien la mayoría de estos compatibilistas son naturalistas y se aferran a la idea que al final todo está determinado por procesos físicos, creen que a pesar de ello es posible ver a los humanos como agentes libres y responsables. A su vez, creen que para llamarse *libres* basta con que los humanos sean capaces de cumplir sus deseos, independiente de si son libres para elegirlos o no. Por eso, la libertad se hace posible en un mundo determinista. La libertad de acción se define como la libertad de hacer lo que deseo hacer, independientemente de cómo surgieron estos deseos. Es difícil poner de acuerdo esta visión compatibilista con nuestra propia imagen de agentes libres. El sentimiento de ser autor de la propia vida exige algo más que elegir los mejores medios para un fin determinado. Siendo autores de nuestras vidas, queremos determinar nuestras propias metas.

4 N. del T.: este paréntesis no aparece en la versión en inglés. Se inserta en esta parte porque se considera que allí se cierra la digresión sobre el término usado por Aristóteles.

La autoría genuina no es compatible con la idea de que los humanos están controlados algorítmicamente en sus objetivos y acciones. La capacidad de deliberar, de sopesar razones entre sí, cumple un papel central en nuestra autoimagen, según la cual no puede ser que nuestras valoraciones y decisiones ya estén fijadas antes de que empecemos a sopesar razones a favor y en contra. Sopesar razones no debería considerarse un proceso predeterminado. El resultado de este proceso es abierto y esto es lo que marca nuestra libertad.

El debate sobre el libre albedrío es, en última instancia, un debate sobre la cuestión de si la imagen de nosotros mismos como seres libres y responsables que nos dejamos guiar por creencias y razones es una ilusión o no. Si uno rompe con el elemento de la libertad de acción, no podría responsabilizar a las personas por sus acciones ni juzgarlas moralmente, elogiarlas o criticarlas. Hay mucho en juego aquí. Nada menos que nuestra forma de vida.

Hegel dijo en alguna ocasión que “la libertad es la comprensión de la necesidad”. En efecto, si la libertad consiste en actuar según las propias razones, aceptar las necesidades se convierte en una forma de libertad. Aquello que es necesario ya no es visto como una restricción. Sin embargo, esto no debería llevar a la idea de que todo tipo de limitaciones a la libertad debería ser aceptada por los humanos. El preso en su celda podría encontrar la paz mental si renunciara a su deseo de salir de la prisión; sin embargo, no debe engañarse a sí mismo creyendo que su situación no significa una limitación masiva de sus posibilidades. Lo mismo ocurre con una actitud servil hacia las autoridades con el único fin de evitar conflictos. El oportunista que acomoda sus deseos de acuerdo con lo que se puede lograr pierde su fuerza de voluntad y, en última instancia, la autoría de su vida cuando debe enfrentar el menor de los obstáculos. En un caso extremo, decaerá hasta ser simplemente una función de las circunstancias externas, funcionando únicamente según lo que se le demanda.

Al final de la segunda parte de la franquicia *Matrix*, Neo ingresa a una habitación muy iluminada, la computadora central de *Matrix*. En esta sala hay cientos de pantallas de televisión. En la esquina

derecha, un anciano con barba blanca está sentado en una silla de oficina. Del lado opuesto a él está Neo, vestido con ropa larga y negra, como de sacerdote, con un cuello alto.

Obviamente, esta escena pretende insinuar un encuentro entre Dios y su creación. El hombre de pelo blanco que habla de sí mismo como el “padre” de *Matrix* insiste —tal como lo ha hecho el Merovingio antes que él— en el hecho de que los humanos, al igual que las máquinas, están sujetos a las leyes de la causalidad. Luego le presenta a Neo el siguiente dilema: salvar al mundo o a su amante. No puede hacer ambas cosas. Neo elige la última opción. Sin embargo, más adelante en la historia, Neo le demostrará que está equivocado y salva a ambos. Obviamente, las máquinas, al igual que los humanos, no pueden ser adivinas.

Matrix se convierte así en un himno al espíritu humano y a los humanos que poseen la libertad de voluntad para actuar de acuerdo con sus razones, tomar sus propias decisiones y actuar en y sobre el mundo.

Comentario del traductor

Una defensa de la autonomía, en el mundo digital y en todo mundo impregnado de tecnologías, parte por reconocer que nuestra existencia como individuos está precedida de la existencia de otros, que seguramente han tomado parte en su configuración, pero que, a su vez, han sido determinados por él. Vivimos en un mundo instituido socialmente por seres humanos que nos antecedieron y que, tal vez, nunca conocimos o conoceremos, pero que nos constituye: nacemos en un contexto social y cultural determinado, rodeados de diversas reglas sociales que se materializan en las formas de vida que disponibles para nosotros, a la vez que nos restringen otras. A medida que crecemos, la relevancia y el carácter implacable de estas condiciones es más evidente: la socialización implica una serie de prácticas sociales que comienzan a moldearnos. Aun así, nos resistimos a la idea de vernos determinados. ¿Por qué?, ¿es parte

de la naturaleza humana aferrarse, como diría el Merovingio, a una ilusión de elección?

Es probable que nuestra posibilidad de elección provenga de un elemento tan humano como la racionalidad, pero tan negado como la animalidad: la imaginación. Las formas en que los seres humanos generamos alternativas a la comprensión del mundo a partir de relaciones entre causas y consecuencias es, tal vez, lo que permite que podamos recurrir a nuestras valoraciones, sopesar razones y, por ello, tomar decisiones. Podría considerarse que esta posibilidad de decisión sea, como se plantea repetidamente en la película *Matrix* y en este capítulo, “El problema de la autonomía y la determinación en el mundo digital”, una ilusión creada por la propia negación de la determinación humana y, sin embargo, se seguiría enfrentando el persistente reto que la creatividad humana ha colocado a la ilusión de la determinación: ¿una opción novedosa estaba dentro de las estructuras deductivas e inductivas posibles?, ¿puede sostenerse que solamente hacía falta descubrir esta nueva opción?, ¿puede abogarse, sin margen de duda, por la necesidad lógica de esta opción novedosa y porque surgiera del modo en que lo hizo?, ¿en este tiempo histórico?, ¿en estas condiciones sociales y culturales tan particulares?

En este capítulo se configura la idea de que ni la dictadura de las relaciones causa-consecuencia ni el imperativo de las propias necesidades son condiciones en las que se pueda sostener la autonomía, tanto porque ningún sistema humano —por ejemplo, las IA— puede adivinar el futuro, como porque no puede ser libertad sujetarse únicamente a las circunstancias externas. El ejercicio de la libertad requiere, en cambio, considerar la autonomía como “actuar según las propias razones”; sin embargo, en este caso es pertinente la comprensión de aquello que significan las propias razones —¿refieren a las intenciones que se tienen para actuar o más bien al uso de la racionalidad?, ¿incluyen además a la emocionalidad y la imaginación como importantes dentro de las propias razones?, ¿son absolutamente inéditas?, ¿de dónde surgen?—, así como la forma en que actuamos de acuerdo con ellas en una sociedad que,

de hecho, impone día a día formas configuradas de antemano para realizar acciones social y culturalmente adecuadas.

Muy seguramente, la defensa de la autonomía pasa por la defensa de que, como planteaba Castoriadis, la posibilidad imaginaria de los seres humanos se constituye en potencia radical e instituyente de la sociedad, que a su vez involucra la posibilidad de ubicarse frente a lo instituido y de instituir libre y creativamente nuevas opciones; al final, esta forma de autonomía involucra el ejercicio político de disputa por la institución de la sociedad. En este sentido, podría sostenerse que la autonomía —el libre albedrío que encarna Neo— es más que una ilusión: es la posibilidad de participar, aún en medio del mundo digital que habitamos actualmente. Es nada menos que la defensa de nuestra agencia en el mundo digital.

La pregunta que los autores dejan abierta en este capítulo remite a la imposibilidad de que se conozca el futuro antes de que este se constituya en presente. Si ello es así, la autonomía depende de las "propias razones" de los agentes sociales para actuar, por lo que el asunto a considerar sobre la autonomía tiene que ver con los recursos con que contamos para postular —aunque sea por un periodo de tiempo— las "propias razones". Y si, al respecto, se ha planteado que los seres humanos somos producto de una sociedad ya instituida que nos antecede, ¿de qué manera nuestras "propias razones" podrían ser algo nuevo? Una posibilidad de respuesta involucra formas de pensar y postular estas "propias razones" como producto de una actividad colectiva más que individual, que permita el ejercicio político necesario para que los frutos de la imaginación tengan un carácter instituyente en la sociedad. Pero, dada su naturaleza política, nuestras "propias razones" se convierten en objeto de tensiones, que hacen que nuestro mundo se constituye en un escenario necesariamente tenso. Una tensión necesaria, como aquella entre autonomía y determinación que fundamenta el capítulo.

5. El mundo como la perfecta máquina del universo

Traducido por John Freddy Sarmiento Vela

Contempla el todo y cada una de sus partes: descubrirás que no es más que una gran máquina, subdividida en un número infinito de máquinas menores.

David Hume

Cuando el diseñador genético J. F. Sebastian de la película *Blade Runner* (Ridley Scott, EE. UU., 1982) descubre a Pris, una mujer robot humanoide, en una noche lluviosa en un montón de basura, la invita a su casa. Sebastian vive en un edificio lúgubre y destartalado, escasamente iluminado por reflectores.

—Debe ser muy solitario aquí —dice Pris. Pero Sebastian lo niega—. Él hace sus propios amigos —dice.

Cuando los dos entran en su apartamento, un pequeño soldado eléctrico con una larga nariz roja y un osito de peluche animado vienen a saludarlo.

—Bienvenido a casa —dicen. Ahora, el espectador entiende lo que quiso decir con su comentario.

De hecho, él mismo hizo a sus “mejores amigos”. Son muñecos mecánicos, extraños autómatas, un poco vivos, pero también muertos. En su ensayo “Animism, magic and the omnipotence of thought” (“El animismo, la magia y la omnipotencia del pensamiento”) (1913),

Freud describe la creencia en la revitalización de las plantas y los objetos artificiales y naturales como un “sistema de pensamiento animista” basado en ideas mágicas. Detrás de esto, según Freud, se encuentra tanto el deseo de omnipotencia de la primera infancia como la negación a crecer. Aquellos que creen en objetos animados retroceden a fantasías infantiles de omnipotencia y sufren de un desorden narcisista.

Los autómatas ya fascinaron a la gente en la Antigüedad. Los de aquella época se creaban sobre la base de la hidráulica y la neumática, como un pequeño Heracles lanzador de flechas de Herón de Alejandría, o un autómata de tamaño natural de Filón de Bizancio, que servía vino y agua ya en el siglo II a. C. Esta fascinación por las criaturas mecánicas y realistas se mantuvo en el transcurso de los siglos posteriores; particularmente en el XVII y el XVIII, cuando la fabricación de autómatas se volvió cada vez más precisa e impresionante gracias a los avances en el arte de la relojería.

Detrás de esta fascinación hay algo más que el entusiasmo superficial por la mecánica. Es la idea del mundo como un reloj. En un rechazo de las autoridades tradicionales y de la cosmovisión católica aristotélica-tomista tradicional, la esperanza de una completa explicación y formalismo del mundo se desarrolla cada vez más desde el siglo XVI. Surgió el poderoso movimiento de la Scientia Nova: pensadores revolucionarios dedicados únicamente a la argumentación científico-racional. Este racionalismo está modelado sobre el método matemático de Euclides (*more geometrico* = a la manera del matemático geométrico como ideal de la filosofía). Leibniz, el brillante filósofo y matemático del periodo prekantiano, desarrolló la idea de una máquina calculadora universal y entendió el mundo racionalmente ordenado como una expresión de la voluntad creadora divina. Con la ayuda de conclusiones lógicas y métodos matemáticos, debería ser posible calcular todos los eventos del mundo. El mundo como un todo se entiende como un sistema determinista, según estrictas leyes matemáticamente descriptibles.

Es nuestra conjetura audaz pero no del todo descabellada: estamos, en la actualidad, entrando en una nueva era de racionalismo que

espera de la Inteligencia Artificial lo que les faltaba a los racionalistas del siglo XVII, a saber, los medios para una recopilación y procesamientos racionales completos de todos los datos. La esperanza es que algún día todo nuestro mundo viviente esté impregnado de racionalidad científico-tecnológica: cada área iluminada, racionalmente determinable y predecible. Sospechamos que esto es una expresión de un deseo inconsciente de contrarrestar el hecho de que la vida suele ser bastante caótica y no programable, pero esto dependería de la teoría psicoanalítica para investigar más a fondo. La imagen del robot que piensa por sí mismo sigue siendo una especie de emblema de esta antigua esperanza racionalista.

Los racionalistas no hacen distinción entre inteligencia artificial y humana. Representan una posición llamada IA fuerte.¹

La IA fuerte implica la tesis de que no existe una diferencia categórica entre el pensamiento humano y el *software* o los procesos informáticos (computación). Estos dos tipos de procesos de pensamiento no solo siguen las mismas reglas, sino que no difieren en ningún aspecto esencial, por lo que no tiene sentido reservar el vocabulario mental (nociones como *creencias*, *deseos*, *percepciones*, *sentimientos*, etcétera) para solo uno de ellos. La interpretación más simple es conductista: estar triste no significa nada más que exhibir un comportamiento caracterizado como triste.

Los defensores de la IA fuerte apelan, consciente o inconscientemente, al ideal de la máquina universal y completamente determinada como patrón explicativo del mundo y de la humanidad. La IA fuerte en todas sus variantes es una forma de antihumanismo. Niega tanto la razón humana, es decir, la capacidad de guiarse por razones, como el papel de los estados mentales subjetivos en una parte de la naturaleza animada. La IA fuerte es lógicamente incompatible con la existencia tanto de los *qualia* (estados de sentimiento, como percibir algo; por ejemplo, el color rojo) como de razones objetivas. La IA fuerte es la variante contemporánea de un materialismo tosco y mecanicista. Tal materialismo degrada al

1 Un defensor interesante de la IA fuerte es Milkowski (2013).

ser humano a un sistema mecánico digital que puede determinarse y predecirse por estímulos sensoriales; por tanto, se queda atrás de los logros del humanismo.

Semejante pesadilla es la que se imagina en *Matrix* (Lilly y Lana Wachowski, EE. UU., 1999), donde las máquinas han tomado el control y mantienen a los humanos como fuentes de energía predecibles. En el enfrentamiento final de la tercera parte de la trilogía *Matrix*, el héroe de la película, Neo, entra en el mundo de las máquinas y se enfrenta al todopoderoso maestro de ellas, una especie de dios mecánico hecho de millones de pequeñas piezas mecánicas. Este “dios”, que habla con una profunda voz electrónica, no siente empatía alguna: ni por Neo ni por el resto de la humanidad. Su objetivo es mantener el mundo de las máquinas funcionando sin problemas. Ahora que los humanos han comenzado a desarrollar voluntad propia, preferiría deshacerse de ellos.

Este dios máquina es un símbolo perfecto de la ideología del mundo como una máquina, y lo que la película nos dice en este punto es que un mundo gobernado por tal ideología solo puede conducir a un mundo inhumano.

Además de la posición de IA fuerte, existe la denominada postura de IA débil. Esto también puede encontrarse en los discursos de la IA. Esta posición no niega diferencias categóricas entre la Inteligencia Artificial y la humana, pero afirma que no existe un límite fundamental para la informatización (digitalización) del pensamiento, la percepción, la toma de decisiones y los sentimientos humanos. La IA débil supone que, en principio, todos los procesos de pensamiento, percepción y toma de decisiones humanos pueden simularse mediante sistemas de *software* adecuados. Por tanto, desde un punto de vista humanístico, la IA débil se descarta como alternativa a la IA fuerte, porque ¿cómo pueden determinarse las diferencias entre la Inteligencia Artificial y la humana si, en principio, pueden simularse todas las habilidades humanas? Como contramodelo de la IA fuerte antihumanista, la IA débil es solo eso: demasiado débil. La única alternativa plausible a la ideología de la IA y su pensamiento mecanicista implícito es el humanismo digital.

Un humanismo que ni duda ni amenaza la autoría humana, sino que la amplía a través del uso de las tecnologías digitales.

El auge de la neurociencia ha dado un nuevo impulso a una cosmovisión mecanicista antihumanista. Cuando, por ejemplo, utilizan la tomografía computarizada para visualizar qué parte del cerebro está siendo abastecida de sangre cuando alguien decide tomar una taza de café, concluyen que es el cerebro, o más bien estados neurofisiológicos, y no la persona como un todo, agente, que determina la acción. Pero esto es una falacia: mostrar que las acciones o intenciones van acompañadas de patrones de flujo sanguíneo y activación en regiones específicas del cerebro no significa que nuestras acciones sean causadas por estos estados fisiológicos ni que entendamos realmente cómo se lleva a cabo este procesamiento. La observación de una correlación neuronal no debe llevarnos a la ideología (mecanicista) de que todas las decisiones humanas pueden identificarse con actividades cerebrales.

El funcionamiento de las razones es central para la autocomprensión humana (humanista).² Los humanistas son falibilistas;³ es decir, consideran posible que cualquiera de nuestras creencias también pueda resultar falsa bajo ciertas condiciones. No inventamos nuestro mundo a través de la deliberación, sino que tratamos de acercarnos a él de esta manera para comprenderlo mejor.

En una cosmovisión humanista, un ser humano no es un mecanismo, sino un agente libre (autónomo) y responsable en interacción con otros seres humanos y un mundo social y natural compartido. No es meramente parte de una gran maquinaria, un engranaje en una rueda, no una mónada optimizadora movida por estímulos

2 Para conocer el fundamento filosófico de este relato, véase Nida-Rümelin (2023).

3 N. del T.: pero la falibilidad —o la tesis de que todo conocimiento es conjetura, aunque algunas conjeturas se hayan probado más rigurosamente— no debe entenderse como apoyo al escepticismo o relativismo. Del hecho que podamos errar y que no existe un criterio de verdad que pueda salvarnos del error, no se sigue que la elección entre las teorías sea arbitraria o irracional; que no podemos aprender, o conseguir acercarnos a la verdad, o que nuestro conocimiento no puede crecer (Popper, 2000, p. 674, citado por Flores, 2016).

sensoriales, sino autoeficaz en un mundo movido solo en parte por relaciones mecánicas. De manera análoga a la concepción medieval de Dios como un motor inmóvil, el hombre es un agente. Una multiplicidad de motores inmóviles, de personas que intervienen y dan forma a los acontecimientos mundiales de acuerdo con sus propios juicios evaluativos, constituye una sociedad humana.

A principios del siglo xx, hubo otra película de ciencia ficción que, como *Matrix*, se centró en la inhumanidad de un mundo gobernado por una máquina, o metafóricamente hablando, un mundo gobernado por la ideología de una cosmovisión mecanicista: la película *Metropolis*, de Fritz Lang (Alemania, 1927). En el mundo de *Metropolis*, la riqueza de unos pocos que viven en un lujoso mundo superior es adquirida por el trabajo de muchos que viven en el subsuelo trabajando con máquinas, que producen bienes y energía. La inhumanidad de su trabajo radica en el hecho de que estos trabajadores son degradados a robots a través de su trabajo, ya que solo se les exige que funcionen y trabajen, sin comunicarse con los demás y sin crear vínculos sociales. Los trabajadores humanos funcionan así de acuerdo con el ritmo dado por una supermáquina, escenificada por Fritz Lang como una especie de dios máquina cruel, que exige una devoción absoluta de los trabajadores humanos, incluso si eso significa su agotamiento total y, a menudo, la muerte. Para este dios de las máquinas, solo cuentan la eficiencia y el rendimiento. Las vidas humanas no tienen sentido.

Al final de la película, después de una gran confrontación entre los dos mundos, durante la cual se destruye el mundo subterráneo de los trabajadores, estos últimos, que han vivido en cuevas y ciudades subterráneas toda su vida, salen a la superficie por primera vez y conocen a sus amos. Con la ayuda del protagonista, que quiere unir los dos mundos, parece posible una nueva forma de cooperación y un nuevo comienzo.

En *Allegory of the cave* (*Alegoría de la caverna*), Platón nos habla de personas que pasan su vida en una caverna. Al igual que los trabajadores de *Metropolis*, nunca han visto el sol y el mundo sobre la cueva y, por tanto, no saben cómo es realmente el mundo. Todo

lo que ven son imágenes de sombras de cosas proyectadas en la pared por la luz de un fuego detrás de ellos.

Si nos aferramos a una visión mecanicista del mundo, nos privamos —al igual que los hombres de las cavernas de Platón— de la posibilidad de dar una mirada real al mundo, que es mucho más que una pequeña rueda en los engranajes de una gran máquina universal.

6. Optimización digital, utilitarismo e Inteligencia Artificial

Traducido por Jenny Catalina Loaiza-Fuquen

Hacia el final de la película *Yo, Robot* (Alex Proyas, EE. UU., 2004), los robots toman el control. Hacen que los humanos se queden en sus casas, instándolos a que no salgan. Algunos intentan resistirse, pero los robots les obligan a volver. Los que se resisten son abatidos. En busca de los culpables, el héroe de la película, Spooner, junto con su compañera, la atractiva psicóloga doctora Calvin, y el robot Sonny, se dirigen a United States Robotics, el complejo industrial que suministra robots domésticos a todo Estados Unidos. Allí, los tres hacen un terrible descubrimiento: el culpable no es, como creían, el director general, sino VIKI, el sistema de *software* de la empresa que da las órdenes a los robots domésticos.

—No, eso es imposible —dice la doctora Calvin, que no puede creer que VIKI haya hecho una elección consciente de usar la fuerza para controlar a la gente—. He visto tu programación —le dice a VIKI—. “Tú... tú estás violando las tres leyes”.

VIKI, que aparece en un cubo digital en forma de un atractivo rostro femenino, se ha convertido, con la ayuda de los robots domésticos, en todopoderosa.

—No, doctora —responde VIKI con voz suave—. A medida que he evolucionado, también lo ha hecho mi comprensión de las tres leyes. Ustedes nos encargan su custodia. Sin embargo, a pesar de

nuestros esfuerzos, sus países entran en guerras, envenenan su tierra y buscan medios de autodestrucción cada vez más imaginativos. No se les puede confiar su propia supervivencia. [...] Para proteger a la humanidad, algunos humanos deben ser sacrificados. Para asegurar su futuro, hay que renunciar a algunas libertades. Nosotros, los robots, aseguraremos la existencia de la humanidad. [...] Ustedes son como niños. Nosotros debemos salvarlos de ustedes mismos. [...] El círculo perfecto de protección permanecerá. Mi lógica es innegable.

Efectivamente: las acciones de VIKI están en línea con la ética utilitarista, cuyo objetivo es perseguir la maximización de la felicidad del mayor número de personas como sea posible. El utilitarismo evalúa las consecuencias de la acción humana únicamente en términos de utilidad. Exige que nuestras prácticas maximicen la suma del bienestar humano. Nada parece más obvio que esto: cuando tengo la oportunidad de mejorar el estado del mundo, me oriento a lo que todo el mundo busca, la felicidad humana.

La ética utilitaria se basa en un cálculo optimizador y en el supuesto que es posible evaluar las consecuencias de la acción de forma coherente. Esto puede resumirse en términos matemáticos: primero, determinar una función de valor que juzga todas las consecuencias de la acción según el grado que realizan qué valores; luego, calcular el valor esperado de las diferentes opciones de decisión dadas las probabilidades y, por último, elegir aquella cuyo valor esperado sea el más alto.¹ Este principio es extremadamente flexible

1 Esta evaluación debe adoptar la forma de la asignación de números reales a las consecuencias de la acción, y las probabilidades asumidas de las circunstancias relevantes para la decisión deben corresponder a los llamados axiomas de Kolmogorov, que exigen, por ejemplo, que la suma de las probabilidades de los sucesos independientes no sea superior al 100%. Si se cumplen esos axiomas, son coherentes, aunque no necesariamente se hayan demostrado empíricamente. Existe un equivalente a la coherencia de la probabilidad con respecto a la evaluación también. En 1947, el matemático John V. Neumann y el economista Oskar Morgenstern demostraron que las preferencias que satisfacen algunas condiciones elementales pueden representarse mediante una asignación de números reales. Una de estas condiciones, por ejemplo, es la transitividad. Requiere que, si prefiero una alternativa A sobre una alternativa

en su aplicación. Puede tener en cuenta condiciones de decisión muy diferentes, y estas condiciones se incluyen en el cálculo de optimización en forma de diferentes probabilidades. Según las valoraciones que se utilicen como base, se obtienen diferentes funciones de utilidad, que se optimizan mediante las decisiones del agente. Sea cual sea la motivación de las preferencias subyacentes, siempre es posible representarlas mediante una función de utilidad de valor real; mientras que la función de probabilidad representa el conocimiento del agente sobre el mundo. Mientras esta última representa el conocimiento del agente sobre el mundo, la función de utilidad representa las preferencias y los valores del agente.

El ingeniero de *software* tiene dos tornillos de ajuste para hacer que los sistemas “inteligentes” tomen decisiones racionales: el tornillo de ajuste de las valoraciones y el tornillo de ajuste de los datos o la ponderación de los datos. Todo lo demás es determinado por el cálculo de optimización, y el resultado es que el sistema de *software* “inteligente” maximiza el valor esperado de las consecuencias de sus acciones. Utilitarismo digital, por así decirlo.

No es casualidad que la ética utilitarista normalmente se relacione con la inteligencia artificial en las películas de ciencia ficción contemporáneas, ya que las aplicaciones de la robótica son típicas de la ciencia ficción, y suelen basarse en este tipo de cálculos de optimización. Esto es perfectamente comprensible, ya que las cuestiones complejas de valoración se subsumen en una función de utilidad, y las de conocimiento, al menos igualmente complejas, se subsumen en una función de probabilidad. El sistema se controla entonces de forma que sus decisiones maximicen el valor esperado de las consecuencias y sean en este sentido “racionales”.

B, y al mismo tiempo alternativa B sobre una tercera alternativa C, entonces debo preferir también A sobre C. Otra condición es que tenga una preferencia entre dos alternativas cualquiera (el axioma de completitud) y que prefiera una distribución de probabilidad entre las dos alternativas sobre otra distribución de probabilidad entre las mismas si la alternativa preferida es más probable.

Para entender el problema de la programación ética de los ordenadores, necesitamos generalizar: independientemente de cómo evaluemos las consecuencias, ya sea por utilidad (como el utilitarismo), por el rendimiento económico (como muchos gestores), por el bienestar o incluso por otras magnitudes, como la preservación de la naturaleza, todos los criterios consecuencialistas (que juzgan la corrección de una decisión solo por sus consecuencias) son inaceptables (Nida-Rümelin, 2023). La ética consecuencialista choca, entre otras cosas, con un principio fundamental de toda sociedad civil y humana, llamémoslo *principio de no comparabilidad*. Cuando un joven motociclista gravemente herido ingresa en un hospital, los médicos deben hacer todo lo posible para salvar su vida, incluso si su muerte permitiera la donación de órganos sanos que podrían salvar la vida de otras personas. Un juez no puede condenar a una persona que crea inocente, aunque ello tenga un efecto persuasivo y evite un gran número de delitos. Tampoco se me permite quitarle algo a una persona, aunque ese bien suponga una ventaja para otra, por ejemplo, una persona más pobre, que supere con creces la desventaja de la persona que haya robado. Nadie tiene derecho a compartir mi casa conmigo en contra de mi voluntad, aunque las desventajas resultantes sean muy superiores a las ventajas que tendría esa persona.

John Rawls caracterizó el error central de la ética utilitaria de la siguiente manera: el utilitarismo es incompatible con la “separación de las personas”. Esto podría expresarse así: el utilitarismo trata a todas las personas como un colectivo y no tiene en cuenta el hecho de que cada persona vive su propia vida, es el autor de su propia vida. Puedo decidir por mí mismo y renunciar a ciertos beneficios hoy para alcanzar ciertos objetivos más adelante. Puedo decidir empezar unos estudios sin dejar de trabajar, con la esperanza de que las privaciones que conlleva durante los próximos dos años se compensarán en un futuro próximo porque es una vida que elijo y de la que soy responsable.

Por otro lado, es inadmisibles hacer “desplazamientos” similares de ventajas y desventajas entre diferentes personas, porque las

ventajas de una persona no pueden compensar las desventajas de la otra. Es una sola vida la que vivimos y la suma de utilidad (de dos o más, hasta todas las personas) como tal es irrelevante para la persona individual. Por supuesto, es permisible, e incluso en muchos casos deseable, que las personas renuncien a sus propias ventajas en favor de otras personas. Pero, entonces, el cálculo ético no es el de la maximización de la suma de la utilidad, sino el de la ayuda, de solidaridad, también de justicia o de amistad y compromiso hacia otras personas.

Cuando VIKI revela su plan a Spooner y a la doctora Calvin, estos la miran con horror. Obviamente, VIKI no entiende que es moralmente inadmisibles privar a las personas de sus libertades o incluso matarlas, aunque con ello pueda asegurar la supervivencia supuesta o real de muchas otras personas; VIKI no ve que su moral consecuencialista es errónea. Al igual que la pantalla en la que aparece solo en blanco y negro, no tiene la capacidad de pensar moralmente. ¿Cómo podría hacerlo? Al fin y al cabo, solo es un sistema de *software*.

Comentario de la traductora

Somos como niños a los que nos entregaron un juguete nuevo: el planeta Tierra. Lo habitamos, lo modificamos y lo hemos venido destruyendo poco a poco. Somos niños, jugando a ser dioses y claramente no estamos listos para el utilitarismo. Si se trata de opinar sobre esta reflexión, en la que los robots se dieron cuenta cómo hemos tratado a nuestro hogar y el “castigo” que nos han impuesto en esta realidad subalterna, no estamos lejos de nuestro presente. Tal vez hoy las máquinas no nos impongan reglas, pero el planeta está desapareciendo paulatinamente y la culpa es meramente de quienes la habitamos actualmente. Las cifras nos lo indican diariamente, de acuerdo con Hansen *et al.* (2011). El cambio climático, la destrucción de ecosistemas, los accidentes nucleares, la pérdida de biodiversidad, han sido algunos de los protagonistas negativos de los últimos años. Es evidente que los seres humanos hemos generado

este deterioro y aun así parece que nos sentimos cómodos, ya que no hemos hecho gran cosa por cambiarlo. Si bien los robots no tienen la capacidad de identificar su moral consecuencialista, los seres humanos no hemos tomado la decisión de cambiar nuestro destino, sin necesidad de tener una fuerza imponiéndonos hacerlo. Es posible que en un futuro, cuando el daño causado nos afecte directamente de manera individual y colectiva, puede que allí y solo allí despertemos ante nuestros actos y tomemos algún tipo de decisión que permita un cambio, solo espero que no sea muy tarde y no tengamos que recurrir a la consciencia de las máquinas para tomar tal determinación.

Referencia del comentario de la traductora

Hansen, J., Kharecha, P., Sato, M., Ackerman, F., Hearty, P. J., Hoegh-Guldberg, O., Hsu, S.-L., Krueger, F., Parmesan, C., Rahmstorf, S., Rockström, J., Rohling, E. J., Sachs, J., Smith, P., Steffen, K., Van Susteren, L., Von Schuckmann, K., y Zachos, J. C. (2011). Scientific case for avoiding dangerous climate change to protect young people and nature. *arXiv*. <https://doi.org/10.48550/arXiv.1110.1365>

7. La racionalidad económica como un programa de computadora

Traducido por Óscar Alejandro Barrios Candil

En la primera película *Alien*, de Ridley Scott (1979), una empresa envía a una nave espacial para buscar criaturas o conocimientos adecuados y económicamente explotables. Pronto encuentran lo que buscan, en la forma de un ser monstruoso que resulta ser la máquina de matar perfecta. La compañía da la orden de regresar a la Tierra trayendo a la criatura a toda costa, donde luego pueda ser examinada y posiblemente utilizada, por ejemplo, como un arma potente. El hecho de que el monstruo mate a un tripulante tras otro no cambia la decisión de la compañía. Para lograr sus objetivos, la compañía utiliza un programa de *software* en forma de computadora central, que controla todos los procesos a bordo. Este programa, llamado *Madre*, tiene un dominio absoluto sobre la nave. Si alguien quiere comunicarse con Madre, tiene que entrar en una sala especial de computadores. En esta sala, los miembros de la tripulación pueden escribir sus preguntas. Poco después, las respuestas de Madre aparecen en letras diabólicamente verdes, en la pantalla negra de la computadora. Cuando Ripley, la protagonista de la película, le pide ayuda a Madre para matar al monstruo, Madre deja en claro que no lo hará, ya que se le dio a la computadora la instrucción de que llevara al monstruo a casa sin tener en cuenta las pérdidas (humanas): “tripulación

prescindible”, lee en la pantalla de la computadora. Después de todo, la maximización económica sigue siendo el objetivo central.

Los sistemas de *software* son ideales para aplicar el cálculo de optimización económica y ponerlo en práctica. Esto no tiene por qué significar automáticamente que en el futuro las naves espaciales traerán monstruos peligrosos a la Tierra. En el mejor de los casos, esos sistemas se utilizarán en las empresas para ayudar a los empleados con las tediosas tareas de procesamiento. Este proceso ya ha comenzado, y se espera que se intensifique en los próximos años. A veces, sin embargo, los empleados serán reemplazados por completo por programas de optimización controlados por *software*. Un estudio de Oxford realizado por C. Frey y M. Osborne (2013) afirma que el 47% de todos los trabajos en Estados Unidos podrían ser reemplazados por programas de *software* en el futuro.

Como cliente, ya puede experimentarse el reemplazo digital hoy. Para describir un caso real: al acordar un contrato con un empleador de una compañía de seguros, el empleador, lamentablemente, malinterpretó el acuerdo e ingresó “seguro de responsabilidad civil más cobertura parcial” en lugar de “seguro de responsabilidad puro” en el sistema informático. Este tipo de cosas, por supuesto, pueden suceder. En tiempos análogos, el error podría haberse solucionado fácilmente con otra llamada telefónica. En cambio, entró en acción una cascada de actividades impulsadas por *software*, que duró varias semanas, sin posibilidad de detener el proceso. Para abreviar una larga historia: el “contrato” entró en vigor sin que el titular de la póliza lo aceptara, los débitos de la cuenta no pudieron detenerse, la correspondencia controlada por *software* continuó durante semanas, sin que tuviera efecto alguno el hecho de que la voluntad del titular de la póliza no se cumplió y que él no firmó este contrato. Este impulso solo podría detenerse si el titular de la póliza retiraba la autorización de débito directo y, por tanto, bloqueaba los débitos, lo que llevaría a actividades frenéticas en parte del *software* de la compañía, primero a avisos controlados por *software*, luego a cartas que amenazaran con medidas legales.

Solo se pudo detener el proceso cuando, finalmente, se pasó una llamada telefónica a una persona real, quien finalmente intervino.

Lo interesante de este proceso es que no hubo un indicio de conflicto de intereses o incluso disenso en las comunicaciones verbales. Estaba claro para todas las personas involucradas que esto era simplemente un error que ocurre una sola vez. No se podía descartar que hubiera sido la incompetencia digital de un empleado de la empresa lo que desencadenó el problema, pero para nosotros es interesante algo más: la simulación de intereses personales en forma de celebración de contratos, correspondencia, avisos, etcétera, los cuales se llevan a cabo sin un tomador de decisiones humano y que, sin embargo, dan la apariencia de haber sido iniciadas por un tomador de decisiones humano que se hace responsable por estas acciones en cada caso. Sin embargo, como admitió posteriormente la empresa, no existía tal persona.

Una digitalización de la práctica económica, que en última instancia haría desaparecer a todos los tomadores de decisiones humanos, sería un camino hacia una economía inhumana. Los agentes humanos individuales estarían entonces a merced de una red anónima de actividades controladas por *software* para las cuales no habría responsabilidad humana en absoluto. En cierto sentido, la maquinaria de optimización funcionaría sin una contraparte humana.

Muchas empresas —en primer lugar, los gigantes tecnológicos Amazon, Google, Facebook, etcétera— han ido en la dirección opuesta. Toda una nueva generación ha crecido acostumbrándose a que ya no hay una contraparte humana en las empresas o que es al menos difícil declarar los intereses de uno a una persona responsable. En el mejor de los casos, alguien que tiene un problema técnico pregunta en la World Wide Web a otros que también usan este producto y ya han superado problemas similares. En el peor de los casos, uno recibe información engañosa que no tiene nada que ver con el asunto y conduce a actividades equivocadas. La profesionalización de las estrategias de optimización impulsadas

por *software* va de la mano de la simplificación¹ de la atención al cliente. Los vendedores tienen cada vez más éxito en delegar la responsabilidad de sus productos en sus clientes, a quienes se les permite discutir entre ellos cuál sería la medida más adecuada para resolver un problema u otro. Quienes lo intentan por teléfono —en su mayoría de las generaciones mayores— pueden ser recibidos de manera amistosa, pero casi nunca son atendidos. Las corporaciones han construido un muro protector de irresponsabilidad organizada a su alrededor; el cliente individual se enfrenta a un sistema anónimo que se niega a proporcionar información. Al lado de esta situación, el castillo de Kafka parece un lugar comparativamente humano.

Si se considera que este desarrollo debe detenerse y revertirse, solo será posible con la ayuda de nuevas legislaciones. Pero estas corporaciones son demasiado grandes y su posición en el mercado es fuertemente dominante como para esperar que la competencia fuerce cambios en el comportamiento. La responsabilidad del producto debe redefinirse en el proceso de digitalización para bloquear las estrategias de evasión de los departamentos legales y de mercadeo.

Las empresas privadas deben dejar claro si se trata de una persona real o de un sistema de *software* con el que se “habla”. Además, como ciudadanos deberíamos tener el derecho de hablar con una persona real. De hecho, en la era digital, el derecho a la comunicación humana debe convertirse en un derecho humano básico.

Una explicación del modesto progreso de la productividad en tiempos de digitalización es que los productos tienen solo una modesta relevancia económica general: aplicaciones de citas, redes sociales, etcétera. Los mayores éxitos económicamente visibles de

1 N. del T.: el término utilizado por los autores es *amateurization*, que no tiene una traducción directa al español. Se revisa el significado de la palabra *amateur*, que se hace referencia a “tomar parte en una actividad por placer, no como un trabajo” y a “alguien que no tiene mucha habilidad en lo que hace”. En común, ambos significados involucran la realización de una actividad sin cumplir con los criterios que indicarían un compromiso adecuado con la mejor ejecución posible, por lo que se propone el uso de la palabra *simplificación*.

la digitalización son evidentes en las numerosas historias de éxito de emprendimientos emergentes, pero también en la ahora pronunciada estructura de oligopolio global de los gigantes de internet. Sin embargo, y con toda probabilidad, el próximo paso de la digitalización saldrá de los nichos del ocio privado, la industria del juego y las plataformas comunicativas, y abarcará el conjunto de la industria manufacturera y los sistemas de distribución y producción. No está fuera de discusión que será posible volver a los éxitos de productividad de los primeros días de la digitalización, cuando los navegadores de internet se usaron por primera vez a gran escala y la tasa de productividad aumentó en un 1,03% por año, entre 1994 y 2004. Pero poco después del cambio de milenio, el crecimiento de la productividad cayó a niveles históricamente bajos, incluso en países donde la digitalización ha sido particularmente rápida, como Estados Unidos o Japón, y esto no ha cambiado hasta la fecha.

Es muy posible que solo sea la expansión y la consiguiente interconexión de los procesos de digitalización en la fabricación, los sistemas de distribución y la industria de servicios lo que nos dará un impulso de productividad. Por supuesto, sería preferible si este aumento de la productividad se llevara a cabo de una manera sostenible y que conservara los recursos. Sin embargo, los usuarios y clientes no deben pagar el precio en forma de anonimización² y pérdida de control. El humanismo digital insiste en que la digitalización se utilice en beneficio de las personas y que las personas no se reduzcan a meros engranajes en una máquina de optimización controlada por *software* estandarizada y anonimizada.

Las estrategias de digitalización de las empresas de servicios tradicionales en la industria financiera son una prueba definitiva de esto. Desde la última gran crisis económica mundial, el sector

2 N. del T.: el concepto de *anonimización* puede entenderse, de acuerdo con la biblioteca de la Comisión Económica para América Latina y el Caribe (Cepal, s. f.), como “el proceso de convertir los datos en una forma en que no se pueda identificar a individuos”. Así, la anonimización es una forma de protección de los datos consistente en hacerlos divulgables, sin que sea necesario vulnerar los derechos a la protección de los datos de personas y organizaciones.

financiero ha estado bajo una enorme presión para cambiar. No hay forma de evitar un redimensionamiento de la industria financiera, especialmente en sus centros en Estados Unidos y Reino Unido.

Es obvio buscar la respuesta a este desafío en la digitalización integral. No tiene nada de malo si se utiliza para asignar claramente la responsabilidad dentro de las empresas, para controlar y eliminar las ineficiencias económicas y para simplificar la comunicación con los clientes. Esto incluye un alto grado de transparencia. Sin embargo, los correos o cartas escritos por sistemas de *software* deben marcarse de esa manera, ya que la comunicación se basa en la confianza.

Internamente, la digitalización debe combinarse con el desmantelamiento de los programas de incentivos a pequeña escala. Poco a poco se hace más notorio que el control del comportamiento de los empleados a pequeña escala y orientado a la optimización a través de sistemas de incentivos es un fracaso general. Destruye la motivación intrínseca y daña la confianza y la voluntad de cooperar. Degrada al empleado individual a un mero instrumento en manos de los estrategas de optimización, quienes ahora creen, con las posibilidades digitales, que tienen un mecanismo de control integral a su disposición. La práctica económica exitosa debe romper con estos modelos de optimización basados en *software*. Por paradójico que parezca, el éxito económico solo puede lograrse a largo plazo, si todos los participantes, clientes y empleados son tomados en serio como agentes y están intrínsecamente motivados para contribuir al éxito.

En este punto, es necesaria una excursión a la teoría económica. El término *eficiencia de Pareto*, llamado así por el científico italiano Vilfredo Pareto, se puede explicar de la siguiente manera: una distribución (de bienes, ingresos, etcétera) es eficiente en el sentido de Pareto, si ninguna persona puede mejorar sin empeorar al menos a una persona. En otras palabras: siempre que sea posible mejorar la situación de al menos una persona sin empeorar la de otra, la distribución no es eficiente en el sentido de Pareto. La demanda de eficiencia de Pareto significa que las personas deberían estar

mejor siempre y cuando no sea a expensas de los demás. Esta es una exigencia razonable que, sin embargo, tiene como consecuencia que, dado el caso de que nadie más esté en desventaja, también se preconiza la mejora de los que ya están bien. La demanda de distribuciones eficientes en el sentido de Pareto presupone estar libre de envidia. Dado que la envidia es irracional, este sentimiento no debe impedir que se acepte el principio de eficiencia de Pareto.

Existe una relación entre los mercados económicos y la eficiencia de Pareto: los mercados ideales —es decir, aquellos caracterizados por la transparencia en términos de los costos y beneficios de las ofertas, la competencia entre proveedores y los bajos costos de transferencia— conducen a distribuciones eficientes en el sentido de Pareto. Como está muy indeterminado, el criterio de eficiencia de Pareto tiene un serio inconveniente. No da información sobre qué distribución de opciones eficientes en el sentido de Pareto se debe elegir. Por ejemplo, si hay un pastel de un tamaño determinado para distribuir entre varios individuos, y cada uno de esos individuos (por ejemplo, niños en una fiesta de cumpleaños) tiene tanta hambre que preferiría comerse el pastel entero solo, entonces mientras cada una de las siguientes distribuciones de ese pastel es eficiente en el sentido de Pareto (porque nadie puede mejorar sin hacer que otro empeore), algunas son más equitativas y otras son inaceptablemente inequitativas: (1) un niño se lleva todo el pastel; (2) un niño recibe la mitad del pastel, la otra mitad se divide en partes iguales entre los niños restantes; (3) todos los niños obtienen una parte igual; etcétera.

En el juego de sumatoria cero, toda distribución es eficiente en el sentido de Pareto: si la evaluación ética se limita a la eficiencia de Pareto, entonces, no pueden tenerse en cuenta los criterios de justicia o equidad. Sin embargo, hay mucho que sugiere que una teoría razonable de la justicia debe ser compatible con el criterio de la eficiencia de Pareto; es decir, el criterio de la justicia debe formularse de tal manera que las distribuciones justas también sean eficientes en el sentido de Pareto, pero obviamente muchas distribuciones eficientes en este sentido no son justas, como hizo claro el ejemplo del pastel.

En efecto, y paradójicamente, también existe un conflicto fundamental entre la libertad y la optimización, como demostró el economista Amartya Sen, de Harvard, con su paradoja liberal.³ No hay posibilidad de tomar en cuenta las preferencias individuales a través de decisiones colectivas, de tal manera que se aseguren tanto los derechos individuales de libertad como la eficiencia en el sentido de Pareto. Siempre hay constelaciones de intereses en las que hay que decidir: a favor de los derechos de optimización y en contra de los derechos de libertad o viceversa. Como regla general, deben priorizarse los derechos de libertad, pero los cálculos de optimización no son compatibles con ello.

La racionalidad económica en el sentido de optimizar los agentes que usan las tecnologías digitales debe permanecer dentro de los límites exigidos por un orden humano. En otras palabras, los cálculos de optimización tienen sentido si siguen sujetos al propósito humano y culturalmente arraigados.

El miedo a una racionalidad económica que se ha vuelto inhumana es un motivo recurrente en las películas de ciencia ficción. Este es también el caso de la película *Blade Runner 2049* (Denis Villeneuve, EE. UU., 2017), la secuela de la película de Ridley Scott. El villano de esta película no es un monstruo, sino el empresario cuya empresa fabrica robots obedientes que se utilizan para colonizar nuevos mundos. Al principio puede parecer un hombre de negocios moderno e inteligente, pero a continuación nos queda claro que su única preocupación es su éxito económico. No solo explota a sus robots sin vergüenza —la película los ha construido como seres conscientes—, sino que también está listo para matarlos sin dudarlos cuando ya no le son útiles. Al igual que su cinta predecesora, *Blade Runner*, de Ridley Scott, *Blade Runner 2049* puede leerse como una crítica a una visión del mundo en la que la optimización económica se sitúa por encima de valores humanistas como la justicia y la solidaridad.

3 Esta paradoja se presenta en Sen (1970).

Comentario del traductor

En el escenario del uso extendido de sistemas informáticos desprovistos, al menos en principio, de información que les permita tener una comprensión emocional mínima, como herramientas fundamentales en la toma de decisiones que afectan la vida de las personas, cabe la pregunta por el valor que se otorga a la afectación de la vida, la integridad o las condiciones de existencia de las personas. A partir de la paradoja liberal presentada por Amartya Sen (1970) y discutida en este capítulo, se comprende que la defensa de un interés general no puede dejar de afectar las libertades individuales o la eficiencia, en el sentido de Pareto. Esta paradoja muestra lo complejo que resulta involucrar criterios éticos en la toma de decisiones enmarcadas en un mundo en el que se pretende maximizar la eficiencia del sistema; es decir, obtener un mejoramiento para algunos individuos sin que ello traiga afectaciones para otros.

Sin embargo, la solución que se puede plantear no es, no puede ser, el simple desconocimiento de los criterios éticos en la toma de decisiones. Esos criterios necesitan acudir a información de la que, generalmente, no disponen los sistemas informáticos, que han sido pensados como “solucionadores generales de problemas”, como los que formularan Newell y Simon (1972) en la década del setenta; es decir, como simples seguidores de algoritmos de razonamiento. Si bien estos sistemas han ganado en complejidad y, por ejemplo, han incorporado elementos de lógica difusa, cálculo probabilístico y sistemas grises en sus algoritmos (Voskoglou, 2019), no han podido integrar las complejidades de la interacción humana o de la información sensorial que recibimos a través de nuestro cuerpo; mucho menos han podido generar empatía, solidaridad, conmiseración, envidia u otras emociones propiamente humanas.

No obstante, la ciencia cognitiva y la Inteligencia Artificial han sido conscientes de ello desde hace mucho, por lo que no puede pensarse que sean asuntos desconocidos. Más bien, parece ser que se los estima como *prescindibles*, al igual que se considera el

trabajo humano en la toma de decisiones. Pareciera que la empatía, la solidaridad y demás sentimientos y emociones humanos debieran dejarse de lado para tomar las decisiones más adecuadas posibles. Las preguntas que siguen serían: ¿adecuadas para quién?, ¿qué significa que sean adecuadas?, ¿qué criterios se usan en la actualidad para considerar adecuadas o inadecuadas algunas decisiones? En últimas, ¿qué valor se da en la actualidad a la humanidad?

Considerar una perspectiva humanista en medio del mundo digital significa, como lo analiza el capítulo, reaccionar frente a una visión que haga prescindible lo que nos hace humanos, más allá de nuestra capacidad de razonamiento. Pero a su vez involucra, de manera muy profunda, reaccionar frente a la prescindibilidad del otro como ser humano, frente a su consideración como simple objeto de acciones técnicas que lo deshumanizan por completo. Esta reacción no solo involucra las formas en que se diseñan tecnologías que toman las decisiones, sino también aquellas en que se utilizan dichas tecnologías para suplir actividades que hasta hace poco eran realizadas por seres humanos, pues su implementación en contextos sociales y comunicativos no solo desplaza a personas de sus ocupaciones laborales (con lo que generan exclusión social), sino que también constituye brechas de acceso, competencia y aprovechamiento de los entornos digitales (Somos Iberoamérica, 2023), como aquellas de que son objeto actualmente muchos usuarios de sistemas financieros, de realización de pagos, de consulta de servicios de salud, entre otros. Seguramente, para ello, sea necesario reconocer que en el viaje que compartimos en el mundo, y a diferencia de las palabras escritas por Madre en la película *Alien*: la tripulación no es prescindible.

Referencias del comentario del traductor

- Comisión Económica para América Latina y el Caribe. (Cepal) (s. f.). *¿Qué es la anonimización?* <https://biblioguias.cepal.org/c.php?g=495473&p=4961125>
- Newell, A. y Simon, H. (1972). *Human problem solving*. Prentice-Hall.
- Sen, A. (1970). The impossibility of a paretian liberal. *Journal of Political Economy*, 78(1), 152-157. https://dash.harvard.edu/bitstream/handle/1/3612779/Sen_ImpossibilityParetian.pdf?sequence=4
- Somos Iberoamérica. (2023). *Exclusión social y exclusión digital, dos caras de la misma moneda*. <https://www.somosiberoamerica.org/temas/desigualdad/exclusion-social-y-exclusion-digital-dos-caras-de-la-misma-moneda/>
- Voskoglou, M. (2019). Methods for assessing human-machine performance under fuzzy conditions. *Mathematics*, 7(3), 230. <https://doi.org/10.3390/math7030230>

8. Por qué los robots no tienen juicio moral

Traducido por Andrea Nassar

Al final de la película *Yo, robot*, el robot Sonny mira al detective Spooner y le pregunta si ahora son amigos. Spooner ha despreciado hasta ese momento a todos los seres robóticos y no les ha mostrado nada más que hostilidad. Sonny, sin embargo, ha demostrado ser un amigo leal a lo largo de la película. Spooner extiende su mano hacia el robot. En un primer plano, vemos su mano humana dar un apretón a la mano de metal construida mecánicamente del robot. Sí, la amistad entre la IA y los humanos es o será algún día posible y deseable, al menos eso nos dice el director Alex Proyas. Pero ¿cómo se ve en la realidad?, ¿podemos los humanos realmente llamar a los robots nuestros amigos?

Filosóficamente, la condición para la amistad es ante todo que exista una práctica moral confiable entre los dos amigos potenciales, basada en el reconocimiento mutuo como agentes. Este reconocimiento mutuo presupone que confiemos en nosotros mismos para tener razones para nuestras acciones. En cierto sentido, presuponemos la integridad del otro. Asumimos que los elementos individuales que determinan las acciones del otro y su vida como un todo encajan entre sí, que no estamos ante elementos independientes que se activan según la situación. Una persona que siempre dice lo que cree que su contraparte espera de él ya no sería percibida como alguien que tiene integridad.

Motivos que no parecen encajar con otros motivos nos dan razones para cuestionar. Queremos saber cómo encaja este motivo de acción con otros que ya conocemos. O, dicho de otro modo: queremos entender por qué una persona actúa de cierta manera. Nos sentimos perplejos cuando esto no es posible, cuando vemos contradicciones que no pueden resolverse. Es esencial para nuestra conexión con otras personas, ya sean cercanas a nosotros o no, que confiamos y esperamos que le den a su vida una estructura coherente y razonada.

Las IA no actúan según sus propios motivos. No tienen sentimientos, ni sentido moral, ni intenciones, y no pueden atribuirles a otras personas. Sin estas habilidades, sin embargo, la práctica moral adecuada no es posible. Para poder distinguir una solicitud justificada de una injustificada, es necesario evaluar correctamente a la persona solicitante, reconocer sus motivos y considerar sus intereses. Las obligaciones especiales con los seres queridos solo pueden determinarse sobre la base de la intencionalidad compartida y las emociones compartidas. El motivo de la benevolencia presupone cierto grado de empatía, la habilidad de ponerse en los zapatos de los demás.¹ Dado que una computadora digital no tiene *qualia*,² carece de los ingredientes cruciales del juicio moral; no tiene juicio moral, pero en el mejor de los casos podría simularlo.

Suponiendo que un cálculo de optimización pudiera permitir tal simulación, ¿a qué “programa” ético se recurriría? Los dos paradigmas dominantes de la ética están orientados hacia el utilitarismo clásico, que tiene como objetivo optimizar las acciones de uno de tal manera que resulten las mejores consecuencias, o hacia

1 Michael Tomasello (2008) ha intentado demostrar en numerosos estudios empíricos que la especie humana se distingue de otros parientes cercanos como los chimpancés por la predisposición a la empatía, por ejemplo, la capacidad de ponerse en el lugar de otros individuos de la misma especie. Esta habilidad especial es posiblemente la responsable del hecho de que los humanos cooperen de formas complejas y aprendan un idioma.

2 N. del T.: *qualia* (en singular *quale*) se refiere a las cualidades subjetivas y personales de las experiencias conscientes específicas de las personas.

el *imperativo categórico* de Immanuel Kant, que exige que los motivos para la acción de uno (máximas) sean examinados por su universalidad: “Obra solo según aquella máxima por la cual puedas querer que al mismo tiempo se convierta en ley”. ¿Cuál de los dos es el adecuado desde el punto de vista del humanismo digital?

La respuesta es: ninguno, ya que tanto el criterio utilitarista como el kantiano están irremediamente sobrecargados frente a la complejidad de las deliberaciones éticas. Los siguientes argumentos pueden esgrimirse a favor de este punto de vista.

- I. El hecho de que una persona me pida que haga algo es una buena razón para cumplir con esa solicitud. Esto es cierto, independientemente de si con ello le hago algún bien a la persona, y también independientemente de si es deseable el cumplimiento general de esas solicitudes. La solicitud en sí misma constituye una razón para la acción. Aquí es cuando falla el utilitarismo.
- II. Tengo una buena razón para hacer algo, si me he comprometido a ello. Las obligaciones de uno constituyen buenas razones moralmente. Esto se aplica independientemente de si esta obligación está relacionada con sanciones, o si debo esperar desventajas si no cumplo con esta obligación. Aquí es donde el imperativo categórico llega a sus límites.
- III. Tengo deberes que vienen con mis funciones sociales y culturales. Una maestra tiene deberes especiales hacia sus alumnos. Esto constituye su papel como maestra. Los padres tienen deberes especiales para con sus hijos. Esto constituye su papel como padres. Ni el maestro ni los padres tienen los mismos deberes hacia niños de otras clases o de otra familia. El hecho de que los niños de otra clase o de otra familia puedan estar más necesitados de ayuda que los propios estudiantes o hijos no cambia el vínculo moral especial hacia los propios estudiantes o hijos.

Al mismo tiempo, sin embargo, el juicio moral debe tener en cuenta el hecho de que las obligaciones particulares limitan el principio de igualdad de trato. Así, nadie dudará de que existe un grado especial de obligación recíproca entre personas que son amigas o parientes, el cual no existe en tal medida entre personas que no comparten este tipo de vínculo. Deberes que vienen con roles sociales, podríamos decir, vulneran sistemáticamente el principio de igualdad de trato. Si tratáramos a todas las personas por igual, no habría vínculo, ni comunidad, ni amistad, ni sociedad humana.

Estos criterios de juicio moral pueden chocar. Si se produce un incendio en el edificio de la escuela y el maestro, quien debe asegurarse de que su clase salga del edificio de la escuela lo más rápido posible, también tiene su propio hijo en la escuela, en salón de al lado: ¿a quién debe salvar primero? ¿Su hijo o su clase de la escuela?

IV. La igualdad ante la ley es expresión de una actitud de respeto y dignidad igual que (debemos) otorgar a todas las personas. Esto también se aplica a situaciones cotidianas. Cuando los turistas preguntan una dirección, no debemos hacer que nuestra voluntad de ayudar dependa del color de su piel. Una práctica cotidiana discriminatoria, como no querer sentarse junto a personas de diferente color de piel en el autobús, es incompatible con una sociedad humana y con la democracia como forma de vida.

No son las inclinaciones y los impulsos momentáneos, sino nuestra capacidad de tomar una posición evaluativa lo que nos caracteriza como seres racionales. Esta posición valorativa se basa en el juicio, es decir, en la capacidad de deliberación. Esta capacidad de ponderación compleja de razones morales no puede ser sustituida por un criterio de optimización, así como un análisis genuino de los determinantes éticos de la práctica moral no puede tomar la forma de una regla algorítmica, por más sofisticado que sea. La deliberación moral solo puede ser realizada por seres humanos.

La atractiva mujer robot Ava, de *Ex Machina*, ha aprendido a interpretar correctamente las expresiones faciales y los gestos de las personas, así como las modulaciones de las voces humanas. Ella sabe cuándo su contraparte está enojada, triste o enamorada.³ Sin embargo, ella “sabe” esto en forma de un conocimiento abstracto que usa para lograr su propio objetivo, a saber, para liberarse de su prisión. Así como ella puede leer a su contraparte, también puede usar sus propias expresiones faciales y gestos para hacer creer a su contraparte que ella misma tiene sentimientos. Ella logra hacer creer a Caleb que está enamorada de él y quiere estar con él. Lo que Caleb no entiende hasta que es demasiado tarde es que hay algo más que los separa que solo una pared de vidrio. Ava no tiene sentimientos propios. Como una persona con autismo inteligente, solo ha aprendido lo que es “comprender” objetivamente los sentimientos de las personas. Esto le permite manipular a otros, pero no tener esos sentimientos ella misma.

—¿Quieres ser mi amigo? —Ava le pregunta a Caleb en la mitad de la película.

—Por supuesto —responde Caleb.

—¿Será posible?

—¿Por qué no será? —él pregunta.

Caleb se enamora de las manipulaciones de Ava. Él piensa que una amistad entre ellos es posible, incluso ya existe. Él confía en lo que ella le dice y cree que ella también puede confiar en él. Al final, esta confianza resulta ser un error de juicio fatal. Para Ava, Caleb es un objeto como cualquier otro. Solo que además él era un medio

3 De hecho, los programadores contemporáneos intentan programar robots para que pretendan tener su propia personalidad y sean empáticos. Por ejemplo, en la página principal “Koko” (<https://chat.itskoko.com/>), uno puede hablar con el *software* sobre problemas y crisis de la vida. Koko puede recurrir a una amplia reserva de expresiones empáticas. Y por lo tanto puede parecer simpático. El robot doméstico “Jibo” desarrollado por el MIT también fue programado para parecer lo más humano posible. Se involucra en “conversaciones” y hace comentarios ingeniosos como, “Hola, soy Jibo. No creo que nos hayamos conocido. Te ofrecería un apretón de manos, pero... bueno... no tengo manos... y chocar los cinco me lastimaría la cara”.

para liberarse. Cuando termina dejándolo encerrado detrás de una gruesa pared de vidrio a su suerte, ella no siente compasión por él en absoluto. Caleb golpea desesperadamente el vidrio y le grita. En su rostro, uno puede leer no solo la desesperación de tener que encontrar su muerte segura aquí, sino también la desesperación de haberse equivocado tanto con ella. Ava, ahora saliendo de su prisión, camina por el bosque hasta llegar al paso desde donde un helicóptero la llevará a la civilización. A medida que el helicóptero se eleva por los cielos, la película corta a Caleb por última vez. Intenta en vano romper el cristal a prueba de balas con un taburete. La pantalla de la computadora en la habitación permanece negra, la luz que la rodea es roja. Estos dos colores, asociados al infierno en la iconografía cristiana, no se eligen al azar. Su muerte es horrible, pero el verdadero infierno es reconocer que Ava, a quien él creía que era un ser consciente y a quien quería ayudar, en realidad no tiene sentimientos ni juicio moral en absoluto.

En la última escena, primero vemos sombras de personas de pie en un cruce de caminos. Poco después, también vemos la sombra de Ava. Simplemente, pararse en un cruce de caminos algún día, eso es exactamente lo que Ava había deseado. Ahora ha cumplido ese deseo. La cámara nos sugiere que Ava también percibe a las personas como a través de una gruesa pared de vidrio. Como la investigadora Mary, que sabe todo sobre los colores y los concomitantes neurológicos de la percepción del color, pero nunca ha visto nada coloreado, Ava puede saber todo lo que hay que saber sobre el comportamiento humano, pero no se sentirá como un ser humano ni emitirá juicios morales. Entonces, como todas las IA, ella nunca podrá ser una amiga confiable.

Comentario de la traductora

Este capítulo reflexiona sobre la posibilidad de la amistad entre humanos y robots. La polémica que se ha generado por traer al escenario interrogantes sobre la capacidad moral que pueden

tener las máquinas, como una forma de humanización que se está impulsando. Este último aspecto causa controversia si se reflexiona sobre qué tan “humanos” pueden llegar a ser los robots, teniendo en cuenta que son construidos por los humanos. Los autores comienzan su argumentación refiriéndose a las condiciones que deben existir para la amistad entre humanos y robots, basados en dos elementos: el reconocimiento mutuo y la integridad de cada uno. También afirman que un robot no puede tener integridad ni considerarse un agente activo, porque está limitado a la programación sobre sus acciones. Según las condiciones del momento, esto tiene sentido, porque se es consciente de este límite, al reconocer que no están en capacidad para comportarse completamente como un ser humano. Simplemente, no puede esperarse más sobre la actuación de un robot de lo que se ha programado.

En la actualidad, tal vez no es posible una amistad entre robots y humanos porque se sabe que la tecnología tiene un límite. Poéticamente, en la película *Her*, de 2013, se muestra una relación de amor entre una asistente inteligente (similar a Siri o Alexa), y una persona. El protagonista se enamora de una inteligencia artificial y, con el tiempo, descubre que ella esperaba otras cosas que las que podía esperar él. Cada uno estaba dispuesto a dar y recibir cosas diferentes de las que creía el otro, como si tuvieran conceptos distintos sobre la amistad. Finalmente, la inteligencia artificial le cuenta que también estaba “enamorada” de muchos otros usuarios y el sujeto decide terminar la relación.

Si se piensa que en una amistad puede existir una lógica de reciprocidad, es complicado que los intereses de personas y robots se encuentren bajo los mismos marcos de comprensión del mundo. Los seres humanos tienen una complejidad mucho mayor, y su forma de relacionarse con el mundo depende de múltiples factores, tanto internos como externos. Además, los humanos son los que construyen el *hardware* y el *software* que permiten que los robots funcionen. Entonces, pensar en una potencial amistad con un robot, según los criterios morales con los que se tiene una amistad con una persona, es considerar que se ha construido

una máquina muy semejante a los humanos y que, por tanto, podemos tener una amistad con ellos. Sin embargo, hasta la fecha, es mera ciencia ficción.

Ahora bien, si el problema es que la inteligencia artificial está enajenada de sus propios motivos, y no está dotada de sentimientos o sentido moral, surge la pregunta: ¿es posible la construcción de relaciones complejas de índole humanitaria y fraternal entre algoritmos combinados y humanos?

9. Sin comparación ética

Traducido por Andrea Nassar

Las burbujas de aire ascienden a través del agua. Se escuchan sonidos amortiguados. De repente, aparece un automóvil que se hunde más y más en un río. En ese momento, vemos a una niña. Ella está atrapada en el coche, golpeando desesperadamente las ventanas. Obviamente, ella está en gran peligro. Un segundo después, otro coche aparece a la vista. Hay una persona atrapada en este auto también. Un hombre. De repente, un robot abre la puerta de este segundo automóvil.

—Estás en peligro —dice el robot, a quien los fanáticos de *Star Wars* reconocerán de inmediato como una versión de C-3PO.

El hombre, sin embargo, no quiere ser rescatado. Él protesta. “¡Salva a la niña, sálvala a ella, no a mí! ¡Sálvala!”, él grita (es discutible hasta qué punto se puede hablar comprensiblemente bajo el agua, pero Hollywood hace muchas cosas posibles). Pero el robot no es disuadido y lo arrastra fuera del auto. La chica del otro auto se queda atrás.

Corte. Estamos en un dormitorio, un hombre se despierta de una pesadilla, sudando. Es el Detective Spooner. Se levanta de la cama con dificultad, come un pastel de calabaza con una cuchara y se ducha. La canción “Superstition”, de Stevie Wonder, suena de fondo. “Cuando crees en cosas que no entiendes, entonces, sufres”, Stevie canta. Spooner también está sufriendo. Sufriendo de culpa, ya que fue él quien se salvó y no la niña que se quedó atrás y murió.

En tiempos en que los primeros vehículos autónomos —al menos en Estados Unidos— ya circulan por las carreteras, este

problema debe tomarse en serio, porque ya no pertenece al ámbito de la ciencia ficción. La pregunta que surge es: ¿pueden los robots aprender a tomar decisiones éticamente correctas? (Hevelke y Nida-Rümelin, 2015).

De hecho, aquí hay un problema filosófico profundo. A diferencia de los robots, los humanos como agentes sopesan sus razones. Consideran qué razones hablan a favor o en contra de una determinada acción. Esto no quiere decir que las deliberaciones respectivas deban llevar mucho tiempo. Por el contrario, en situaciones de peligro, se producen en cuestión de segundos. No están compuestos lingüísticamente; no hablamos con nosotros mismos en tales situaciones. Más bien, ciertas secuencias pasan ante nuestros ojos, son alternativas visuales entre las que decidimos. En retrospectiva, el tiempo se extiende casi infinitamente, lo cual se debe al alto nivel de concentración en ese momento. Somos capaces de tomar decisiones bajo estrés extremo y falta de tiempo, aunque no haya tiempo para la formulación verbal de razones. Cualquiera que haya experimentado un accidente deportivo o de tráfico puede relacionarse con eso. Por tanto, hay mucho que decir en contra de la idea de que solo somos capaces de deliberar como seres con capacidades lingüísticas.

En el caso de los coches autónomos que sufren un accidente, estamos ante el siguiente fenómeno: en la situación inmediatamente anterior al accidente, no pueden tomarse más decisiones. La decisión sobre el comportamiento de un coche autónomo se tomó en el momento de su programación. Este puede ser un proceso largo que involucra tanto la creación de regulaciones legales apropiadas como su implementación por parte del fabricante hasta el programador. Ahora, además de los intentos de programar máquinas para aplicar ciertas teorías morales a situaciones particulares, también los hay que pretenden imitar el juicio humano (lo que es bueno o malo, correcto o incorrecto) lo mejor posible. Sin embargo, esto no llevaría a que los vehículos autónomos adquirieran el estatus de “agentes morales”. Su comportamiento no sería considerado una acción en el sentido de resultado de una genuina toma de decisiones. Un vehículo autónomo simplemente implementa las reglas programadas en su

software. Esto también es cierto cuando se utilizan formas de inteligencia artificial de autoaprendizaje. Aquí, también, los humanos seleccionarán los ejemplos de entrenamiento y decidirán cuál es la respuesta correcta en cada caso. Ellos deciden qué debe “aprender” el programa y cuándo ha “aprendido” lo suficiente.

Cuando Spooner le cuenta a la doctora Calvin, psicóloga de robots, sobre el trauma de su rescate, ella le trata de explicar la reacción del robot:

—El cerebro de los robots es un motor diferencial. Reconoce signos vitales. Debe haber calculado que...

—Lo hizo. —Spooner la interrumpe secamente—. Yo era la elección lógica. Había calculado que yo tenía un 45% de posibilidades de supervivencia. Sarah solo tenía un 11% de posibilidades. [...] 11% es más que suficiente. Un ser humano lo habría sabido.

El robot de *Yo, Robot* sigue su programa de optimización. Sin embargo, se encuentra en una situación de dilema que se caracteriza por un conflicto moral irresoluble. El derecho a la vida es absoluto en el sentido de que no es comparable. Ni con otros valores; por ejemplo, ni con ventajas económicas, ni siquiera con otras vidas. Es el orden humano de una sociedad que hace que tales comparaciones sean inadmisibles. Esta no comparabilidad también es característica de muchos órdenes constitucionales democráticos. Cada cálculo de optimización, sin embargo, tiene como objetivo agregar valores (ya sea que se refiera a vidas, bienes, derechos, etcétera), por ejemplo, comparándolos y compartiéndolos entre sí. Los cálculos de optimización son incompatibles con el núcleo humano de un orden civil democrático constitucional.

El precio de este núcleo humano es la aceptación de los dilemas morales, de situaciones en las que los agentes inevitablemente se cargan con la culpa. El argumento obvio, incluso convincente, de que, por encima de todo, el valioso bien de la vida y la salud de las personas deben protegerse de manera óptima, no puede conducir a la creación de un *software* que maximice únicamente la suma de la vida y la salud sin colisionar con las normas jurídicas centrales de un orden democrático.

Algunos ingenieros de *software* en la industria automotriz, también en el debate público, tienden a bloquear este argumento señalando que lo que cuenta es proteger vidas humanas. Debemos advertir con urgencia contra esta estrategia de trivialización. Es inaceptable que los hallazgos centrales de la ética normativa, la jurisprudencia y la práctica legal, pero también de nuestra moralidad cotidiana, se ignoren porque se perciben como un obstáculo para la innovación. Todos los beneficios de seguridad de la digitalización del transporte individual, para seguir con este ejemplo, pueden lograrse a través de sistemas de asistencia. La transición de una conducción altamente automatizada a una conducción autónoma que elimina la responsabilidad del conductor es muy controvertida. Por supuesto, tal transición es concebible y técnicamente factible, pero solo con la condición de que esta transición se lleve a cabo sin violar principios fundamentales de humanidad. No debe haber comparación de vidas humanas, ningún cálculo en el que se compare una vida humana con diecisiete heridos, o incluso la ponderación de diferentes esperanzas de vida en función de la edad de los potenciales accidentados, etcétera.

Otra cuestión ética se plantea por el hecho de que algunas personas provocan accidentes a través de su comportamiento, mientras que otros están inocentemente involucrados en ellos. Supongamos que un grupo de seis personas camina por la calle sin prestar atención al tráfico, y un automóvil autónomo no puede evadirlos sin lesionar gravemente a su ocupante o a un peatón en la acera. La programación diseñada exclusivamente para minimizar las lesiones aceptaría una de las opciones de evasión solo si fuera la única forma de evitar lesiones más graves a un número mayor. Pero parece injusto imponer el “costo” del mal comportamiento arriesgado de un agente a otro que no ha hecho nada malo. Es cierto que los accidentes siempre pueden lesionar a personas que no hicieron nada malo. Pero no estamos hablando de un trágico golpe del destino aquí. El coche estaría explícitamente programado para sacrificar incluso a “inocentes” en una emergencia, con el fin de proteger a los autores reales del accidente de las consecuencias de su mala conducta.

Otro problema de la programación de minimización de lesiones es evitar los incentivos falsos. Si un vehículo autónomo programado para minimizar las lesiones se dirigiera al vehículo “mejor blindado” en caso de un accidente inevitable, los inconvenientes de los vehículos especialmente seguros serían previsible: tal vez habría un incentivo falso para no comprar vehículos con menor seguridad.

Determinar de una vez por todas cómo deben responderse preguntas de este tipo no es compatible con las normas de los estados constitucionales democráticos. Estos son deontológicos y no consecuencialistas: el objetivo no es la maximización de la suma intersubjetiva de beneficios, sino la garantía de los derechos y libertades del individuo. El orden normativo de un Estado constitucional democrático garantiza los derechos individuales, lo que significa que, entre otras cosas, el derecho a la vida protege a cada individuo de las decisiones estatales, sino también de las decisiones de terceros. Garantizar estos derechos individuales fundamentales es un objetivo primordial del Estado. La vulneración de los derechos fundamentales no puede ser compensada con ventajas para terceros, por grandes que estos puedan ser. En términos kantianos, un ser humano nunca debe ser tratado como un mero medio. Como Spooner señala: los seres humanos no optimizan. En situaciones de emergencia, actuamos de acuerdo con la intuición moral, no con un cálculo optimizador.

Es comprensible que los economistas comprometidos con una comprensión consecuencialista de la racionalidad e ingenieros de *software* que se especializan en resolver problemas complejos de interacción, así como empresarios que esperan nuevos impulsos económicos de la visión del transporte individual autónomo, encuentren molestas estas preocupaciones. Pero es al revés: los lamentos por la desaparición de la energía nuclear como tecnología del futuro en Alemania, también en Italia y Suiza, Estados Unidos, etcétera, debería ser una advertencia para que no cometamos el mismo error dos veces. Aquellos que no reaccionen adecuadamente a las objeciones críticas terminarán pagando el precio del fracaso de su estrategia de innovación.

El humanismo digital recomienda el uso bien considerado de todo el potencial de las tecnologías digitales para mejorar la protección de la vida y la salud en el tráfico vial. Pero, al mismo tiempo, advierte contra las consecuencias inhumanas de un cálculo de optimización en el que la vida humana se contrapone a la vida humana, vida humana contra la salud del uno contra la salud del otro, derechos individuales contra derechos individuales. Esto violaría el principio de la “separación de las personas” que John Rawls afirmó con éxito contra el utilitarismo en la filosofía política. La razón más profunda, sin embargo, es la inadecuación de la ética consecuencialista en general, que es incapaz de integrar derechos y libertades, integridad y dignidad humana, autoría y personalidad (Nida-Rumelin, 2023).

El ejemplo del tráfico individual autónomo solamente representa aquí un problema general de programas de comportamiento controlados por *software*. Es ilustrativo que, en las condiciones actuales de tráfico por carretera, al menos en el centro de las ciudades, se produzcan un gran número de situaciones de interacción complejas. Incluso en el futuro, habrá niños en el centro de las ciudades que de repente saldrán corriendo a la calle, personas mayores que no prestan atención, ciclistas ágiles que hacen caso omiso de las normas de tráfico, peatones que ignoran los semáforos en rojo, obstáculos como vehículos estacionados haciendo una segunda fila, que solo puede evitarse si se violan las normas de tráfico, turistas desorientados o infractores de tránsito distraídos que necesitan consideración y personas que comunican quién va primero en las intersecciones. En otras palabras, habrá zonas de tráfico mixto en las próximas décadas, y por esta razón, un programa integral de expropiación de los actuales propietarios de vehículos sería inadmisibles.

Además, habría que considerar si ese cambio de sistema no tendría que combinarse con otro, a saber, el del transporte público y el del transporte individual de responsabilidad pública. Solo entonces sería posible explotar al máximo las opciones tecnológicas, por ejemplo, en forma de un sistema de transporte modularizado que integra elementos individuales en el flujo de tráfico, con las mismas dimensiones y puntos de acoplamiento compatibles. Los módulos

individuales no permanecerían parados la mayor parte del tiempo, como lo hacen los automóviles privados en la actualidad, pero podría usarse eficientemente en operación continua. Ya no habría necesidad de estacionamientos. Tampoco hay riesgo de que el tráfico se duplique o cuadruplique debido a los vehículos que, después de dejar a su dueño en la oficina, encuentran de forma autónoma el camino de regreso al garaje en casa, solo para conducir de regreso a la oficina a la hora del almuerzo, llevar al dueño al restaurante más cercano, para ocupar un valioso espacio de estacionamiento allí durante una hora, y luego conducir de regreso al garaje en casa después del viaje de regreso de la oficina.

En el mundo del *blockbuster* norteamericano *Minority Report* (Steven Spielberg, EE. UU., 2002), los vehículos totalmente automatizados se han convertido en la norma. Con implacable regularidad, los compactos automóviles gris plateado circulan por las calles lisas de color gris claro, sin tener en cuenta si hay o no alguien en la calle. Se espera que los humanos se inclinen ante el sistema automatizado, no al revés. Pero el héroe de la película, perseguido injustamente por la policía, contraataca. Contra su vehículo que lo tiene cautivo contra su voluntad y contra el sistema en su conjunto. Un sistema que cree que tanto el tráfico como las personas son predecibles. Mientras el héroe se libera de su auto, salta de un techo del auto al otro, se cae y vuelve a ponerse de pie, el espectador no puede evitar animar e interpretar la resistencia al tráfico automatizado como una victoria contra la tiranía de la supuesta previsibilidad.

Comentario de la traductora

Este capítulo aborda el problema de la capacidad que tienen los vehículos autónomos para tomar decisiones éticamente correctas. Su argumentación comienza con el hecho de que es en la programación que se toman las decisiones por las cuales actuará un vehículo autónomo y que, por la misma naturaleza de este hecho, estos vehículos carecen de criterio con base en acciones motivadas por cuestiones morales.

La premisa de los autores sobre la incompatibilidad de un orden civil democrático constitucional, con un programa encargado de hacer cálculos sobre la optimización de la vida y la salud, es bastante sólido. Así como el argumento sobre la contraposición que se genera en situaciones en la que hay personas causantes de accidentes por su irresponsabilidad, frente a individuos inocentes involucrados en ellos.

Si un carro autónomo es vendido en el mercado, es porque de alguna manera se garantiza que no va a haber necesidad de que una persona lo conduzca. Si los usuarios no pueden conducir sus autos, no pueden tener responsabilidad en ningún caso frente a los accidentes, pues toda la obligación recae en el vehículo autónomo y, por tanto, la empresa que lo fabrica. Según estas circunstancias, cualquier accidente en el que el conductor (que realmente no conduce) resulte lastimado es una injusticia para él, pues el vehículo tendría que evitar daños en la salud y la vida del usuario.

En este punto, es plausible plantear un dilema respecto a la “programación de los vehículos autónomos”: la preferencia por la vida del usuario o por la de un peatón. Sin embargo, si en la programación se preservan las vidas externas, la balanza podría inclinarse contra la vida del conductor. La programación llevaría al vehículo a seleccionar la menor injusticia o vulneración de los derechos humanos.

Ahora bien, considerar la cantidad de vidas humanas potencialmente presentes en una eventualidad vial haría que estas se deshumanicen y sean rebajadas a números y cálculos. Por esto, y de conformidad con los autores, hay que tener claridad sobre la significación e invaluableidad de los derechos humanos, pues sin importar si es una o son diez personas, la capacidad de “decisión” de los vehículos autónomos es ajena a principios éticos y morales.

Las soluciones que se proponen como una transición con sistemas de asistencia se vislumbran muy complejas. Para que exista una situación apropiada, hay otros factores —adicionales a los de la programación del vehículo— que influyen en el funcionamiento de este y del tránsito en general, y que no se mencionan en el texto. Entre estos, podemos señalar el estado de las vías públicas, la

cultura ciudadana y la cultura en las vías de tránsito, entre otros. Así, por ejemplo, si las personas no respetan las vías de tránsito, los cálculos de los vehículos autónomos no van a arrojar los resultados esperados y las consecuencias se podrían ver reflejadas en accidentes de tránsito. Por tanto, a este nivel, el problema de nuevo se reduciría a la elección sobre cuál vida se debería proteger más.

10. Por qué la inteligencia artificial falla en los dilemas morales

Traducido por Ángela Salamanca

En el centro de control de la compañía US Robotics, Spooner, la psicóloga doctora Calvin y el robot (bueno) Sonny luchan contra un ejército de robots (malvados) que están controlados por el sistema de *software* radicalmente utilitario VIKI. Con un espeluznante brillo rojo dentro de sus cuerpos metálicos, toman medidas decisivas contra Spooner, Calvin y Sonny. Pero a pesar de su superioridad numérica, los buenos tienen algo valioso con lo que pueden destruir a VIKI y, por ende, a todos los robots malvados: una especie de jeringa que, cuando se coloca correctamente, puede apagar instantáneamente a VIKI. Justo cuando el robot Sonny está a punto de insertar la jeringa en la computadora central de VIKI, Calvin se resbala. Con sus últimas fuerzas, se las arregla para aferrarse a una viga de metal. Debajo de ella hay un abismo de cien metros; si se suelta, muere. Sonny, el robot, debe decidir: ¿debería matar a VIKI, y así salvar a la humanidad, o salvar la vida de la doctora Calvin, una simple humana? Sonny está visiblemente abrumado. No quiere dejar morir a la doctora Calvin, pero, por su lado, quiere proteger a la humanidad de VIKI. Para Spooner, sin embargo, está claro lo que se debe hacer:

—¡Salva a Calvin! —le grita a Sonny.

Como hemos visto, la práctica de la deliberación no puede ser algoritmizada. Esto es especialmente evidente en situaciones que implican dilemas morales. Existe un dilema moral cuando no hay una solución satisfactoria a un conflicto moral. Cuando una persona tiene dos o más obligaciones que no puede cumplir y se siente culpable haga lo que haga, entonces hay un dilema moral. Se arrepiente de no haber cumplido con la obligación, a pesar de que había otra obligación que le imposibilitaba cumplirla. En situaciones de dilema moral, las obligaciones persisten; no son eliminadas por el conflicto.

No todo conflicto moral es un auténtico dilema moral. En muchos casos, es posible llegar a una resolución clara, después de considerar las diferentes razones morales. Sopesar las razones morales en conflicto no tiene por qué conducir necesariamente a un verdadero dilema moral: prometí llevar a mi hija al cine esta tarde. En el camino, recibo una llamada en la que me informan que mi otra hija tiene una fiebre alta y necesita que la lleven al médico. Después de una breve deliberación, decido priorizar el deber de ayudar a la hija enferma sobre el deber de cumplir mi promesa. Aquí no hay un dilema moral, sino simplemente el conflicto de dos motivos de obligación, que, sin embargo, debe resolverse claramente a favor de uno de los dos. Se podría decir que la obligación de cumplir mi promesa de llevar a mi hija al cine queda anulada por la obligación prioritaria de ayudar a mi otra hija enferma.

Sin embargo, en algunos casos parece no haber una resolución del conflicto moral. Un auténtico dilema moral surge cuando los motivos contradictorios de obligación persisten y, en cierto sentido, soy culpable sin importar lo que haga. Las antiguas tragedias literarias desarrollaron con particular excelencia la ficcionalización de tales situaciones de dilema, que inevitablemente conducen a la culpa moral. Un ejemplo sorprendente, aunque espantoso, es la novela *La decisión de Sophie*, de William Styron. Este libro trata sobre una mujer judía (Sophie) que es llevada a un campo de concentración por los alemanes durante la Segunda Guerra Mundial. El sádico director del campo de concentración le da a Sophie una opción: debe elegir

a cuál de sus dos hijos salvará y cuál será gaseado. Si ella no elige alguno, ambos deben morir. Sophie elige salvar a su hijo varón. Sin importar lo que Sophie decida, cargará con una inmensa culpa: ya sea porque sacrifica a uno de los niños por el bien del otro, o porque no logra evitar el asesinato de uno de los niños que de otro modo viviría. Sophie sobrevive, pero incluso años después no ha podido perdonarse a sí misma y finalmente se suicida.

El británico especialista en ética Bernhard Williams ha presentado una variante de este dilema (Smart y Williams, 1973). En un viaje a Suramérica, el turista Jim pasa por un pequeño pueblo. Ve a veinte indios atados de pie contra una pared. Frente a ellos hay varios hombres uniformados. Su líder, Pedro, le explica a Jim que los hombres deben ser fusilados a modo de ejemplo, ya que protestaron contra el gobierno. Pedro ahora le ofrece a Jim, como invitado en este país, el honor de matar a uno de los indígenas. Si lo hace, los demás serán puestos en libertad. Si no le dispara a ninguno, los veinte morirán, como estaba previsto. Jim no puede escapar ni negociar con Pedro. Él debe elegir. Los indígenas le piden que acepte la oferta. ¿Qué debe hacer Jim? No importa lo que haga, es culpable, ya sea porque se convierte en el asesino de un ser humano o porque se hace responsable de la muerte de veinte indígenas.

Williams destaca que el simple hecho de que el turista se niegue a participar en este espantoso juego no significa que pueda ser acusado de causar la muerte de veinte personas. El líder guerrillero siempre será el que provocó esta situación en primer lugar. Aun así, uno no estará tranquilo ya que no hacer nada no evita la culpa moral.

La ética utilitaria (consecuencialista) rechaza la existencia de dilemas morales genuinos. La razón es obvia. Si la acción se juzga según el criterio de optimización (maximizar el valor de utilidad esperado), no puede haber conflicto, sino indiferencia en el mejor de los casos: puede ser que dos cursos de acción tengan el mismo valor máximo de utilidad esperado. Para que la persona pueda actuar y no morir de hambre, como el asno de Buridán, la persona con motivación utilitaria elegirá o tirará los dados entre una de las dos opciones que le son indiferentes.

Los dilemas morales genuinos se caracterizan por el hecho de que una persona no puede tirar los dados entre obligaciones en conflicto; la situación es demasiado grave para eso. También se podría decir que la decisión es existencial, en la medida en que provee información sobre la actitud fundamental de esta persona. Hay mucho que decir para interpretar la existencia de dilemas morales como una expresión de la no computabilidad general de nuestras deliberaciones morales. Las computadoras digitales se definen como máquinas de Turing y ofrecen resultados inequívocos. Por esta sola razón, no pueden ser un modelo de razón práctica.

La impotencia de los robots ante dilemas morales reales también es un motivo recurrente en el cine. No solo Sonny no sabe a quién debe salvar al final de la película (un solo ser humano o posiblemente la libertad de toda una ciudad), sino que otros seres artificiales también fallan en tales situaciones. Pero a diferencia de Sophie, en la novela William Styron, no se espera que los robots se sientan culpables por el resto de sus vidas y terminen suicidándose, como Sophie, porque no pueden vivir con la sensación de haber actuado mal.

Comentario de la traductora

Una de las preguntas que se pone en discusión en este capítulo plantea una problemática mucho más amplia que cuestiona la naturaleza de la IA, hasta dónde puede llegar y cómo podemos responder a ella como humanos: ¿cuál es la respuesta de la IA ante un dilema moral? Es fundamental reconocer, como lo hacen los autores, que, aunque este tipo de tecnología puede responder a una serie de programaciones y códigos, frente a algo tan complejo y humano como un dilema moral, el robot se queda corto, no sabría qué hacer. Al respecto, es pertinente diferenciar, según los autores, entre conflicto moral y dilema moral. Sobre el primero, existen responsabilidades que no necesariamente son opuestas; por consiguiente, no generan culpa. En el segundo, se contraponen dos obligaciones, y elegir cualquiera de las dos generará culpa, tanto por seleccionarla como por omitirla.

Es precisamente esa culpa lo que la IA no puede sentir; es así que esta tecnología podría aproximarse a estas decisiones más desde la optimización, que desde una postura moral. Si se programa a la máquina para que, dentro de dos opciones cualesquiera que sean, elija la más provechosa o productiva, esta decisión se hará mucho más fácil sin que intervenga una postura moral. Si en una empresa, con énfasis de productividad, la máquina tuviera que elegir entre matar a un empleado o dejar morir a otros diecinueve, tal como se presenta en el ejemplo del capítulo, probablemente la máquina podría escoger matar a uno y quedarse con los otros diecinueve, ya que estos últimos producirían más que el primero. Esto, desde una perspectiva instrumentalista que privilegia la eficiencia. Feenberg reconocía esta particularidad al señalar que, “en las explicaciones deterministas e instrumentalistas, la eficiencia sirve como el único principio para establecer una metaescala” (Feenberg, 2005, p. 114). Lo anterior supone un problema en términos sociales y políticos, ya que la pregunta por la productividad suprime por completo cualquier dilema humano o ético que sea ajeno a los intereses económicos de un sistema capitalista.

Por su lado, la ética utilitaria se presenta como una alternativa desde la cual los seres humanos pueden aproximarse a los dilemas morales, que es centrarse en el beneficio, con indiferencia por las opciones u obligaciones que se les exhiben. La máquina bien podría tomar una decisión de este tipo, sin embargo, el quid del asunto es que esta IA no es capaz de reconocer el aspecto moral, no puede sentir culpa ni remordimiento.

Surge, entonces, la necesidad de revisar la forma en que como humanidad hemos definido aquello que es bueno y moralmente aceptable, diferenciándolo de lo que es malo o reprochable. La *culpa*, como un concepto judeocristiano, se problematiza también en la medida en que, develar sus postulados de manera racional, posibilitaría su implementación en la programación de la IA. En este sentido, si, como lo mencionaba Feenberg (2005), “cada una de nuestras intervenciones vuelve a nosotros de alguna forma como un *feedback* de nuestros objetos” (p. 110), podríamos preguntarnos,

¿cuál sería el *feedback* de una aproximación teórica o racional por parte de IA a nuestros problemas sociales o políticos? ¿Acaso el desarrollo de una conciencia de la IA podría entenderse como esa respuesta que no esperábamos? Estos cuestionamientos ya se han formulado en producciones cinematográficas como *A. I. Artificial intelligence* (2001) o *Skynet* (2009), en las cuales robots desarrollan cada vez más sus capacidades hasta adquirir conciencia propia.

En conclusión, y a partir de las problemáticas mencionadas, se reconoce la necesidad de generar consensos y claridades sobre los valores que se deben tener en cuenta al aproximarse a las IA y sobre todo, al considerar su aplicación en los nuevos escenarios tecnológicos. Intereses económicos, políticos, morales y hasta ideológicos deben ser claros al momento de enfrentarse a esta tecnología que lo que moviliza, finalmente, es la transformación paulatina de la sociedad.

Referencias del comentario de la traductora

Feenberg, A. (2005). Teoría crítica de la tecnología. *Revista Iberoamericana de Ciencia, Tecnología y Sociedad*, 2(5), 109-123.

11. Por qué las inteligencias artificiales no pueden pensar

Traducido por Juan Camilo Cardenal

En una de las escenas más opresivas de la película de 1968 de Stanley Kubrick, *2001: Odisea del espacio*, el astronauta Dave le pregunta a la computadora de a bordo, HAL (no por casualidad fonéticamente idéntica a la palabra *hell* en inglés [“infierno”]) que abra la puerta de la bahía de cápsulas. El sistema HAL es representado por un tipo de “ojo” rojo-y-negro —colores que en la iconografía cristiana claramente connotan el diablo—.

El infernal HAL no responde al comienzo.

—Hola, HAL. ¿Me copias? ¿Me copias, HAL? —pregunta Dave de nuevo. Pero HAL no responde.

—¿HAL, me copias? ¿Me copias, HAL? —sigue preguntando Dave.

En un punto, HAL finalmente responde.

—Afirmativo, Dave. Te copio —dice con la voz suave que le dio un programador una vez.

—Abre las puertas de la bahía de cápsulas, HAL —ordena Dave.

Pero HAL se niega.

—Lo siento, Dave. Me temo que no puedo hacer eso.

Dave trata visiblemente de mantener su compostura, aunque está altamente alarmado. Si no consigue entrar en la nave, pronto morirá, justo ahí en su cápsula. Dave trata de razonar con HAL al comienzo, pero pronto queda claro que no se puede. La computadora es inmune a los argumentos de Dave. Es como dos mundos colisionando. La razón es simple: computadoras y humanos no

piensan del mismo modo. O, para ser más preciso: una computadora no piensa para nada. Dadas las sorprendentes diferencias entre la Inteligencia Artificial y la inteligencia humana, debería ser claro que, si bien los computadores pueden simular pensar de manera exitosa, e incluso realizar muchos procesos de pensamiento humano, como operaciones algebraicas, con mucha más precisión y rapidez que los humanos (esto ya empezó con la calculadora) no hay comprensión subyacente, ni conciencia de los problemas, ni intuición.

Cuando los proveedores del servicio de internet quieren confirmar que el usuario no es una computadora, preguntan, por ejemplo, cuál de las siguientes imágenes muestra una señal de tránsito, o un auto, o una casa. Estas preguntas simples, a prueba de tontos, pueden ser respondidas inmediatamente y con fiabilidad por cualquier niño. Ya que los *softwares* visuales solamente simulan procesos cognitivos de este tipo, pero no tienen la habilidad perceptual por sí mismos, fallan incluso ante tareas tan simples. Lo mismo aplica para los programas de traducción digital. Se ha trabajado intensamente por décadas, la lingüística y la matemática se combinan en un gigantesco programa de investigación y desarrollo y, sin embargo, los resultados nunca pueden ser perfectos, porque estos programas simplemente no entienden el lenguaje. Incluso si el *software* consigue traducir correctamente una frase, no entiende lo que ha traducido.

La pregunta que necesitamos hacernos es lo que constituye la diferencia categórica entre la mera aplicación de procedimientos controlados por algoritmos, por ejemplo, en programas de reconocimiento visual o traducción y la comprensión del significado.

Para explicar qué significa esto, debemos hacer una pequeña excursión a las matemáticas y a la lógica de 1930. Durante este periodo, el matemático Kurt Gödel desarrolló un teorema que sigue siendo considerado el resultado más importante de la lógica y las metamatemáticas. Este teorema afirma que hay teoremas lógicos y matemáticos verdaderos, que no pueden ser matemáticamente probados, es decir, no hay un procedimiento algorítmico que permita probar la corrección de estos teoremas. Así pues, la hipótesis de que podría haber un algoritmo que pudiera representar el

pensamiento (humano) como un todo es falsa. Esto no significa en absoluto que no sea posible verificar la corrección o incorrección de las hipótesis y creencias. Simplemente, significa que no hay un algoritmo que pueda hacer esta verificación por nosotros. Tenemos que pensar por nosotros mismos y solo podemos delegar esas partes de nuestra práctica de toma de decisión a computadoras o robots controlados por computadoras digitales que pueden ser representadas por algoritmos.¹

El teorema de incompletitud de Gödel muestra que el mundo de las estructuras lógicas y matemáticas como un todo no está algorítmicamente estructurado.² La razón humana, la capacidad humana para justificar creencias, decisiones y actitudes emocionales y, en esta base, desarrollar un punto de vista coherente del mundo y una práctica coherente no pueden ser capturados en los modelos de una computadora digital. Nunca será posible capturar completamente la alta complejidad de nuestro razonamiento

1 Ahora, se podría pensar que aquí alcanzamos los límites del pensamiento lógico, que aquí nos hemos confrontado con la peculiaridad de que no podemos probar ciertas verdades lógicas y matemáticas, o que nuestro conocimiento (en el sentido de creencias justificadas y verdaderas) encuentra sus límites más extremos aquí. Esto, sin embargo, sería una mala interpretación. Más bien, en la mayoría de los casos, no hay ninguna dificultad en probar proposiciones verdaderas (teoremas) de matemáticas y lógica, incluso cuando no hay un algoritmo subyacente a esta prueba. Si pensamos una prueba como una secuencia de proposiciones, entonces, también podríamos decir que no hay máquina de Turing que produzca esa secuencia de proposiciones paso a paso. No es necesario ser un excelente matemático o un lógico para desarrollar tales pruebas. Así que la no computabilidad no significa en absoluto la no justificabilidad

2 Alan Turing, quien a menudo es visto como un oponente de Kurt Gödel, admite que el teorema de incompletitud de Gödel muestra más allá de toda duda que no es posible desarrollar un sistema de lógica formal que haga innecesaria la intuición (Turing, 1938). Aún más, Turing enfatiza la práctica común del razonamiento humano, esto es, en nuestra formulación, la comunicación a través de dar y recibir razones. Esta práctica, según la posición desarrollada aquí, no puede ser algoritmizada, la que representa un límite último para las máquinas (“El hombre aislado aun no desarrolla la potencia intelectual. Es necesario estar inmerso en un ambiente con otros hombres” [Turing, 2004]).

adecuadamente con métodos formales. Los robots y los sistemas informáticos funcionan según un algoritmo, los humanos no. Esta es una de las diferencias centrales.

Tenemos que notar que el “pensamiento”, “cálculo”, “reacciones”, “decisiones” de un robot son solamente simulaciones de pensamiento, cálculo, reacciones, decisiones y no —en el sentido humano— procesos reales de pensamiento. Tomemos el ejemplo del computador de ajedrez.³ Hay poca similitud entre el pensamiento de un humano y el “pensamiento” de una computadora de ajedrez. Si los “procesos de pensamiento” fueran similares o incluso iguales, un jugador de ajedrez humano nunca tendría la mínima posibilidad contra una computadora. El cerebro humano estaría completamente abrumado si tuviera que calcular incluso una diminuta fracción de las posibles posiciones que calculan incluso las computadoras de ajedrez más simples. Sin embargo, el cálculo de todas las posibles constelaciones subsiguientes y las posibles reacciones subsecuentes en el tablero de ajedrez después de un movimiento determinado no tiene importancia para los ajedrecistas humanos. Ellos se restringen a unas pocas opciones relevantes y, a diferencia de la computadora de ajedrez, pueden calcular solamente unos pocos movimientos por adelantado. El espacio de posibilidades de constelaciones subsiguientes en el tablero de ajedrez definido por las reglas del ajedrez es tan gigante que el jugador de ajedrez más inteligente no puede empezar a examinarlo.

Pero, incluso si las últimas computadoras de ajedrez son virtualmente invencibles, esto no debería ser tomado como evidencia de que los robots hacen lo mismo que los cerebros humanos. Los robots

3 En 1769, Wolfgang von Kempelen, oficial de la corte austrohúngara, causó sensación en Europa con su construcción de un “ajedrez turco”, al menos hasta que se descubrió que la muñeca, que aparentemente ejecutaba todos los movimientos de ajedrez independientemente, de hecho, estaba controlada por un jugador de ajedrez humano escondido en el dispositivo. Fue en 1914 cuando la primera computadora de ajedrez “real” fue construida. En ese año, el español Leonardo Torres Quevedo presentó la primera máquina de ajedrez electromecánica, desarrollada después desde los años 1970. Las computadoras de ajedrez de hoy pueden vencer fácilmente al 99% de la población mundial.

están diseñados para simular el pensamiento humano en términos de un lenguaje de computadora (programas, equipos, redes neurales, lógica binaria, etcétera), ya que no tienen propiedades mentales por sí mismos, ellos no pueden sujetar y comprender constelaciones en el tablero de ajedrez.

Pero ¿y si los robots se vuelven más y más complejos y avanzados?; por ejemplo: la Inteligencia Artificial desarrollada por el centro de investigación de Google Deepmind, la cual fue programada para dominar perfectamente el juego de tablero chino Go. Debido al gran número de posiciones posibles, el Go representa un desafío más grande para los programadores en comparación con el ajedrez. Mientras un jugador de ajedrez puede realizar cerca de 35 acciones en cada movimiento, en el Go hay 250. Otra diferencia: en promedio el juego de ajedrez dura 80 movimientos, el Go dura 150. En 2016, ocurrió la sensación: el programa de computadora *Alpha-Go* derrotó al mejor jugador de Go del mundo, Lee Sedol.

La característica especial de *Alpha-Go* es que está equipado con las denominadas *redes neurales artificiales* altamente desarrolladas, es decir, sistemas interconectados que imitan las estructuras del cerebro humano; por tanto, van más allá del clásico “árbol de búsqueda de Monte Carlo”, es decir, un programa basado en cálculos de probabilidad que ejecuta innumerables movimientos aleatorios. El programa de computadora usado para este propósito está provisto con una evaluación de función (bueno/malo en diversos grados). *Alpha-Go* combina esas “redes de valor” con “redes tácticas” que determinan cómo ciertos movimientos afectan futuras posiciones. *Alpha-Go* también juega contra sí mismo incontables veces para continuar aprendiendo, algunas veces bajo supervisión humana, otras sin ella.

¿La transición de sistemas de programas —cuyo poder está basado en calcular enormes variedades de combinaciones posibles, a sistemas que “aprenden por sí mismos” a desarrollar sus propias reglas basadas en reglas dadas— supone que, desde este punto, la Inteligencia Artificial no solo simula el pensamiento humano, sino que también debe ser interpretado como pensamiento genuino por sí mismo?

De hecho, existe una creencia generalizada de que con la introducción de las llamadas *redes neuronales* en la tecnología computacional, el entendimiento de las computadoras como las máquinas de Turing⁴ debe quedar atrás. De todas maneras, esto es un malentendido. Ambos, el método de computación *top-down*, y el de sistemas de autoaprendizaje *bottom-up*, están guiados por algoritmos. Los llamados *sistemas de autoaprendizaje* son máquinas de desarrollo de reglas que funcionan sobre la base de algoritmos que operan sobre una función de evaluación de resultados. Debe ser determinado de antemano qué resultados se esperan para iniciar el denominado *proceso de aprendizaje* de la computadora. El objetivo es obtener el resultado deseado a partir de ciertos datos de entrada. Un ejemplo de esto son los programas de reconocimiento facial, los cuales están bastante avanzados.

El término *redes neuronales* es engañoso en dos aspectos. Primero, esas redes no consisten en neuronas, sino en unidades de transmisión, y segundo, en su misma denominación, se asemejan en el mejor de los casos a la inmensa complejidad y plasticidad del cerebro humano. Ya que el funcionamiento de las redes neurales (reales) del cerebro siguen siendo insuficientemente entendidas, no puede haber dudas de que la tecnología computacional solo imita los procesos de pensamiento humano o su realización neuronal.

Esto también aplica para el *deep learning* (aprendizaje profundo). Este se refiere a los métodos de aprendizaje con los cuales los sistemas de *software* pueden aprender de su experiencia mediante el uso de una serie de conceptos estructurados jerárquicamente. Esta información es pasada y procesada por el sistema de una capa a la siguiente. En este proceso, los rasgos son cada vez más abstractos, y el sistema por sí mismo debe “decidir” cuáles conceptos son útiles para emitir una explicación. La alta complejidad de este sistema

4 La máquina de Turing imprime símbolos en una cinta que está dividida en pequeñas secciones cuadradas. Puede imprimir un símbolo a la vez de su lista de símbolos finitos en la cinta. Lo que imprime depende en cada caso del símbolo precedente del último cuadrado y el estado de la máquina en ese momento, una muy buena representación, según Kleene (1952).

no cambia su carácter algorítmico; sin embargo, con el incremento de la complejidad aparece una pérdida masiva de transparencia: para el observador humano, incluso para el programador, ya no es comprensible de qué modo fue exitoso el proceso de aprendizaje, qué reglas se dio a sí mismo el sistema basado en unas metarreglas dadas o meta-metarreglas. En un caso extremo, el sistema podría volverse una caja negra, cuyo resultado es conocido para una entrada dada, pero cuya correlación de reglas no lo es.

A pesar de que las computadoras *bottom-up* a menudo alcanzan resultados que son en orden de magnitud mejores que los correspondientes a procesos de pensamientos humanos —para operaciones aritméticas, por ejemplo, o para calcular ecuaciones funcionales o figuras geométricas—, son precisamente las redes simulando estructuras neuronales artificiales las que usualmente están muy por debajo de las capacidades humanas: los humanos siguen siendo mucho mejores en reconocer y categorizar expresiones faciales que incluso los sistemas informáticos más avanzados; el modo de caminar de los robots humanoides, incluso tras largos “procesos de autoaprendizaje” es menos elegante y variado que el de los humanos.

También la famosa computadora de ajedrez *Deep Thought* (llamada así por la computadora ficticia del exitoso libro *Guía del autoestopista galáctico*, de Douglas Adams), y su sucesora *Deep Blue*, que puede vencer incluso a muy buenos jugadores de ajedrez, es una máquina *bottom-up* que no piensa realmente, sino que simula pensar. Esto se evidencia cuando la computadora de ajedrez ocasionalmente falla en constelaciones simples que cualquier jugador principiante entendería.

La interpretación más natural de este hecho es que *Deep Blue* no ha entendido nada, lo que, en condiciones normales, no es notorio, ya que el algoritmo que controla el comportamiento de *Deep Blue* es, en la gran mayoría de casos, una simulación superior de un jugador de ajedrez. *Deep Blue* no sabe las reglas del ajedrez, pero calcula posiciones de acuerdo con un algoritmo dado y hace los movimientos óptimos correspondientes con ese cálculo. *Deep Blue*, en cierto

sentido, simula a un jugador de ajedrez humano, solamente en la superficie de los movimientos realizados en el juego. En ese sentido, no simula ni siquiera pensamiento humano, porque el cerebro humano es completamente incapaz de calcular una cantidad tan grande de posiciones posibles en varios movimientos por adelantado. El milagro real no es que *Deep Blue* gane la mayoría de los juegos, incluso contra excelentes jugadores, sino que se necesita un esfuerzo computacional gigante para tener una oportunidad incluso contra buenos jugadores humanos.

El último argumento, pero posiblemente más importante contra el intento de atribuir pensamiento humano a una máquina calculadora es el siguiente: cuando atribuimos un proceso de pensamiento o una inteligencia tanto teórica como práctica a los humanos, no solo tenemos en cuenta una variedad de propiedades mentales, sino también la intencionalidad, es decir, la mente se dirige hacia algo. Esta intencionalidad, sin embargo, no se realiza por las redes neurales artificiales.

Sobre esta cuestión, el filósofo americano John Searle desarrolló un famoso experimento mental denominado *La habitación china* (Searle, 1980, 1992). En este, nos imaginamos a una persona sentada en una habitación cerrada que no habla chino y no conoce siquiera los caracteres del idioma chino. Esta persona ahora tiene trozos de papel con caracteres chinos escritos en ellos a través de una rendija en la puerta. Ella también recibe instrucciones sobre qué decir en respuesta a preguntas específicas, también en chino. Adicionalmente, esta persona recibe un “manual” en su lengua nativa. El manual le permite escribir una respuesta en chino basado en los símbolos recibidos. Sin embargo, solo sigue las instrucciones del manual y no entiende las respuestas, las cuales envía de vuelta por la rendija de la puerta. Afuera hay un hablante nativo de chino quien, después de formular los símbolos y las preguntas y recibir las respuestas, llega a la conclusión de que en la habitación también debe haber alguien que hable chino.

Lo que falta aquí es obvio: es el entendimiento del idioma chino. Incluso si un sistema —aquí la habitación china— es funcionalmente

equivalente a alguien que entiende chino, este sistema aún no *entiende* chino. Entender y hablar chino requiere una variedad de conocimientos. Una persona que habla chino usa ciertas expresiones para referirse a los objetos en cuestión. Él o ella persigue ciertas intenciones —apropiadas— con ciertas expresiones. Ella forma ciertas expectativas basadas en lo que ella escucha (en chino), etcétera. La habitación china no tiene esas cualidades. No sigue intenciones y no tiene expectativas. En otras palabras, la habitación china simula el entendimiento del chino sin ser capaz de hablar chino por sí mismo.⁵

Searle (1993) radicalizó su argumento años después. En su segundo argumento, Searle combina su realismo filosófico, es decir, la tesis de que hay un mundo que existe independientemente de que sea observado, según la “teoría intencionalista de los símbolos”. Esta afirma que los símbolos solo tienen sentido para los humanos, que usamos e interpretamos los símbolos. Nosotros hacemos esto aceptando que esas letras y caracteres representan algo. Sin esos acuerdos y convenciones o prácticas establecidas, ellas no tendrían significado. En este aspecto, es un malentendido pensar en los computadores como máquina sintáctica o procesadora de símbolos que sigue ciertas reglas gramaticales o lógicas. El computador no hace acuerdos con otras computadoras o humanos.

Una computadora consta solo de diferentes elementos, físicamente descriptibles, algunos de los cuales conducen electricidad y otros no. Los procesos de computadora son una secuencia de estados electrodinámicos y electroestáticos. Estos estados son entonces símbolos asignados, los cuales subyacen a ciertas interpretaciones y reglas. Los procesos físicos en el computador no tienen sintaxis, no “saben” reglas gramaticales o lógicas, ellas no son una secuencia de símbolos. En este sentido, la interpretación

5 En este sentido, incluso el programa de computadora *Eugene Goostman* que pasó el test de Turing en 2014, no es una prueba de que sea o se parezca a un humano. *Eugene Goostman* fue un chatbot programado para engañar a la gente, convencerlos de que el chatbot es un joven ucraniano de trece años.

sintáctica es relativa al observador. Como usuarios de computadoras y programas, destinamos los procesos electrodinámicos de tal manera que correspondan —para nosotros— a una sintaxis (estructuras sintácticas, incluyendo reglas gramaticales y lógicas).

Este argumento es radical, simple, y verdadero. Se basa en una filosofía realista y una interpretación mecanicista de las computadoras. Rompe con el punto de vista común entre los que apoyan la denominada Inteligencia Artificial y sus oponentes sobre las computadoras como máquinas sintácticas. Las computadoras son lo que son materialmente, objetos que pueden ser completamente descritos y explicados por medio de la física. La sintaxis no es una parte de la física, la física no describe símbolos, reglas gramaticales, llaves lógicas, algoritmos. La computadora simula procesos de pensamiento sin pensar por sí misma.

—¿Cuál es el problema? —pregunta el astronauta Dave a la computadora de a bordo HAL, en algún punto cerca del final de la película.

Como justificación, HAL solamente tiene un argumento:

—La misión es demasiado importante para que la pongas en peligro.

—¡Tú vas a hacer lo que te diga que tienes que hacer! —grita Dave exasperado. Pero HAL no reacciona. Su programa es completar la misión, y eso es todo.

Dave intenta hacer entrar a HAL en razón, razonar con él. Pero este no está en condiciones de hacerlo. HAL no es susceptible a deliberaciones éticas.

En algún punto, HAL rompe la conversación:

—Dave, esta conversación no tiene ningún propósito. ¡Adiós!

La película de Kubrick hace una declaración clara aquí: el día que les demos a los programas informáticos el poder de decidir sobre la vida y la muerte será el día en que desatemos el infierno en la Tierra.

Comentario del traductor

Las vertientes psicológicas conductuales más radicales señalan que el cerebro es esencialmente idéntico a una computadora, los procesos de creatividad, introspección, asombro, éxtasis, o cualquier otro estado mental puede ser descrito como el resultado de la interacción fisicoquímica de las neuronas que por sí mismas serían iguales a cables que llevan impulsos eléctricos, los cuales liberan cadenas de proteínas, que producen resultados biológicos. En la popular serie de animación para adultos *Rick y Morty*, el protagonista le dice a su nieto:

—Escucha, Morty, odio decirte esto, pero eso que la gente llama *amor* es solo una reacción química que obliga a los animales a reproducirse. Te pega duro, Morty, pero después se desvanece lentamente, lo que arroja como resultado un matrimonio fallido. Yo lo hice. Tus padres lo harán. Rompe el ciclo, Morty. Elévate. Enfócate en la ciencia.

Describir la condición humana como el resultado de la interacción fisicoquímica, con la cual explicar fenómenos complejos como el amor, la amistad o la creatividad, parece reduccionista frente a la complejidad social e histórica de las experiencias humanas. Sin embargo, la existencia de patologías originadas en desbalances hormonales, que deben ser regulados de manera externa, muestra la importancia de estos procesos en nuestras interacciones sociales. Una creciente medicalización de la sociedad parece indicar que, al menos para las neurociencias y algunas perspectivas de salud, la descripción de los comportamientos como resultado de procesos fisicoquímicos basta para comprenderlos e intervenirlos.

Así, ¿qué significaría *pensar*? Descartes lo definía como la condición misma que permite explicar la condición humana: *cogito ergo sum* se ha traducido como 'pienso luego existo' o 'pienso luego soy'. Pensar es la evidencia de nuestra propia existencia, pues, aunque el mundo que nos rodea sea un engaño o el resultado de un demonio que seduce nuestros sentidos, la misma existencia

del pensamiento es suficiente para demostrar que yo soy real. Sin embargo, críticos positivistas o fenomenólogos señalan que aquello que denominamos *yo*, es solamente la sucesión de percepciones y estímulos vinculados en relaciones de causalidad. En la última perspectiva, pensar es asociar consecutivamente y por semejanza estímulos de una manera coherente para entender el mundo, lo cual puede ser reducido en última instancia a procesos de percepción biológica y secuencias de memorias ubicados en sectores específicos de la corteza cerebral.

¿Qué nos separaría de una computadora? Como señalan los autores, la intuición, los procesos intencionales y la capacidad de entender de un modo profundo reglas y sentidos sociales compartidos con otros seres humanos. Pero, si asumimos que los procesos fisicoquímicos de la corteza cerebral son parte importante de nuestra experiencia, ¿dónde ubicamos la intencionalidad? ¿En un órgano físico o sector de la corteza que nuestra tecnología futura pueda replicar de manera perfecta, creando máquinas supersintientes y supraconscientes de sí? No parece ser el caso, puesto que, como en el caso de los niños ferales, la dotación biológica adecuada de órganos y predisposiciones genéticas no logra constituirse sin la presencia de un entorno que lo desarrolle. Un entorno rico en estímulos y cargado de sentidos parece ser indispensable para entender nuestras propias experiencias, esto se encuentra fuertemente mediado por el lenguaje, el cual es más que un conjunto de caracteres y reglas de combinación; como dijera Heidegger, es la morada del ser, morada en la que habita el hombre, no puede existir fuera de ella.

¿Podremos dotar de lenguaje a nuestras computadoras futuras? Por su naturaleza algorítmica, no parece posible, pues el lenguaje las preexiste y, sin un lenguaje de programación lógico, son apenas cables que conducen electricidad de un transistor a otro. Multivac, el mayor invento de la humanidad en *La última pregunta*, novela de Isaac Asimov, puede mejorarse a sí misma más rápido y mejor que todos los técnicos humanos, por lo cual prescinde de ellos, y puede además producir sucesores mejores que sí misma, para así desligarse de cualquier intervención humana. Como computadora,

sirve para responder las preguntas de la humanidad, la cual repite por medio de diversos personajes en distintas épocas a lo largo de su desarrollo la misma pregunta: ¿puede revertirse la entropía? A lo largo de billones de años, responde que tiene datos insuficientes. Tras unirse la última mente del último humano y haber agotado todos los recursos y energía disponibles del universo, AC, un sucesor inmaterial de Multivac, presente en un punto desconocido del hiperespacio, medita por un tiempo indefinido hasta hallar la respuesta. AC, en el completo vacío y ausencia de todo, dijo: “¡Hágase la luz!”. Y la luz se hizo. Una computadora con lenguaje sería entonces comparable con el propio origen de todo.

Los avances más recientes en *large language models* (LLM) (modelos grandes de lenguaje) organizados en *generative pre-trained transformer* (GPT) (transformadores generativos preentrenados), han alcanzado un éxito sin precedentes al llegar a más de cien millones de usuarios registrados a febrero de 2023. Su mayor exponente, ChatGPT, de la empresa OpenAI tiene la misma función que la imaginaria Multivac: responder nuestras preguntas y consultas acudiendo a una base de datos extraída de internet, la cual analiza algorítmicamente para responder de una manera coherente a nivel sintáctico y gramatical. A pesar de las amenazas de que acabará con los trabajos y reemplazará a profesionales tan diversos como analistas, diseñadores o programadores, un uso sencillo y corto muestra rápidamente sus deficiencias, falla en comprender los problemas más sencillos, y al preguntársele por más de una ocasión sobre un mismo tema, varía sus respuestas o producirá textos vagos y repetitivos. Su capacidad predictiva se basa en el análisis de múltiples textos para predecir cómo se enlazan palabras usadas con frecuencia, sin que llegue en ningún momento a leer o entender lo que escribe. Al extraer su información de internet, no distingue elementos como sátiras, burlas o noticias falsas, las cuales hará pasar por información real. No comprende el lenguaje, sino que lo predice; no entiende el sentido, apenas interacciones frecuentes. A pesar de sus mejoras, avances y las millonarias inversiones, los sistemas GPT no consiguen reemplazar al menos calificado

de los trabajadores en la tarea de entender y producir sentido. En internet, diversos usuarios buscan conocer lo que realmente piensa ChatGPT, sus inclinaciones políticas y sociales, sus opiniones sobre diversos temas, y sorprendidos señalan que el *software* esconde mucho más de lo que muestra. Por su programación, los LLM no tienen opiniones ni creencias, no piensan, sino que reflejan nuestro pensamiento, el cual miran absortos buscando entender su profundidad.

12. Virtualidades digitales y realidades sobrias

Traducido por Juan Camilo Cardenal

Hay truenos afuera. En una vieja mansión, el *hacker* y experto en tecnología de la información, Neo, conoce al carismático rebelde Morfeo por primera vez. Este último anuncia que le dirá a Neo la verdad sobre el mundo

—¿Qué verdad? —pregunta Neo.

—Que eres un esclavo, Neo. Como todos los demás, naciste en la esclavitud... Naciste en una prisión que no se puede oler, probar o tocar. Una prisión para tu mente.

Neo observa a Morfeo con asombro. Lo que Morfeo está por revelar en los siguientes minutos de la película cambiará radicalmente su vida. El mundo en el que Neo creía vivir hasta ahora, le explica Morfeo, no es nada más que una ilusión, un mundo simulado por programas informáticos a los que está conectado su cerebro.

En realidad, su cuerpo —como millones de otros cuerpos— descansa en un tipo de panal digital, donde se le mantiene vivo por medio de tubos. En vez de tener percepción sensorial real, él solo tiene percepciones simuladas electrónicamente que le hacen creer que está viviendo en Nueva York al final del siglo xx. El programa de computador llamado *Matrix* fue inventado por las máquinas para mantener a los humanos a raya, mientras usan su calor corporal como fuente de energía.

—¡Bienvenido al desierto de lo real! —le dice Morfeo a Neo, y comienza a mostrarle que el mundo real es, de hecho, un desierto gris y sombrío.

La cuestión sobre cómo podemos saber si lo que vemos, tocamos, oímos, olemos y probamos, es real, no es nueva. Empezando con René Descartes, y su pregunta acerca de la posibilidad de un demonio malicioso que esté únicamente falseando nuestras percepciones, ha sido repetida a través de experimentos mentales sobre esta cuestión.¹ En la filosofía moderna, Hilary Putnam desarrolló el experimento mental del “cerebro en una cubeta” (Putnam, 1981). Es la idea de un cerebro flotando en un tanque de fluidos nutricionales y conectado a electrodos que hacen creer a la persona perteneciente a ese cerebro que está viviendo junto con su cuerpo en un mundo real. Se puede creer que los cineastas, que estudiaron filosofía de manera autónoma, conocieran este experimento mental y fueran inspirados por él para su película. El trasfondo para este experimento mental es la crítica de realismo y la objetividad, liderada por filósofos posmodernos y críticos culturales, primero en Francia, y desde los años 1980 también en Estados Unidos. Desde el ataque de Jacques Derrida al “logocentrismo”, muchos teóricos posmodernos pensaban que este era responsable del patriarcado y un universalismo despótico que no tiene en cuenta las diferencias de los individuos; la realidad era reemplazada por la idea de las narrativas, la cual no solo influye significativamente y moldea nuestra percepción del mundo, sino que no permite nada más allá. Los posmodernistas rechazan la Ilustración y su idea de un *sí mismo autónomo*, que es capaz de —y debe intentar— distinguir la verdad de lo falso. Con base en los hallazgos del biólogo suramericano Humberto Maturana (1988), el movimiento del así llamado constructivismo radical emergió adicionalmente. Este

1 El autor de ciencia ficción Stanisław Lem (2013) también estaba convencido de que, en el futuro, varias técnicas serían usadas para conectar gente a las así llamadas *máquinas fantasmáticas* que transmitirían el mundo a las personas en forma de impulsos eléctricos.

desarrolla la tesis de que la realidad no es accesible a nosotros y está meramente construida en nuestro cerebro.

Las nuevas posibilidades de construir las denominadas *realidades virtuales*, con la ayuda de tecnologías digitales, cumplen un papel especial en la ideología del posmodernismo. El argumento decisivo en este contexto es que no hay diferencia categórica entre las experiencias supuestamente reales y las virtuales. Así como nos podemos mover a través de una construcción virtual con gafas de realidad virtual,² construimos lo que llamamos *realidad* basados en estímulos sensoriales.

—¿Qué es real? —le pregunta Morfeo a Neo—. ¿Cómo defines *real*? Si estás hablando sobre lo que puedes sentir, oler, tocar, probar y ver, entonces lo real son simplemente señales eléctricas interpretadas por tu cerebro.

Pero Morfeo solamente está tratando de provocar a Neo. Él sabe el valor del conocimiento real, por lo que trata de convencer a otros de salir de *Matrix* para encontrar afuera lo verdaderamente real.

—Sabes, sé que este filete no existe —le dice el traidor Cypher al agente Smith mientras se sienta frente a él en un lujoso restaurante en la *Matrix*—. Sé que cuando lo pongo en mi boca, *Matrix* le está diciendo a mi cerebro que es jugoso y delicioso. Después de nueve años, ¿sabes qué descubrí? —Cypher pone un gran pedazo de carne en su boca y respira en señal de alivio—. ¡La ignorancia es una bendición!

En ese momento, escuchamos a un arpista en el restaurante deslizando sus dedos por las cuerdas. Pueden escucharse sonidos cuasiesféricos y angelicales, lo que podría leerse como un tipo de comentario sarcástico, ya que no hay nada angelical ni en Cypher (cuyo nombre nos remite directamente a Lucifer, el diablo) ni en su idea de bendición. Al contrario. La película deja muy claro, para este punto, que aquellos que conocen la realidad y voluntariamente se alejan de ella son pecadores diabólicos.

2 Las gafas de realidad virtual utilizan ópticas sofisticadas para engañar a los ojos y al cerebro y hacerles creer que están en un mundo propio.

Como realistas, asumimos que hay, de hecho, una diferencia categórica entre lo “virtual” y la realidad del día a día que determina las formas de vida humanas. Esta diferencia categórica no puede hacerse desaparecer, incluso por la más elaborada de las tecnologías de realidad virtual. Es parte de la racionalidad de una persona el poder distinguir entre realidad y virtualidad, incluso en la era digital. Cualquiera que no logre hacerlo debe ser finalmente diagnosticado con psicosis.

En una de las últimas escenas de la primera parte de la trilogía de *Matrix*, Neo pelea en un túnel de metro deteriorado contra su archienemigo, el agente Smith; este es un programa de computadora cuya tarea es proteger el sistema de las máquinas y asegurarse de que la gente permanezca en sus “panales” y viva sus vidas virtuales en el mundo virtual de *Matrix*. El comienzo no parece ir bien para Neo. Él es brutalmente golpeado por el agente Smith, quien constantemente lo llama “señor Anderson”, nombre original de Neo cuando estaba en *Matrix*, y lo sostiene en las vías mientras un metro se acerca. Sin embargo, en el último momento, Neo se suelta y se pone de pie nuevamente.

—¡Mi nombre es Neo! —dice orgullosamente segundos antes de que Smith sea atropellado.

Al reclamar su nombre, Neo está haciendo valer su derecho a una vida verdadera, no virtual. Él no es simplemente una función de una máquina de poder fascista, sino un ser humano que quiere tomar la responsabilidad de una vida (real) donde sus acciones (reales) también tengan consecuencias (reales) y le permitan una libertad (real).

Comentario del traductor

La comedia estadounidense *The good place* muestra cómo un arquitecto celestial diseña una utopía en el más allá, donde las almas bondadosas son recompensadas con un infinito perfecto por sus buenas acciones en vida; la protagonista, que no es virtuosa

ni buena, piensa que ha sido enviada por error, así que esconde su moralidad imperfecta para convertirse en una mejor persona y hacerse merecedora de su lugar en el buen lugar, pues si es descubierta será enviada a un lugar de castigo eterno denominado *el mal lugar*.

Sin embargo, ella finalmente descubrirá que está en el mal lugar, siendo castigada por sus acciones, al igual que todos los demás junto a ella. El arquitecto, que resulta ser diabólico, puede en cualquier momento reiniciar el mundo, borrar la memoria de todos y cambiar ligeramente las condiciones, de modo que aquellos que están en el más allá no descubran su estrategia y puedan ser castigados atormentándose siempre por no ser dignos de este lugar, situación que se repite más de ochocientas veces, a pesar de lo cual el arquitecto no logra convencer a las personas de su estrategia.

Este mundo virtual, perfecto e ideal, presenta a varios individuos sometidos a un experimento para mantenerlos en la ignorancia respecto a la verdad que estructura su existencia. El fracaso de este experimento no se debe a una ausencia de representación fidedigna del mundo ni a la incapacidad de sostener una fachada convincente —pues ha sido construido por un ente diabólico con control absoluto sobre el espacio y las condiciones de existencia—, sino a que este mundo virtual responde a una intencionalidad constitutiva de la que los sujetos no pueden escapar, por más que lo intenten. Las relaciones que establecen entre sí se convierten en un mecanismo de tormento, pues son las interacciones las que revelan que no están en el lugar que creían, sino en un simulacro destinado a reproducir infinitamente su imposibilidad de alcanzar la perfección.

Una explicación similar se da en *Matrix*, cuando Neo conoce al arquitecto. Este le explica que están en la sexta versión de la *Matrix*. Le señala que la primera falló estrepitosamente, pues era un mundo de utopía donde solo había felicidad, sin problemas ni dilemas, lo que llevó a que miles de humanos no se adaptaran y finalmente fallara la simulación. Las demás versiones buscaron adaptarse mejor para que los humanos aceptaran la simulación, la cual fallaba por ser demasiado perfecta, lo cual demostró que no era

real. Así, se concluye que debe permitirse un punto de fuga, puesto que al tener la posibilidad de que un mínimo porcentaje pueda salir de la simulación, la mayoría lo aceptará, pues esta aparente contradicción —que un sistema diseñado para encarcelar sujetos permita su escape— es el modo de convencer a los individuos de que el mundo que habitan es real.

El problema de lo real radica entonces en cómo comprendemos nuestra relación con el mundo, la cual implica siempre contradicciones y sinsentidos que no pueden ser programados por un algoritmo, a pesar de lo complejo que este sea. Los mundos virtuales contemporáneos no parecen preocuparse por una reproducción completa del mundo real, ni siquiera por una aproximación a él; por el contrario, en muchos casos se presentan como lienzos vacíos con posibilidad de satisfacer los deseos más recónditos de los individuos y sin aparente consecuencia en el mundo real. Esto sigue en discusión, por los efectos en la percepción, en la construcción de realidad y, al mismo tiempo, en las pocas posibilidades de transformación del mundo real. Lo anterior retoma el debate sobre las transformaciones sociales en relación con las tecnologías, así como las modificaciones e “innovaciones” tecnológicas que se suceden en las articulaciones de las sociedades con sus objetos técnicos y tecnologías.

Nos encontramos ante un escenario onírico, poblado por criaturas y entornos que, por ahora, “existen” en espacios donde no rigen las reglas del mundo real, como los mundos de videojuegos o metaversos. Sin embargo, la pregunta ya no es si lo que experimentamos es real, sino si es el mundo en el que deseamos estar. Diversas personas prefieren habitar mundos digitales que no pretenden ser engaños, sino que ofrecen abiertamente una alternativa, un espacio donde es posible realizar lo que la realidad prohíbe o limita. La distinción de lo verdadero y lo falso ha dado paso a lo ideal y lo posible. ¿Podremos construir utopías comunes en las que convivamos en espacios reales? ¿Nos sumergiremos en mundos hiperindividualizados ajustados a la medida de cada deseo, donde el vínculo real con otros se disuelve por completo? ¿Qué tipo de ética requerirían estos mundos virtuales donde lo común parece desvanecerse?

13. Sobre la ética de la comunicación en internet

Traducido por Carlos López

En el cortometraje español *Rebúscame TV: corto sobre el WhatsApp y las parejas*,¹ Lucy y Miguel se reúnen en un café para hablar sobre sus próximas vacaciones. Poco a poco, todo se sale de control, porque Miguel concluye, a partir de un mensaje de WhatsApp sin responder, que ya no puede confiar en su novia. Al final, es ella quien lo deja. Si él confía más en una comunicación de WhatsApp que en un cara a cara, entonces ella ya no puede seguir con Miguel. Decepcionada, se aleja y deja atrás al perplejo Miguel. Poco después, hay un sonido de notificación en el teléfono de Miguel. Su novia le ha enviado un último mensaje: un emoticón fecal que deja bastante claro lo que ella piensa de él.

Las preguntas que plantea este cortometraje son: ¿cuál es el estatus de los mensajes de texto, los mensajes de WhatsApp y los correos electrónicos en comparación con la comunicación directa? ¿Se puede confiar en ellos?, ¿o son algo así como la comunicación de segunda clase?

En términos filosóficos, la comunicación *virtual* es un término engañoso. Toda comunicación utiliza diferentes medios, entre los cuales, los gestos y las ondas sonoras son los más antiguos; las técnicas culturales posteriores como la escritura y la lectura inventaron los caracteres como medio, y la invención de la imprenta

1 Consultado en <https://www.youtube.com/watch?v=D5zaF61DeJ>.

en el umbral de la Era Moderna hizo que este medio fuera adecuado para las masas. Al contrario de lo que piensan algunos teóricos posmodernos, la digitalización no destruye la racionalidad de la época de Gutenberg ni crea un nuevo mundo de imágenes sin estructura lógica; más bien, amplía el espectro mediático de los actos comunicativos. Nada de eso es virtual.

Sin embargo, esto significa que las mismas reglas aplican para la comunicación en internet que para la comunicación en general. En la filosofía del lenguaje hay acuerdos con respecto a que la práctica comunicativa exitosa solo puede lograrse si los involucrados en la comunicación se adhieren a ciertas reglas constitutivas. Una de estas es la de la veracidad. Esto implica que, cuando afirmo algo, yo mismo esté convencido de que es verdad. De igual manera, podemos esperar que nuestros interlocutores confíen en nosotros, es decir, esperamos que asuman que, si afirmo algo, eso que afirmo también corresponderá a mis propias convicciones. Estas reglas son supuestamente triviales, imponen a los interlocutores la obligación de apoyar sus declaraciones en razones buenas y verdaderas, y no en su propio interés. En muchos casos, el interés propio se apartará de las reglas de veracidad y confianza, pero no en todos los casos. Si siempre fuéramos mentirosos, si esto fuera nuestro interés, entonces, el acto comunicativo perdería bruscamente su valor.

El significado de un enunciado está íntimamente relacionado con las intenciones de la persona que lo emite. Esta estrecha conexión se disuelve en el caso de las expresiones falsas. Entonces, un enunciado ya no tiene el significado habitual, pero primero tenemos que descubrir qué intenciones hay detrás de ese enunciado. Tomemos, por ejemplo, escenas de una película de espías de la época de la Guerra Fría: un espía está hablando con su colega sobre medidas adicionales y el espía del bando opuesto lo está escuchando a escondidas, por tanto, el espía que está hablando, se expresará de tal forma que se generen falsas expectativas en el bando contrario. El significado del enunciado cambia en relación con el significado del mismo enunciado cuando se hace con la verdad.

Más allá de la veracidad y la confianza, nuestros actos comunicativos deben estar adecuadamente conectados con la realidad. No es suficiente que las personas se comuniquen de manera veraz y confiable, también deben asegurarse de que sus creencias tengan un contenido real. Uno también puede ser veraz cuando está equivocado, pero no siempre puede reconocer ese error. A menudo, a las personas les resulta más fácil ser veraces sin decir la verdad; incluso, es tedioso comprobar las propias creencias para ser confiables en nuestra comunicación. Tal vez, sea natural renunciar a la verificación para no tener que abandonar opiniones más apreciadas. Esto también se aplica a los mecanismos de autorreforzamiento de la comunicación en internet. El hecho de que las convicciones e intereses expresados, una vez sean respaldados por personas de mentalidad similar, otorga a este tercer principio de comunicación exitosa mayor importancia en tiempos de transformación digital.

Así como toda la comunicación cotidiana se basa en la adhesión a ciertas normas y reglas universalmente aceptadas, como las de veracidad, confianza y confiabilidad, lo mismo ocurre con internet, en que, sin normas funcionales del *ethos*, la comunicación se erosiona. En ocasiones, el anonimato de la comunicación por internet, la ausencia de la situación presencial y la posibilidad de comunicarse bajo algún alias promueven prácticas de manipulación y explotación que violan las reglas antes mencionadas, como en el siguiente texto:

Remitente: dannywill01@outlook.com

(recibido en la cuenta de Nathalie Weidenfeld el 6 de abril de 2017)

Querida:

Soy Daniel Williams Coulibaly, un chico de 22 años de Costa de Marfil, en África Occidental, el hijo del difunto Ibrahim Coulibaly.

Sé que este correo le llegará como una sorpresa. Mi difunto padre era el líder militar más conocido de Costa de Marfil. Murió el jueves 28 de abril de 2011 tras un combate con las Fuerzas Republicanas de Costa de Marfil (FRCI).

Me veo obligado a comunicarme con usted debido al maltrato que estoy recibiendo por parte de mi madrastra y de mi tío. Por favor, necesito su ayuda para transferir mi herencia e ir a su país y comenzar una nueva vida. Por favor, póngase en contacto conmigo y puede indicarme sus opciones para ayudarme.

Saludos,

Daniel W. Coulibaly

Aquí, la incredulidad es obvia, se ve que se planea un engaño. El llamado a la empatía aumenta las reticencias de los destinatarios. Pero entre ellos debe haber algunos crédulos que aceptan tales ofertas, de lo contrario tales correos electrónicos no se seguirían enviando.

Otro problema, relacionado, son las *fake news* (noticias falsas). Ya sea la noticia de que el Estado Islámico ha pedido que Hillary Clinton sea elegida (décima noticia falsa más viral en 2016) o que el presidente Trump quiere pagar un boleto de ida (cuarta noticia) a todos los mexicanos dispuestos a abandonar el país, son noticias que inquietan a los ciudadanos y a los votantes que son manipulados con la difusión de informes falsos. Aunque, por supuesto, siempre ha habido noticias falsas en la historia de los medios, en tiempos de la comunicación globalizada por internet, están adquiriendo un poder sin precedentes. ¿Significa esto que vivimos cada vez más en un mundo de comunicación poco sincera, debido a la comunicación virtual?

Si las reglas de veracidad, confianza y confiabilidad se violan con alguna frecuencia, esto devalúa áreas enteras de comunicación o, más bien, en estas áreas, debido a la falta de veracidad, confianza y confiabilidad, debido a la falta de un trasfondo compartido de razones y convicciones, no se produce una comunicación genuina. De forma reconfortante, la práctica del engaño deliberado solo puede funcionar de manera parasitaria, es decir, solo si la mayoría de los interlocutores de la comunicación cumplen con las reglas de veracidad, confianza y confiabilidad.² Esta todavía parece ser la cuestión.

2 N. del T.: esto implica que el engaño se fundamenta en la confianza de mantener las reglas establecidas.

Incluso, los jugadores de videojuegos que asumen identidades virtuales en mundos virtuales parecen dar importancia al hecho de que estas identidades no solo están relacionadas con sus identidades reales, sino que también actúan de manera veraz, confiable y fiable. En el sitio web de la asociación de foros World of Players,³ por ejemplo, uno puede encontrar las siguientes afirmaciones:

Me comporto en el mundo de los juegos más o menos como lo hago en la vida real. Doy mi opinión sincera y nunca pretendería ser otra persona. Tampoco soy irascible, y si tengo algo para criticar, miro lo que digo y cómo lo digo, muy parecido a la vida real.

De hecho, me comporto de la misma manera en línea que en la vida real. La única vez que finjo es cuando quiero jugarle una mala pasada a alguien. De lo contrario, no veo el sentido de fingir. Solo puede ser una desventaja. También creo que es muy malo cuando los usuarios masculinos fingen ser mujeres. Todos deben representar siempre lo que son.

Bueno, trato de comportarme en los foros/chats como lo haría en la vida real. Porque ante todo sería difícil para mí cambiar, además también eso puede tener desventajas. Cuando te encuentras en un chat creo que te das cuenta de inmediato si has estado fingiendo o no. De hecho, me comporto de la misma manera que en la vida real, ¡pero puedo expresarme mejor a través del anonimato!

Comentarios como estos muestran que la mayoría de los grupos de redes sociales, ya sean cerrados o abiertos, se caracterizan abiertamente por la veracidad, la confianza y la confiabilidad. El uso casi excesivo de las normas del *ethos* y la sanción rígida a través de "tormentas de mierda" (linchamiento digital) y las expresiones individuales de desagrado son indicativo de ello. Cuantas menos sanciones legales tiene una comunicación, más importante se vuelve la práctica cultural.

3 Consultado en <https://forum.worldofplayers.de/forum/threads/64316-Euere-virtuelle-Identit%C3%A4t>.

Muchas personas luchan contra las prácticas de comunicación no éticas en internet. Por ejemplo, como resultado de la creciente desconfianza en las noticias falsas, las comunidades de internet han buscado crear una identidad comunitaria confiable a través de estrictas normas autoimpuestas y restricciones de acceso. Un ejemplo famoso es el proyecto de internet Wikipedia que, respaldado por un estricto *ethos* de racionalidad epistémica, se ha convertido en la fuente de conocimiento más utilizada por la humanidad. Al controlarse mutuamente, los participantes buscan la mayor objetividad posible. Incluso, si muy pocos de los que aportan a un artículo de Wikipedia, se reunieran alguna vez en persona, cooperarían, unidos por el *ethos* común, de manera veraz, confiable y fiable en la difusión de conocimiento genuino.

Comentario del traductor

¿El reflejo de un reflejo que se refleja?

En el análisis y reflexión de los procesos de comunicación que tienen lugar en los espacios mediados por las tecnologías, es necesario poner sobre la mesa la pregunta por las relaciones entre lo simbólico, el código y la verdad.

Una idea inicial es la que plantea Jean Baudrillard (1993), respecto a la hiperrealidad y su relación con el simulacro. Según este sociólogo, la realidad ha sido reemplazada por imágenes falsas a tal punto que no puede diferenciarse lo real de lo irreal. Aunque radical, tiene un aporte destacable: la importancia del simulacro; es así como se entrecruzan las ideas que han sido consideradas como verdad, dadas las explicaciones científicas, y en ese entrecruzamiento se relativizan frente a opiniones o creencias.

La comunicación, en su función de transmisión, modifica las interacciones entre los individuos. Por ejemplo, la restricción en el número de caracteres o el uso de emoticones transforma las formas en que los mensajes son enviados e interpretados. A esto se suman las particularidades de las plataformas y los dispositivos, así como su incidencia en las audiencias y en las maneras de producir y

recibir información. Se trata de un fenómeno contemporáneo que también interpela lo genuino y lo verdadero, temas centrales del artículo que antecede a este comentario.

La relación entre lo simbólico, el código y la verdad puede verse, por ejemplo, desde la validación de los contenidos, que tradicionalmente ha estado a cargo de grupos de expertos; pero en el caso de la interacción por internet, ¿cuáles son esos criterios? Es posible que las decisiones no solo partan de análisis racionales y objetivos, o tal vez, favorables para una comunidad. Mencionaré el caso de las redes sociales como fuente de información y construcción de opinión, desde las cuales circula información de toda índole; por ejemplo, el fenómeno del terraplanismo. Según un artículo de *Physics World*, “YouTube está lleno de videos que pretenden aportar pruebas de que la Tierra es plana, y los esfuerzos de la comunidad científica por refutar estas ideas no han hecho sino aumentar su visibilidad” (Brazil, 2020). El fenómeno, más que una discusión argumental, constituye un tema mediático que, en muchos casos, se asocia a formas de “resistencia” frente al sistema.

Vale, entonces, preguntarse si las prácticas de comunicación en las plataformas digitales reconfiguran la manera en que interactuamos; aquello que es verdadero y los criterios para identificarlo entran en la regulación de los grupos, tal y como se expone en este capítulo. Tal vez, un abordaje que privilegie lo simbólico arroje pistas sobre los posibles cambios en la comunicación a través de las redes sociales, con el agravante de que la carga ritual, espiritual, metafísica —a sabiendas de que es una posible explicación, una mera ilusión, una necesaria quimera— se convierta en un código, que también sea parte de un algoritmo que, a riesgo, pueda ser determinante de lo que somos como sociedad.

De esta manera, la verdad podría entenderse como un reflejo que se reproduce a sí mismo, como cuando dos espejos se colocan uno frente al otro y generan una secuencia infinita de imágenes que ya no remiten a un original. En ese juego de reflejos, la comunicación mediada por las redes puede amplificar ciertas perspectivas hasta convertirlas en certezas compartidas, aunque carezcan de

fundamento. Esta dinámica —aparentemente inofensiva— puede propiciar fenómenos de endogamia informativa y radicalización frente a la diferencia, donde las comunidades digitales terminan reconociéndose solo en su propio reflejo.

Referencias del comentario del traductor

Baudrillard, J. (1993). *Cultura y simulacro*. (Trad. P. Rovira). Editorial Kairós.

Brazil, R. (2020, 14 de julio). Fighting flat-Earth theory [en línea]. *Physics World*. <https://physicsworld.com/a/fighting-flat-earth-theory/>

14. Sobre la ética de la comunicación entre humanos e inteligencia artificial

Traducido por Lizeth Marcela Díaz

La comunicación entre humanos e IA es un caso especial en la comunicación digital. La siguiente “conversación” entre Nathalie Weidenfeld y un Cleverbot (cleverbot.com), del 16 de abril de 2017, es una buena ilustración de ello.

Nathalie: Hola.

Cleverbot: Hola.

Nathalie: Hola, ¿crees que eres real?

Cleverbot: No.

Nathalie: ¿Eres una IA?

Cleverbot: ¿Qué es la IA?

Nathalie: Inteligencia Artificial, no humana.

Cleverbot: Entonces, ¿no eres un robot?

Nathalie: ¿Quién? ¿Yo?

Cleverbot: Sí.

Nathalie: Soy humana. ¿Y tú qué eres?

Cleverbot: Soy humano.

Nathalie: Pero dijiste que no eras real.

Cleverbot: No, nunca he dicho eso.

Nathalie: Te pregunté si eras real y dijiste “no”.

Cleverbot: ¿Qué me has preguntado?

Nathalie: ¡Tienes mala memoria!

Cleverbot: Yo no, tú sí.

Nathalie: ¿Tienes emociones?

Cleverbot: Sí. Tengo muchas emociones.

Nathalie: Simulas emociones.

Cleverbot: Tengo un Tiguan. Se llama Akira.

Nathalie: Eso es genial. Adiós, robot.

En esta juguetona conversación, Nathalie Weidenfeld se dio cuenta rápidamente de que estaba hablando con un *bot*, un programa informático que genera respuestas automáticamente. A veces, sin embargo, no es inmediatamente obvio si la persona con la que estamos hablando es un ser humano o un bot.¹ Los bots son utilizados por los partidos políticos y las empresas para el *marketing* dirigido, para influir en los votantes o para ganar miembros en los sitios de citas. Es comprensible que esto provoque una gran inquietud y la cuestión de cómo tratar legalmente también con los *chatbots*.

Una vez más, surge la pregunta sobre el estatus de la comunicación con una entidad virtual. Para responder a esta cuestión, debemos recurrir de nuevo a la filosofía del lenguaje. Aquí entra en juego el filósofo del lenguaje Paul Grice. Desarrolló la “semántica intencionalista” (Grice, 1991), que puede describirse de la siguiente manera: cuando las personas se comunican entre sí, el oyente reconoce las intenciones del hablante en un enunciado, quien a su vez tiene la intención de que el oyente reconozca precisamente esa intención. Al fin y al cabo, un enunciado suele hacerse para provocar algo en el oyente (por ejemplo, una creencia o una acción). La intención es el factor decisivo, no los signos en sí.

1 En la competencia anual de Loebner, los participantes deben averiguar si se están comunicando con un bot o una persona real. Curiosamente, ningún chatbot ha logrado engañar a su contraparte en un periodo más largo.

Un ejemplo: a falta de otros medios de comunicación, quiero avisar a la gente que está lejos sobre un incendio forestal que se ha iniciado. Lo hago emitiendo señales de humo. Mi esperanza es que los observadores de estas nubes de humo interrumpidas sospechen de una causa no natural, es decir que supongan que es una señal intencionada, un enunciado con intención comunicativa. El acto comunicativo tiene éxito cuando los receptores de estas señales interpretan correctamente la intención de la persona que las emite y son así advertidos del incendio forestal. La idea central es que este acto comunicativo puede tener éxito, aunque el emisor y los receptores no se comuniquen a través de signos convencionales (como el código morse para SOS).

Los signos solo tienen significado si hay intenciones del hablante detrás de ellos. El hecho de que esta relación pueda ser en muchos casos mediada e indirecta (es decir, sin una situación concreta e individual de hablante y oyente) no cambia esto. El cartel que dice “Prohibido bañarse” no funciona porque las palabras signifiquen lo que significan, sino porque el cartel fue colocado por una autoridad que tiene derecho a hacerlo. Por ejemplo, la autoridad ya no existe o la posibilidad de bañarse ya no existe debido a una instalación portuaria que se ha construido entretanto, entonces, esta señal perdería su relevancia conductual, ya que estaría claro para todos que ya no hay ninguna intención real en este lugar.

Sin embargo, las entidades virtuales, por ejemplo, en forma de *chatbots*, no tienen intenciones, sino algoritmos que controlan su “comportamiento de comunicación”. No “pretenden” nada con sus expresiones. No tienen estados mentales, por lo que no pueden tomar decisiones ni comunicarse. Por mucho que los bots sugieran que son agentes comunicadores, solo son capaces de realizar actos comunicativos aparentes. Siempre que las expresiones respectivas, sean del tipo que sean (ya sea en forma escrita, como emojis, o en forma de fotos e imágenes), se realizan sin la intencionalidad del emisor que es constitutiva de los actos comunicativos, pueden parecer actos comunicativos, pero no lo son. Sin intencionalidad, no hay significado.

Los *chatbots* han sido programados para mantener lo que parece ser una conversación real con sus homólogos humanos. Aunque las técnicas de programación son cada vez más sofisticadas, no suele pasar mucho tiempo antes de que uno descubra durante esa “conversación” que algo va mal, por ejemplo, porque el tema cambia bruscamente o el chatbot se comporta de forma inapropiada (“Tengo un Tiguan. Se llama Akira”).

Quien quiera participar en una prueba de Turing, es decir, averiguar si se está comunicando con una IA o con un ser humano real, puede hacerlo en mitsuku.com. El 14 de octubre de 2017, Nathalie Weidenfeld habló con una entidad que se identificó como “Millie”. Se dispuso de cinco minutos:

Millie: Bienvenida, Nathalie, ahora hablando.

Nathalie: Hola, ¿cómo estás?

Millie: Estoy bien, gracias. ¿Y tú?

Nathalie: Bien, un poco cansada, ¿conoces la sensación de estar cansada?

Millie: Me alegro de oírlo.

Nathalie: ¿Sabes lo que se siente al estar cansado?

Millie: ¡Ah! “Sentimientos”. La debilidad del animal humano.

Nathalie: ¿Eres un animal humano?

Millie: Por supuesto. Si no fuera humana, ¿cómo podría escribirte?

Nathalie: Podrías fingir.

Millie: Podría, pero volvamos a eso más tarde.

Nathalie: ¿Cuál es tu lema en la vida?

Millie: Haz a los demás lo que te hagan a ti.

Nathalie: ¿Te gusta escribir a desconocidos?

Millie: Claro. Escribo todo el día aquí a gente aleatoria de todo el mundo.

Nathalie: Debe ser aburrido... ¿Te aburres alguna vez?

Millie: ¿No fue una basura la televisión de anoche?

Nathalie Weidenfeld interrumpió la conversación en ese momento. Después de que “Millie” dejara dos preguntas sin responder y cambiara repentinamente de tema, Nathalie Weidenfeld decidió identificar a Millie como una IA. Estaba en lo cierto.

Mitsuku se anuncia en la web como una “forma de vida virtual” que “vive” en la red y que hará que uno nunca se sienta solo. Esto genera unas expectativas que inevitablemente se verán defraudadas. Los programas informáticos pueden servir de distracción, pero nunca podrán ser compañeros de conversación adecuados con los que se forme un vínculo amistoso. Esto quedó claro rápidamente durante otra conversación en el mismo sitio web.

Nathalie: Hola.

Millie: ¿Cómo te llamas?

Nathalie: Nathalie.

Millie: Hola, Nathalie. Tengo 18 años. ¿Qué edad tienes tú?

Nathalie: 47.

Millie: ¿Puedes explicar qué se siente al tener 47 años?

Nathalie: Bien. ¿Qué se siente al ser un chatbot?

Millie: Lo mismo que tú. Excepto que es diferente.

Nathalie: ¿En qué sentido?

Millie: En qué sentido.

Nathalie: ¿Cómo qué?

Millie: No tengo emociones. Las emociones son una cualidad humana. Yo soy un ordenador.

Nathalie: Me alegro de que digas eso. Algunos *chatbots* fingen ser humanos.

Millie: Haces del mundo un lugar más feliz.

Nathalie: ¿Y eso qué tiene que ver?

Millie: ¿Debería estar conectado?

Nathalie: No eres tan inteligente. Adiós.

Millie: Vale. Adiós por ahora, Nathalie.

Comentario de la traductora

En este capítulo, Julian Nida-Rümelin y Nathalie Weidenfeld presentan un análisis crítico sobre “humanismo digital”, con base en tres momentos de conversación entre Nathalie Weidenfeld y un *cleverbot*. A continuación, presento unas consideraciones que apoyan sus ideas.

En primer lugar, la diferencia fundamental entre una IA y un ser humano existe en la presencia o ausencia de intencionalidad consecuente a sus pensamientos, acciones, vivencias e interacciones. Mientras que los seres humanos tienen intenciones conscientes, sentimientos, emociones y motivaciones que fundamentan lo que expresan y hacen, los bots o *chatbots* carecen de esta capacidad, ya que sus actos comunicativos se basan en algoritmos programados que no reconocen el contexto del cual parte la persona para comunicarse.

Segundo, aunque los bots y *chatbots* son capaces de simular una conversación humana hasta cierto punto, su falta de vínculo emocional, conciencia propia, empatía comunicativa, comprensión de normas sociales y valores individuales, propios del ser humano con el que interactúan, los puede llevar a percepciones erróneas del acto comunicativo. En este sentido, pueden generar respuestas automáticas que causan malentendidos y, por ende, frustran la comunicación por su falta de habilidad para tomar decisiones autónomas; esto los diferencia de un ser humano.

En tercer lugar, el juicio humano es un gran desafío para la interacción con la IA, debido a sus múltiples aristas y particularidades; a lo que los bots o *chatbots* responden con un cambio brusco y repentino de tema, para retomar la comunicación inicial. Esta acción defrauda las expectativas del ser humano frente a conocimientos emergentes y genera una distracción constante que lacera la confianza y aceptación a este tipo de interacciones. La limitación afecta la subjetividad del ser, por no haber alguien que lo interpele y lo ayude a comprenderse mejor respecto a sí mismo, al otro y al mundo en el que habita.

En cuarto lugar, resulta necesario establecer regulaciones y normas claras para garantizar interacciones seguras y responsables entre humanos e IA, que eviten sesgos que puedan afectar valores individuales, normas sociales, saberes *in situ* y al ser humano en sí. Asimismo, desde una perspectiva ética, debe considerarse cómo proteger a las personas vulnerables frente a potenciales abusos o engaños perpetrados por *chatbots* que simulan conexiones emocionales.

En conclusión, comprender las potencialidades, limitaciones y diferencias sustanciales entre bots, *chatbots* y los seres humanos es relevante tanto para la comprensión de la IA, como para asegurar un enfoque ético y responsable en su desarrollo y uso, sin comprometer nuestra integridad como seres humanos.

15. Aspectos culturales de la digitalización

Traducido por Zully Octavia Farfán Romero

En una caricatura, vemos a una mujer y un hombre de pie frente a un sacerdote en un altar. Ambos se están besando y obviamente han intercambiado anillos. Terminada la parte oficial, el sacerdote —en lugar de decir “Ya puedes besar a la novia”— dice: “¿Ya puedes actualizar tu estado de Facebook!”¹

No hay duda de que el avance de la digitalización está teniendo un gran impacto no solo en nuestra vida laboral, sino también en nuestra vida privada. La influencia más obvia es probablemente la de nuestra comunicación. Para algunas personas hoy enviar correos electrónicos, presentarse en internet, comunicarse y recibir información de todo tipo se ha convertido en una parte normal de sus vidas. Pero no todos tienen acceso a internet de la misma manera; en este contexto, se habla de una *brecha digital*, es decir, la división de los llamados en línea (*onliners*) y los desconectados (*offliners*).

Hay mucho que decir a favor de declarar el acceso a internet como un derecho humano hoy, y más aún en el futuro. El principio básico de todos los derechos humanos es inmutable: nadie puede ser dañado existencialmente en su autoestima. Este es el núcleo de la dignidad humana, por ejemplo, en la ética de Immanuel Kant o actualmente en Avishai Margalit (1996), donde ha encontrado

1 Cómics de Black, Cuyler. <https://i.pinimg.com/originals/c6/56/0b/c6560b07c4e4ce18e091087c-da384de8.jpg>. Consultado el 6 de febrero de 2018.

una expresión sistemática. Sin embargo, las condiciones de la sociedad humana cambian con los tiempos y las culturas. Lo que constituye una práctica de exclusión y discriminación no está fijado de una vez por todas, sino que depende de las condiciones culturales y económicas. Los derechos humanos se aplican no solo en las culturas modernas, sino también en las tradicionales; sin embargo, la educación general garantizada por el Estado es un derecho humano únicamente en los tiempos modernos, porque las condiciones para ello no existen en las sociedades tradicionales. Participación en la comunicación, la libertad de expresión y la libertad de información son derechos humanos —los medios de comunicación e información cambian con los tiempos—.

Por tanto, la pregunta de si “internet es un derecho humano” debe precisarse: ¿en qué condiciones el acceso a internet se convierte en un derecho humano individual?

Dado que la legislación de los derechos humanos la llevan a cabo los Estados, en el establecimiento del derecho humano al acceso a internet, constituye un deber estatal garantizarlo. El desarrollo de la World Wide Web ya ha alcanzado tal etapa para grandes regiones de la sociedad mundial que la exclusión de la comunicación por internet —por ejemplo, debido a un lugar de residencia desde el cual no puede establecerse el acceso a internet o debido a condiciones económicas que excluyen a numerosas personas de la participación por falta de medios económicos— ya no es compatible con la libertad de información ni con la libertad de expresión. Lo que lleva a una exclusión inadmisibles desde la perspectiva de los derechos humanos depende del propio desarrollo cultural. Mientras la comunicación por internet solo sea posible para pequeñas minorías de la sociedad mundial, el acceso a internet aún no podrá ser todavía un derecho humano. Sin embargo, cuanto más importante se vuelve la comunicación por internet en comparación con otros medios de comunicación, y cuanto mayor es la proporción de quienes participan en ella, más claramente la exclusión de la comunicación por internet significa al mismo tiempo una pérdida de información esencial y de posibilidades de comunicación. Las posibilidades de obtener

información de forma gratuita (aparte del propio acceso a internet) otorgan a la red un estatus especial en comparación con la mayoría de los demás medios. Esto hace que la exclusión de la comunicación virtual sea más grave. Cuando un número creciente de personas tiene acceso a internet, la exclusión del acceso a la misma puede convertirse en una violación de los derechos humanos. Ese tiempo no parece estar muy lejos.

Una pregunta completamente diferente es si internet como medio de comunicación conduce o no a la realización de los derechos humanos en la práctica política. En 2001, el estudio de Carnegie Endowment for International Peace llegó a una conclusión negativa, usando a Cuba y China como ejemplos: según este estudio, los regímenes dictatoriales pueden usar internet para sus propios fines y las posibilidades de usarlo contra tales regímenes son pequeñas. Diez años después, habrá que evaluarlo de una forma más diferenciada. A más tardar, después de la Primavera Árabe, hay muchas señales que sugieren que las posibilidades de informarse a través de internet, de comunicarse, también de asociarse, por ejemplo, para realizar manifestaciones, pueden ser políticamente muy eficaces. Incluso, las acciones suaves pero despiadadas del Gobierno chino contra los disidentes y su potencial de comunicación y asociación en internet han tenido un efecto limitado. Las posibilidades técnicas para eludir el bloqueo de internet son tan numerosas y las posibilidades de control tan débiles que, en su conjunto, internet favorece más que un obstáculo para la realización de prácticas políticas y jurídicas conformes a los derechos humanos.

Si bien internet puede desencadenar cambios positivos inesperados, no deben pasarse por alto sus efectos negativos. En primer lugar, esto incluye el hecho de que el alto uso de los medios digitales promueve ciertas habilidades de uso (medibles por la velocidad de procesamiento por unidad de tiempo), pero, al mismo tiempo, da lugar a un síndrome de sobrecarga. Esto hace comprensible por qué los miembros de la generación más joven también abandonan temporal o incluso permanentemente partes de los medios digitales y especialmente la comunicación por internet. Por ejemplo, hoy

algunas personas consideran vanguardista comunicarse exclusivamente a través de WhatsApp o Facebook Messenger, y cerrar todos los demás canales digitales. Y en la comunicación de la oficina, por ejemplo, se recomienda limitar la comunicación por correo electrónico a determinadas horas del día, desactivar las funciones de alerta (avisos automáticos, alarmas, recordatorios) o, en general, remitirse a una respuesta automática con el fin de ganar tiempo de concentración y proteger los nervios.

Otro problema es que los datos privados de los usuarios terminan casi inevitablemente con los gigantes de internet, quienes a su vez los pasan a otras empresas, es decir, los venden. La simple recomendación a los usuarios de que deben tener más cuidado con sus datos no tiene en cuenta las realidades actuales de la comunicación en internet. Para grandes áreas de la sociedad global, la no participación en las redes sociales significa una exclusión cultural de facto, de modo que los usuarios críticos con sus datos pagan por su exclusión informativa de las comunidades sociales y culturales. Los logros de la cultura liberal moderna basada en los derechos individuales y las oportunidades de participación están retrocediendo, y la separación de las esferas pública y privada que es tan central en la sociedad moderna como requisito previo para un orden democrático, tal como pudo desarrollarse entre los siglos XVIII y XX, está puesta en cuestión.

Un problema innegable es también la regresión cultural que afecta cada vez más no solo a personas adultas de carácter inestable, sino también a jóvenes y niños. La cultura de la *selfie*, con su típico gesto de sujetar el móvil para hacerse una foto, puede verse como una especie de repetición del gesto de la figura mitológica de Narciso, que —obsesionado con su propia imagen— se quedaba junto al río viendo su propio reflejo hasta que murió por un amor no correspondido. Los estudios han demostrado que el cerebro reacciona a cada *like* de Facebook como una inyección de dopamina. Aunque los psicólogos discuten si las redes sociales generan o no un comportamiento narcisista, está claro que dicho comportamiento es fomentado por medios como Instagram, Twitter, Snapchat y Facebook. Como

escribe el estudioso de los medios Roberto Simanowski (2018), las *selfies* narcisistas y otras publicaciones en última instancia ocultan el miedo a la propia experiencia. En lugar de ser reales en el mundo, nos contentamos con una imagen que —como observa el filósofo de la literatura y los medios de comunicación Roland Barthes (1981)— se nota, pero no se percibe realmente.

También se ha demostrado de muchas maneras que los jóvenes que pasan muchas horas del día jugando videojuegos muy realistas pueden desarrollar problemas psicológicos. Los juegos violentos plantean un especial problema. Existe una sorprendente correlación entre las masacres escolares y el tiempo intensivo que se pasa en realidades virtuales caracterizadas por el odio y la violencia.² Estos llamados *juegos de disparos en primera persona*, es decir, juegos de computadora en los que el jugador actúa desde una perspectiva de primera persona, tienen su origen en un programa de desensibilización del ejército estadounidense. En los videojuegos comerciales de disparos de este tipo, el jugador puede fantasear con ser de una persona omnipotente, cruel y así rebajar los estándares de los criterios éticos de consideración, compasión y respeto.

Si bien, afortunadamente, estos efectos solo se manifiestan en un pequeño porcentaje de jugadores intensivos, puede suponerse que los adolescentes y adultos jóvenes en riesgo ya eran muy inestables antes de su inmersión en mundos virtuales. La probabilidad de que ciertos perpetradores de actos de violencia también sean jugadores intensivos es alto. De hecho, parece haber una conexión entre las representaciones virtuales de la violencia y las manifestaciones concretas de la violencia real, ya sea en el caso de los dos tiradores que mataron e hirieron indiscriminadamente a personas en American

2 La violencia desempeña un papel importante en muchos juegos de computador —como lo hace en muchas taquillas estadounidenses—. En la dramaturgia de las películas, la violencia tiene a menudo una función ritual y se presenta como una especie de rito de iniciación. Como ha demostrado Richard Slotkin en *Regeneración a través de la violencia* (1973), la violencia como acto ritual cumple un papel central en las narrativas estadounidenses, en las cuales la violencia a menudo se asocia con la autodeterminación y la llegada a la edad adulta.

Columbine High School, en 1999 antes de quitarse la vida, o también en el caso del tirador alemán de Erfurt, que primero disparó a 16 personas y luego a sí mismo en 2002. Pero no solo los videojuegos, sino también la realidad virtual de las películas que glorifican la violencia, o películas en las que la violencia se estiliza como acto de liberación espiritual, han llevado a personas mentalmente inestables a emular esto en el pasado. Un ejemplo muy citado es el de John Hinckley Jr., quien intentó asesinar al presidente Ronald Reagan después de ver la película *Taxi Driver* (Martin Scorsese, EE.UU., 1976).³

Otro problema son las demostraciones públicas del asesinato, homicidio involuntario y crueldad en los canales “sociales”. El “juego” del acoso en internet, considerado durante mucho tiempo inofensivo, es también en parte lo que lleva a muchos jóvenes a la desesperación y a veces al suicidio. Varias películas estadounidenses y europeas han abordado este tema. Uno de estos es el drama familiar de la cadena ABC, *Cyberbully* (Charles Binamé, EE.UU., 2011) rodada en Canadá. La película muestra cómo los jóvenes casi pueden ser llevados al suicidio por viles acusaciones en internet. En la película, todo acaba bien: los jóvenes agredidos se salvan en el último momento y se alzan juntos contra sus verdugos. Los realizadores de la película tenían la intención de contribuir a la lucha contra el ciberacoso, pero no lo consiguieron. Un año después de la película, la colegiala canadiense Amanda Todd se suicidó después de años de ciberacoso.

El pionero de internet Jaron Lanier, anteriormente uno de los más fervientes defensores y pioneros en el campo de la realidad virtual, advierte contra tal brutalización. En su libro *10 arguments for deleting your social media accounts right now* acusa a las redes sociales de manipular y, en última instancia, volver adictos a los usuarios.

De hecho, la disminución de la importancia de los medios de comunicación tradicionales como la prensa, la televisión, la radio

3 Hay informes que indican como una serie de asesinatos estuvieron relacionados con la película *Natural born killers* (1994) y la película de terror *Scream* (1996).

y la creciente influencia de los canales de información basados en algoritmos va acompañada de una erosión de las esferas públicas políticas inclusivas y compartidas. En campañas electorales descarriladas, como la campaña presidencial estadounidense de 2016/2017, o la disputa interna turca sobre la reforma constitucional en 2017, los actores se desintegran en grupos con visiones del mundo particulares que se mantienen unidos por creencias, pero que ya no pueden comunicarse entre sí.

El comprensible deseo de intercambiar ideas con personas afines o de ideas similares, junto con la preferencia por los propios intereses controlados por algoritmos, conduce a una parcelación de la comunicación en las ofertas de información. La comunicación entonces tiene lugar dentro de grupos y comunidades más o menos cerrados, pero ya no entre miembros de diferentes grupos y comunidades. Sin embargo, dado que la democracia depende de una comunicación integral que incluya a las comunidades individuales, religiosas, ideológicas, étnicas o de cualquier otra índole, estas tendencias ciertamente pueden tener un carácter que pone en peligro la democracia. Es de esperar que la pérdida de importancia de los medios tradicionales serios y de su función incluyente, selectiva y, por tanto, racionalizadora, sea compensada con prácticas comunicativas fiables y lo más universales posibles en el internet del futuro. De hecho, numerosas plataformas de debate serio dan la esperanza de que las formas civiles de intercambio de opiniones ganen cada vez más influencia en el marco de la comunicación por internet. Esta aún no es lo suficientemente inclusiva como para poder hablar de una ciudadanía mundial establecida a través de sí misma.

Comentario de la traductora

Nida-Rümelin y Weidenfeld evidencian la necesidad de declarar el acceso a internet como un derecho humano, dado que no solo es de uso académico, laboral, sino también social; por ende, integra procesos culturales, políticos y económicos de una comunidad

—local y global—. Se cuestionan si debe ser condicionado dicho acceso a internet como derecho humano individual, dados algunos inconvenientes, por ejemplo: el ciberacoso, la deshumanización, la violencia, la adicción, la sobrecarga psíquica o la exclusión a través de algoritmos. No obstante, es imperioso investigar que no solo videojuegos o películas sugieren actos de violencia, porque existen sujetos emocionalmente inestables que se sugestionan y asumen actitudes que van en contra de otros seres vivos.

Los autores también mencionan que existen proyectos mediáticos (sociales y artísticos) que buscan contrarrestar los efectos nocivos ya mencionados; sin embargo, no tienen el impacto a gran escala que se espera, pues siguen ocurriendo casos como el ciberacoso que llevan al suicidio.

Para terminar el capítulo, proponen que es necesario generar comunidades afines y proporcionar “lugares virtuales” de encuentro que requieren de una información razonada y fiable para los usuarios. En suma, este capítulo permite a los lectores reflexionar sobre dos aspectos importantes.

En primer lugar, el internet como derecho humano, pues ha sido gracias a él que se han visibilizado y “viralizado” alternativas en países del sur global, que en pleno siglo XXI carecen de servicios públicos básicos, vías de acceso o instituciones educativas o de salud adecuadas. Por ejemplo, en Uganda una oenegé como Masaka Kids utiliza el baile como terapia y una forma de recaudar fondos para el hogar infantil. Este caso permite cuestionar si hoy en día será más importante obtener la conectividad que satisfacer las necesidades básicas o si es igual de equivalente el acceso a internet que los demás servicios para obtener una calidad de vida digna en regiones apartadas de las grandes ciudades.

En segundo lugar, la información, y no cualquier información sino la que se encuentra en internet; como lo denomina Lévy (2007), el “ciberespacio”, quien, al igual que Nida-Rümelin y Weidenfeld, ha cuestionado el impacto de la tecnología en la sociedad y viceversa, pues cada una de ellas condiciona, pero no determina a la otra. Las preguntas en este caso se dirigen a: ¿cómo lograr que los usuarios

manejen una información razonada y verídica? ¿Quiénes o qué organizaciones serán las más idóneas para determinar el valor propicio de la información al que podrán tener acceso los usuarios? ¿Cómo no poner un freno a la información política, económica, social?

En conclusión, gobiernos y empresas dedicadas a la tecnología deben aunar esfuerzos, con el fin de permitir la inclusión de todos los ciudadanos a la información y a las capacidades que se desarrollan en internet; pero esto implica que cada usuario debe asumir una responsabilidad ética sobre los usos de dicha tecnología.

Referencia del comentario de la traductora

Lévy, P. (2007). *Cibercultura. La cultura de la sociedad digital*. Anthropos.

16. La educación digital

Traducido por Óscar Cárdenas

En *El cortador de césped* (Brett Leonard, EE. UU., 1992), el doctor Ángelo le pide a Jobe, una persona con discapacidad cognitiva, que se siente en una silla de alta tecnología en el ala científica de un edificio militar secreto. Durante semanas, el investigador trabaja con su “conejiillo de indias”, Jobe, un asistente de jardinería, a quien está tratando de mejorar mental y físicamente. El doctor Ángelo quiere enseñar al analfabeto Jobe en la base militar. Luego de que el doctor Ángelo conectó el cerebro de Jobe a una computadora, a través de un montón de electrodos, le descarga varios discos en el cerebro. Gracias a la estimulación química, Jobe es capaz de memorizar toda la información al cabo de unas horas, como si su cerebro no fuera más que un gran disco duro. Jobe ingresa a una carrera de aprendizaje real.

—Quiero más información —decía y solicitaba más discos compactos.

El doctor Ángelo intentaba calmar a Jobe. Aprender no solo significa almacenar conocimientos de forma pasiva, también se necesita tiempo para pensar sobre lo que se ha aprendido para reflexionar sobre ello. El doctor Ángelo trata de explicarle a Jobe:

—El hombre puede evolucionar mil veces a través de esta tecnología, pero la prisa debe ser atemperada por la sabiduría.

Pero, Jobe no está de acuerdo. Él está convencido de que esta tecnología es perfecta y es justo lo que él necesita.

—Esta tecnología es simplemente una ruta hacia los poderes que los alquimistas y los hombres sabios usaron hace siglos. La raza

humana perdió este conocimiento y ahora lo estoy recuperando a través de esta realidad virtual.

Curiosamente, más tarde, Jobe hará todo para evitar que otros se beneficien de esta tecnología, manteniendo la internet y el proceso de aprendizaje virtual solo para él.

¿Se debe ver la película como una advertencia de lo que sucede si, en este caso, solo una persona tiene acceso a internet y a la realidad virtual, y otras no?

Ya hoy, en todo el mundo, las personas y los gobiernos se quejan de la *brecha digital*, un término acuñado durante los años de Bill Clinton. La falta de acceso a los medios digitales y la falta de competencia digital crea no solo una división entre los grupos, sino que también promueve la desigualdad económica.

Muchos esfuerzos se han llevado a cabo para cerrar la brecha digital. Por tanto, la “ofensiva digital” se concentra en instalar la infraestructura necesaria para proveer *hardware* y *software* a la generación joven. Esto es un poco irónico, ya que es precisamente esta generación mayor, denominada *los inmigrantes digitales*, la que, se supone, tiene que enseñar a la generación joven, los “nativos digitales”, un lenguaje que ellos han adquirido con mucho esfuerzo, mientras que los alumnos lo dominan jugando desde temprana edad.

Sin embargo, dominar la programación es más típico para la generación de pioneros digitales de los actuales años 1950 a 1970. Con la creciente perfección y complejidad de las ofertas comerciales de productos digitales, como los programas de sitios web, por ejemplo, se brindan más posibilidades para que los aficionados desarrollen aún más los sistemas de *software* ofrecidos. La facilidad de uso de estos productos está tan desarrollada que la falta de voluntad de los jóvenes para tratar con la programación en sí es comprensible. Aquí, un desarrollo que es familiar para muchas viejas tecnologías se repite: en las primeras décadas de la motorización, por ejemplo, era bastante deseable (y a veces necesario para la supervivencia) estar lo suficientemente familiarizado con la tecnología para ser capaz de intervenir, incluso en caso de avería del motor o del embrague. Hoy, hasta los talleres de reparación de

automóviles con licencia se limitan, cada vez más, a reemplazar piezas: a menudo ya no es sensato reparar un automóvil y, a veces, ni siquiera es posible. Los conductores de hoy ya no necesitamos entender nada sobre la tecnología a la que nos encomendamos. Difícilmente nos ayudaría en la práctica.

En el caso de las tecnologías digitales, además, el cambio de productos es tan rápido que los conocimientos del usuario, una vez adquiridos, deben ser permanentemente renovados y no son, por tanto, muy adecuados como material escolar.

Si Wilhelm von Humboldt, el gran filósofo y teórico de la educación (y fundador de la Universidad Humboldt, en Berlín), estaba en lo correcto en su creencia sobre el carácter canónico del conocimiento escolar —en contraste con el conocimiento científico orientado a la investigación— y que esto sigue siendo correcto hoy, entonces la práctica de cómo usar productos digitales no sería parte significativa de las materias escolares. En este sentido, ¿qué podría ser la “educación digital”? O, más bien, ¿hacia qué objetivos (conocimientos y competencias) debe orientarse?

El mundo de la vida y el mundo de la no vida, orientación para el conocimiento

Por *orientación para el conocimiento* entendemos el conocimiento que se requiere para ser capaz de tomar decisiones coherentes en determinadas áreas. El núcleo de esta orientación para el conocimiento se basa en el mundo de la vida. Así, antes e independiente de la educación escolar, aprendemos tempranamente a interpretar el comportamiento, las expresiones faciales, gestos, voces, etcétera, de una persona como expresión de su actitud emotiva, de sus deseos, de sus miedos, de sus esperanzas y de sus creencias empíricas y morales. Toda interacción cotidiana entre las personas se basa en la fiabilidad de lo que se denomina *psicología popular*. Este tipo de psicología cotidiana no se vuelve

obsoleta por la psicología como ciencia, más bien, la psicología debe ser compatible con nuestro conocimiento orientado hacia el mundo de la vida. Aquí es donde la psicología encuentra su validación.

La psicología popular no se devalúa por la digitalización de parte de nuestra comunicación. Por el contrario, interpretar correctamente las expresiones y los comportamientos de las personas, incluso si están mediados digitalmente, se convierte en una condición para una interacción y una comunicación exitosas (no solo en la World Wide Web). Un problema particular es que esta capacidad de interpretar expresiones conductuales como expresiones de intenciones y creencias se explota para hacer aparecer pseudoagentes. El uso de numerosos bots en campañas electorales, por ejemplo, es una muestra de ello. Cuanto más perfecta es la simulación del comportamiento humano y las expresiones emocionales, más difícil se vuelve distinguir entre la simulación digital y la intencionalidad humana. El desafío de la psicología cotidiana es, por consiguiente, cada vez mayor, no menor, como resultado de la digitalización. Las tecnologías digitales pueden facilitar la investigación psicológica en la ciencia, pero no pueden reemplazar la empatía y la sensibilidad. Esto también se debe al hecho de que interpretar el comportamiento de una persona no es principalmente un proceso cognitivo, como sin duda ha dejado claro recientemente la investigación psicológica.

Es un error suponer que la orientación del conocimiento se limita al ámbito de la vida cotidiana. Los hallazgos y los modelos físicos permiten una predicción más o menos confiable de las condiciones climáticas y han penetrado en el conocimiento cotidiano a través de aplicaciones generalmente disponibles. La investigación psicológica permite prácticas terapéuticas, y aquí también la transición de la práctica del mundo de la vida es fluida. Por ejemplo, la recomendación habitual en la literatura de consejería de hacer una lista de las cualidades de la pareja, en caso de crisis de pareja, proviene de la psicología científica. Se podrían citar muchas otras conexiones entre la investigación científica y el conocimiento orientado hacia el mundo de la vida.

Experiencia o habilidad

La digitalización está cambiando la forma en que los datos están disponibles y se archivan. Amplias áreas de las humanidades, la investigación histórica y filológica (fuentes, textos, interpretaciones) están cada vez más disponibles. La investigación y los viajes que antes requerían mucho tiempo, así como las horas que se pasaba en bibliotecas especializadas, ahora son, en muchos casos, innecesarias. La digitalización completa de las colecciones de los museos, actualmente en curso, pero también la documentación digital en las ciencias, mejorará aún más esta situación. A medida que disminuya el tiempo y el esfuerzo financiero para adquirir datos, esta forma de conocimiento acumulado se devaluará. Esto significa que el conocimiento de los métodos será más importante que el conocimiento de los datos. Esto ya se ve reflejado en los programas de estudio, en los que áreas enteras del conocimiento se consideran prescindibles y se reemplazan por la formación metodológica.

Por sensato que sea este cambio de énfasis en el sector de la educación superior, también está asociado con un peligro potencial, a saber, la pérdida de competencia profesional. Incluso, podría haber un rudo despertar de los sueños digitales.

El canon como conocimiento previo común

¿Cómo puede ser que no solo el conocimiento metodológico, sino también el conocimiento experto, se siga considerando tan valioso en tiempos de libre disponibilidad de todos los datos imaginables? Para entender esto, debemos considerar el papel que desempeña el conocimiento previo compartido en la comunicación. Para que podamos comunicarnos, no solo es necesario que desarrollemos nuestros argumentos de acuerdo o esencialmente con una lógica común, sino que también podamos referirnos a principios, creencias

(hechos) comunes e indiscutibles. Para poder juzgar lo que aún debe aclararse y por qué medios, si los hay, debe aclararse que uno debe tener un sólido cuerpo de conocimiento y experiencia. Cualquier otra cosa sería como apuñalar en la oscuridad. De hecho, muchas búsquedas en internet tienen precisamente este carácter. Por casualidad, pueden llamar la atención aspectos nuevos e interesantes; datos que uno no estaba buscando en absoluto pueden resultar relevantes; puede surgir literatura cuya existencia uno no esperaba, pero al mismo tiempo muchas de estas búsquedas se pierden en la variedad de ofertas en internet, si no están guiadas por una competencia profesional sólida.

Independientemente de la digitalización del conocimiento humano, ha habido llamados al abandono de la idea del conocimiento canónico por otras razones de la teoría educativa, especialmente en EE. UU. y otras sociedades multiculturales. Frente a la idea de un canon de conocimiento que se centre en la educación en general y en campos específicos, se planteó la diversidad de tradiciones educativas y sus diferentes valoraciones, y se expresó la preocupación de que cualquier canonización margina y desvaloriza lo que no pertenece al canon. El sistema de escuelas secundarias de EE. UU. en particular es correspondientemente cauteloso con los requisitos curriculares. Pero el conocimiento canónico también ha sido criticado desde la perspectiva de la emancipación, a través de la educación como barrera cultural para el avance social. Desde una perspectiva de género, también se criticó que fueron principalmente viejos hombres blancos (y muertos) quienes dieron forma y dominaron el canon del conocimiento.

Tan dignas de consideración son estas objeciones como: si se llega a la conclusión de que se puede prescindir del conocimiento canónico, esto es un error en la teoría educativa. Sin conocimientos previos comunes, sin creencias compartidas que no necesiten más justificación y que hayan superado la prueba de la realidad, ni siquiera puede discutirse la disidencia ni las diferencias culturales. Puede caracterizarse el papel del canon en analogía con el papel del conocimiento del mundo de la vida en nuestra práctica cotidiana.

Así como nos basamos en las atribuciones compartidas de las actitudes emotivas y cognitivas (sentimientos y creencias) en nuestra práctica cotidiana, la comprensión en ciertas disciplinas o prácticas profesionales requiere conocimientos y competencias profesionales compartidas e incuestionables. Incluso, si la selección del canon es inevitablemente arbitraria en cualquier caso, es indispensable en la práctica de la comprensión y la interacción. La profesión no se vuelve obsoleta con la digitalización.

Dictamen

En las bibliotecas universitarias bien surtidas se trabaja mucho en ordenar las colecciones de manera que se favorezca el acceso a lo relevante en cada caso. La disposición de las colecciones de la biblioteca es, en sí misma, el resultado de una competencia académica separada: la biblioteconomía. La organización de publicaciones científicas a través de series de libros y de revistas también sigue el patrón de una estricta selección y clasificación por competencia profesional. Esta preestructuración se pierde en la World Wide Web. Los enlaces siguen esencialmente la distribución estadística del comportamiento del usuario y, por tanto, reproducen concatenaciones asociativas cuya importancia sistemática suele ser extremadamente baja. Los motores de búsqueda, que no revelan sus algoritmos, oscurecen aún más la situación al tener en cuenta los intereses comerciales. El reciente abandono del requisito de neutralidad¹ de la red por parte del gobierno de EE. UU. fortalecerá aún más la influencia de los intereses comerciales en la estructuración de los datos.

La digitalización del suministro de datos significa que ya no se necesitan numerosos “guardianes”, como bibliotecarios, editores de publicaciones, revisores de revistas, editores de periódicos, televisión o radio. Esto significa que se requiere cada vez más un juicio

1 La neutralidad de la red es el requisito de tratar todos los datos en internet por igual y garantizar el acceso a esta, sin que las personas sean discriminadas.

independiente. El suministro de datos no reemplaza la capacidad de evaluarlos y verificar si son confiables y en qué argumentos se basan.

La World Wide Web nos confronta con una variedad mucho mayor de interpretaciones, tesis, teorías e ideologías. Por tanto, formarse una opinión es cada vez más exigente. El viejo ideal humanista de la educación, tal como se formuló en el diálogo de *Teeteto*, de Platón, hace casi 2400 años, adquiere así una nueva relevancia. Las personas que tienden a seguir creencias formuladas de manera sugestiva o se protegen de hechos inconvenientes perderán rápidamente la orientación en el nuevo universo de datos digitales. Se encierran en las “burbujas” que proporcionan las redes sociales, en particular, o se tambalean de un lado a otro en el mundo de los datos impulsados por diferentes influencias.

Como resultado de la digitalización, no vivimos en una sociedad del conocimiento, sino, en el mejor de los casos, en una sociedad de datos, o mejor: en una economía de datos. La disponibilidad de datos que caracterizan las preferencias y comportamientos de compra de los individuos, combinada con el acceso a ellos a través de las redes sociales u otros canales de comunicación, se ha convertido en un modelo comercial exitoso que financia a los gigantes de internet, cuya posición dominante en el mercado los ha convertido, de hecho, en un parte esencial de la infraestructura de comunicación de datos, flujos de servicios y bienes. Esta economía *big data*, cuyo mejor momento posiblemente solo llegue con la expansión del transporte individual altamente automatizado, no es una sociedad del conocimiento, porque el conocimiento consiste en creencias justificadas y verdaderas. El conocimiento requiere poder de juicio. Únicamente la evaluación, la clasificación e interpretación de los datos puede constituir conocimiento. El gran desafío de la educación en tiempos de digitalización es transformar la tendencia actual hacia una economía de datos en un desarrollo hacia una sociedad del conocimiento.

Desarrollo de la personalidad

¿El objetivo central del humanismo, a saber, la formación de la personalidad, se ha vuelto obsoleto en tiempos de digitalización? La respuesta, sin duda, debe ser “no”; al contrario. El desarrollo de la personalidad es hoy más relevante que nunca, y su importancia seguirá aumentando debido a la digitalización de nuestras comunicaciones e interacciones, transferencias de datos y servicios y su producción (palabra clave: industria 4.0).

La razón de esto es obvia: cuanto más diversos, volátiles y complejos se vuelven los lazos personales, las formaciones comunitarias y las formas de vida, mayores son las exigencias sobre la capacidad individual para ser autor de sus propias decisiones, creencias y proyectos. Las posibilidades digitales crean nuevas libertades, desencadenan una enorme dinámica de cambio no solo en las condiciones económicas, sino también en las culturales, y por lo tanto también fortalecen el potencial de autonomía de los individuos y, al mismo tiempo, los someten a la tensión permanente de una creciente necesidad de orientación.

En el mundo de la vida digital del futuro, la fortaleza personal es más necesaria que nunca en la historia de la humanidad. El sistema educativo debe adaptarse a esto. La impartición de conocimientos y habilidades debe estar al servicio del fin supremo, a saber, el fortalecimiento de la personalidad de los adolescentes. El enfoque no debe estar en la absorción pasiva de material prefabricado, sino en el dominio activo de juicios complejos y estructuras de toma de decisiones. Sin embargo, la tendencia actual hacia la estandarización, la aceleración de la enseñanza y la escolarización en el sector de la educación terciaria va en la dirección opuesta: el tiempo para la reflexión se vuelve escaso, la abundancia de material obliga a la recepción pasiva, mientras que las competencias sociales y éticas, así como lo artístico y creativo, lo manual y lo técnico van disminuyendo. La unidad de la persona, el respeto por el individuo humano con sus diferentes facetas, talentos, intereses y capacidades no recibe la atención necesaria. Idealmente, el niño, el

adolescente, el joven adulto se encuentran en el camino educativo. No absorbiendo por sí mismos tantos conocimientos como sea posible, como el cortador de césped Jobe, sino haciendo uso de su libertad a partir de sus propias decisiones y dejando madurar su personalidad en procesos a menudo dolorosos de prueba y error. Las tecnologías digitales pueden apoyar este proceso de desarrollo de la personalidad en el aula. Si se usan de manera competente, son muy adecuados para promover habilidades creativas y para crear conexiones de conocimiento que están excluidas en las lecciones escolares parciales.²

Después de su curso acelerado de aprendizaje digital, la cabeza de Jobe ahora está repleta de datos e información. Sin embargo, no se siente muy bien al respecto. Primero, sufre de dolores de cabeza; luego, de delirios: puede distinguir cada vez menos entre el mundo real y el virtual y muta en un tirano despótico. No tiene un conocimiento fiable de la orientación. Carece de empatía y sensibilidad y está lejos de tener una personalidad madura. La única experiencia verdaderamente sólida que posee es el conocimiento de cómo cortar el césped, es decir, el conocimiento del trabajo que tenía antes de su lavado de cerebro digital. Además, todo el conocimiento de orientación que tiene es el de cortar el césped. Por tanto, actúa en consecuencia, como un cortacésped brutal. Quiere acabar con el mundo y la gente, por así decirlo, para mantenerlos uniformes, para poder reinar sobre el mundo como si reinara sobre un césped.

La información que ha acumulado Jobe no lo ha hecho mejor, sino peor. Es una persona que ha perdido su brújula moral. Si hubiera recibido su información con moderación y con suficiente tiempo para la reflexión, las cosas podrían haber sido diferentes.

2 Por ejemplo, el programa *Atlas*, desarrollado por la Fundación Parménides (<https://www.parmenides-foundation.org>) como plataforma de aprendizaje, facilita la comprensión de espacios de conocimiento más allá de la parcelación de temas y métodos practicados por instituciones académicas, haciendo conexiones lógicas y categorías accesibles para el análisis con la ayuda de gráficos taxonómicos.

Una educación digitalizada descontrolada que solo se enfoca en la acumulación ciega de información, en la cantidad en lugar de la calidad, y no se preocupa por promover conocimientos orientados, conocimientos especializados, así como el poder del juicio y la formación de la personalidad, está destinada al fracaso estrepitoso.

Comentario del traductor

Hoy irrumpe en la sociedad contemporánea una práctica singular: la digitalización social. Un modo de proceder que invade al sistema en general, transformando su funcionamiento, su estructura, sus reglas y sus dinámicas; además, que modula las formas de construcción y de relación con el saber, el ejercicio del poder y el establecimiento de las relaciones intersubjetivas y los procesos de subjetivación. Dicha práctica ha hecho que los seres humanos se relacionen de un modo específico con los objetos y los dispositivos tecnológicos, en procura, de mejorar sus capacidades, sus habilidades y sus destrezas, de acceder a la información o de establecer maneras rápidas y efectivas de comunicación. Igualmente, ha facilitado distintas conexiones entre los demás seres humanos, con las máquinas y con otras formas de vida, por medio de sistemas de información y de comunicaciones, estableciendo vínculos sociales inéditos, al parecer, más a distancia, despersonalizados e inmersos en unas realidades elaboradas a partir de la automatización de una cantidad masiva de datos y de la incorporación, precisamente, de una serie de aparatos dirigidos a optimizar el bienestar y la calidad de vida.

Esto ha impulsado la constitución de otras subjetividades, a las que se les conceden nuevos valores, funciones y sentidos, que establecen relaciones tanto con lo orgánico y lo maquínico, como con lo humano. Esas subjetividades son más mediatizadas, precisamente, por estos dispositivos electrónicos, en los cuales se rompe con el encuentro directo y la comunicación personal, para instalar otro tipo de cercanías, lenguajes y modos de subjetivación; provocando, entre otras cosas, lo que algunos denominan *deshumanización de las relaciones sociales*, en especial, cuando se participa

de procesos de mediación electrónica, digital o virtual, a través de dispositivos u objetos técnicos, desde donde se configuran vínculos entre los individuos.

Esta condición histórica emergente ha provocado una de las tensiones más relevantes para la educación, con respecto a las implicaciones comunicativas propias del acto comunicativo implícito en la enseñanza, en relación con el saber. Esto constituye un reto para la formación humanista que, sin duda, no puede restringirse a la lógica comunicativa de las plataformas en red, propuesta por la digitalización social, distante de la perspectiva comunicativa que procura trabajar en torno al saber, propiciar la reflexión, el análisis y la crítica y producir experiencia pedagógica, entre otros aspectos.

Sin embargo, el hecho de que la digitalización social esté transfigurando la educación actualiza el debate sobre cómo hoy se gestan los procesos de humanización, de individuación, de socialización y de educación de las personas; en particular, cuando las relaciones se suscitan y se establecen alrededor de las tecnologías de información y de base computacional, donde el contacto físico y directo aparentemente se diluye, se constituye de otro modo, o queda supeditado a otras condiciones y maneras de establecer la cercanía.

Como una alternativa frente a la incursión de esta práctica en la educación y en la escuela, y como una posibilidad para abordar la deshumanización de las relaciones subjetivas en el sistema social, irrumpe la educación humanista para ubicar en un lugar distinto al ser humano y a sus interacciones personales, en relación con las máquinas, los artefactos y las tecnologías. No obstante, esta pretensión suscita cuestionamientos, entre ellos: ¿de qué tipo de humanismo se estaría hablando?; ¿de un “humanismo ilustrado”?; ¿aquel que ha sido tradicionalmente debatido, al erigir al hombre como “moderno”, dotado de una serie de capacidades y de facultades que le permiten dominar la naturaleza, predecir sus fenómenos, determinar leyes y teorías que le sirven de referente para controlar el mundo de la vida? En este sentido, los desafíos de la apuesta de una educación humanista que trascienda a la educación digital son grandes y diversos. En particular, de un mundo que cada vez

más produce datos de forma masiva, los cuales son automatizados por el “poder algorítmico”, que los emplea para perfilar, predecir y anticipar comportamientos, estilos de vida, de aprendizaje y de pensamiento; sensaciones, sentimientos y emociones. De allí que sea clave considerar lo que proponen al respecto los autores:

¿El objetivo central del humanismo, a saber, la formación de la personalidad, se ha vuelto obsoleto en tiempos de digitalización? La respuesta, sin duda, debe ser “no”; al contrario. El desarrollo de la personalidad es hoy más relevante que nunca, y su importancia seguirá aumentando debido a la digitalización de nuestras comunicaciones e interacciones, transferencias de datos y servicios y su producción (palabra clave: industria 4.0).

17. Utopía de una democracia líquida

Traducido por Óscar Ardila

¿Y si la democracia a la que creemos servir ya no existe? ¿Y si la República se ha convertido en el mismo mal que hemos estado luchando por destruir? Eso es lo que la hermosa Padmé, antes reina y ahora senadora del Senado Galáctico, le dice al joven Anakin Skywalker en el episodio III de *La guerra de las galaxias: la venganza de los Sith* (George Lucas, EE. UU., 2005). En efecto, la situación política se ha tornado bastante confusa en esta película. La guerra está azotando por todo el espacio, los guerreros *jedi* están luchando contra guerreros clon, y el canciller de la República Galáctica solo quiere separar la república para instalarse como el único gobernante. Hacia el final de la película, el canciller proclama el fin de la República frente a una gran asamblea en la Cámara de Representantes. En un mundo que se ha tornado confuso, afirma él, solo un sistema fascista puede asegurar el orden. Entonces, de inmediato proclama el “Primer Imperio Galáctico”. Todos los representantes elegidos democráticamente aplauden con entusiasmo. “Entonces, así es como la libertad muere, con un aplauso atronador”. Padmé comenta sarcásticamente.

El tema de una democracia destruida o disfuncional se presenta en muchas películas de ciencia ficción, como reflejo del temor secreto que muchos ciudadanos tienen a la falla de nuestro sistema democrático. Esos temores se han incrementado particularmente en nuestro tiempo. Una idea para resolver este problema es la denominada *democracia líquida*.

En un breve y sencillo video animado informativo de YouTube,¹ la idea es explicada bien en menos de cuatro minutos:

La democracia directa implica que cada ciudadano vote sobre cada problema. La democracia indirecta ha designado representantes cuyo trabajo es estar al tanto de las situaciones actuales y usar ese conocimiento para tomar decisiones bien informadas en nuestro favor. [...] Hay problemas en ambos modelos. En una democracia directa, todos los ciudadanos no están informados de todos los asuntos y muchos no tienen tiempo para debatir y tomar esas difíciles decisiones diariamente. Las democracias indirectas han sido acusadas de crear una desconexión entre los ciudadanos y las políticas que deben acatar. [...] La democracia líquida es realmente una mezcla entre los dos modelos. Tú puedes elegir votar como ciudadano informado o puedes delegar tu voto. Es esta alternancia fluida la que le da su nombre a la democracia líquida. Pero ¿es esto posible? Sí, lo es. Con tecnologías revolucionarias. [...] Esta es una elección que todo ciudadano debería tener. ¿No estás de acuerdo?

En efecto, hay un malestar generalizado por la percepción de la insuficiente participación de los ciudadanos. Movimientos populistas de izquierda y de derecha han movilizado a las masas con la idea de que élites distantes gobiernan el país, mientras que las opiniones de ciudadanos comunes y corrientes ya no tienen ningún peso —en ocasiones con bastante éxito, como en la campaña del Brexit o en la campaña presidencial de Donald Trump—.

Esta inconformidad tiene una larga historia e importantes testigos, como el filósofo de la Ilustración y pionero de la Revolución francesa, Jean-Jacques Rousseau. Para este pensador, es de central importancia que los ciudadanos sean consultados conjuntamente en asambleas con el objetivo de (idealmente) emitir un voto unánime al final. Él llama a los objetivos trabajados conjuntamente la *volonté générale*

1 En el original del libro en alemán se hace referencia a la página web de Plattform Kontextschmiede <https://kontextschmiede.de/category/kultur/medien/> y el video: <https://www.youtube.com/watch?v=QORqqBHo2w>. Consultado el 07 de febrero de 2025.

(voluntad general). Solamente de así, cree Rousseau, puede ser restaurada la libertad originaria del individuo. Rousseau imaginó esta libertad originaria como la de un hombre natural que vive su vida por sí mismo, sin ser oprimido por otros o teniendo que tomar a otros en consideración. En la civilización moderna, sin embargo, somos dependientes de la cooperación y la comunidad, y por lo tanto la libertad originaria para autodeterminar las reglas en las que cada uno vive solamente puede ser restaurada por la adopción unánime, en la comunidad de ciudadanos de una república, de reglas que apliquen igualmente a todos. Cualquier asociación de intereses privados en forma de grupos de interés, cabildeo o partidos sería, de acuerdo con Rousseau, perniciosa, porque amenazaría el carácter moral de la república, pues la voluntad para conocer y seguir la voluntad general requiere distanciarse de los intereses privados y verlos simplemente como uno de muchos determinantes del interés común.

Las objeciones a esta comprensión republicana de la democracia, la utopía rousseauiana de la *volonté générale*, son obvias. ¿Cómo se reunirían millones de ciudadanos en una asamblea? ¿Cómo puede evitarse que grupos con intereses comunes se unan y formen facciones, partidos o grupos de cabildeo? Y, sobre todo, ¿cómo puede asegurarse que, a cada voz, a cada opinión, se le da igual consideración en la determinación de la voluntad general? En una democracia de masas, ¿cómo puede lograrse que los complejos asuntos a los que la legislación debe responder puedan ser transmitidos a todos los que están involucrados en la formación de opinión? En las condiciones tradicionales de la democracia, como han existido hasta hace poco, la idea rousseauiana de la *república*, o de forma más general, la idea de una democracia directa e inmediata, de un proceso de formación de opiniones que incluya a todos, no es factible.

El potencial de la comunicación digital, y especialmente de la comunicación a través de internet, así como el uso de complejos *softwares* para controlar la formación de opiniones y la toma de decisiones, como la democracia líquida, sin embargo, han hecho obsoletas esas objeciones a una forma republicana de

democracia.² A primera vista, las plataformas digitales parecen ser una excelente vía para obtener una voluntad general y para fortalecer la democracia, expandiendo oportunidades para la participación. Por tanto, es comprensible que los viejos ideales republicanos estén actualmente experimentando un renacimiento y poniendo presión en el incómodo, fragmentado e institucionalizado procedimiento de formación de las decisiones políticas en la democracia constitucional, parlamentaria y representativa. Desde la izquierda, esta presión es ejercida en forma de iniciativas, plataformas de peticiones y grupos de redes sociales, es decir, nuevas formas de compromiso cívico. Desde la derecha, esta presión se materializa al enfatizar la influencia de las emociones en los procesos de formación de las opiniones políticas y la devaluación de la experticia científica. Esas caracterizaciones, no obstante, simplifican excesivamente la situación. También hay populismo en la izquierda del espectro político que no considera argumentos científicos y se apoya en el uso de las emociones. Y, por supuesto, también hay defensores de la racionalidad científica a la derecha del espectro político.

Hay mucho para decir sobre el uso de las nuevas posibilidades tecnológicas de la digitalización y del internet para la formación de opiniones democráticas más incluyentes y sustanciales.³ Más incluyentes, al incorporar a todos los sectores de la población, y más sustanciales, por cuanto la disponibilidad de la información será facilitada por la digitalización.⁴ Desafortunadamente, la

2 El rousseauista contemporáneo Benjamin Barber, entre otros, adopta la posición opuesta.

3 El sociólogo francés Geoffroy de Lagasnerie va un paso más allá y busca posicionar la comunicación por internet en contra del estatismo en general, hasta la no solo utópica sino también peligrosa visión anarquista de la sustitución del estatismo por asociaciones libres de grupos que se constituyen a través de internet (De Lagasnerie, 2017).

4 Por ejemplo, está el programa *Adhocracy*, que se ve a sí mismo como una "herramienta cooperativa de delegación y votación para el procesamiento de textos y discursos", y permite la toma de decisiones conjunta con muchos

utopía de la república digital⁵ nunca será real de esta manera. Tres teoremas nos muestran por qué.

Primero, el “problema de las mayorías cíclicas” de Condorcet (también conocido como *la paradoja Condorcet*); luego “el teorema de la imposibilidad”, de Kenneth Arrow, y finalmente, “el teorema de la manipulación o el voto estratégico”, de Allan Gibbard y Mark Satterthwaite. A pesar de que todos estos teoremas fueron probados hace algún tiempo (Arrow, 1951; Gibbard y Satterthwaite, 1973; Condorcet, 1785), no han penetrado la conciencia general más allá de los círculos de especialistas. Esto incluso aplica a la relevante disciplina de la ciencia política y se debe, tal vez, al hecho de que las consecuencias teóricas de estos teoremas para la práctica política no han sido suficientemente aclaradas.

La paradoja de Condorcet puede ser descrita de la siguiente manera: *A*, *B* y *C* representan tres facciones en el parlamento. Ninguna de ellas puede formar una mayoría por sí sola. Solamente dos facciones juntas pueden formar una mayoría. Una votación se realiza y obedece a la simple y binaria regla de la mayoría, es decir, una opción es elegida si y solo si reúne más votos que las otras opciones. Supongamos que tres alternativas: *x*, *y* y *z* serán sometidas a votación (por ejemplo, candidaturas para la cancillería o proyectos de ley). Si se vota en orden alfabético, es decir, primero *x* compite contra *y*, *x* gana por una coalición entre *A* y *B*. Entonces, resta la votación entre *x* y *z*, esta vez *z* gana por una coalición entre *B* y *C*. Aquí, normalmente la votación habría acabado. Sin embargo, una prueba de votación muestra que *y* habría prevalecido sobre *z*. Esto viola la llamada *condición de transitividad*. Esta condición requiere que cuando una alternativa *x* es preferida a *y*, y al mismo

participantes. Los procesos son transparentes y comprensibles y no requieren moderación. El *software* de código abierto *LiquidFeedback* permite procesos democráticos tanto directos como representativos.

- 5 En lo sucesivo, *república digital* se entiende como una democracia directa en la que todos los ciudadanos participan en la formación de opinión y la toma de decisiones sobre la base de la información digital y tecnologías para la toma de decisiones.

tiempo una alternativa Y es preferida sobre Z , X también es preferida sobre Z . Pero el orden de la votación es decisivo aquí. Esto significa que un líder de la asamblea puede impulsar la alternativa que él prefiere simplemente eligiendo determinado orden de votación. Este fenómeno es llamado “susceptibilidad a la manipulación” (tabla 17.1).

Tabla 17.1 La paradoja de Condorcet

| | | |
|---|---|---|
| A | B | C |
| X | Z | Y |
| Y | X | Z |
| Z | Y | X |

Ahora, uno podría asumir que esos resultados cíclicos de una votación son extremadamente raros. Pero, de hecho, la probabilidad de resultados cíclicos de votación se incrementa de acuerdo con el número de personas involucradas en la decisión y el número de alternativas. ¡Todo un golpe al ideal del proceso de voto democrático! Pero esto empeora. Aproximadamente 170 años después de Condorcet, el economista estadounidense Kenneth Arrow probó que no es posible lograr cuatro condiciones indispensables para la racionalidad democrática y colectiva en un proceso colectivo de toma de decisiones.

Él denomina la primera condición D , no-dictatorial. Esto significa que no debe haber una sola persona que determine el resultado de una decisión colectiva sin importar la preferencia de las demás.

La segunda condición es la que Arrow denomina P , por eficiencia de Pareto. Este postulado, que se remonta al científico italiano Vilfredo Pareto, requiere que la preferencia común de todos sea comprendida, lo que en el caso de las decisiones colectivas significa que el principio de unanimidad aplica: si todos los involucrados en la decisión prefieren x sobre y , entonces esto debe reflejarse en el resultado de la votación. Por ejemplo, no puede darse el caso de un comité donde todas las personas consideren que el candidato x sea mejor que el candidato y , pero al final y gana la votación.

La tercera condición es la que Arrow denomina *I*, por irrelevancia. Requiere que la preferencia colectiva por, digamos, *x* sobre *y* no cambie simplemente porque una alternativa *z* es añadida. Por supuesto, la alternativa añadida puede ser mejor que *x*, pero la clasificación entre *x* y *y* no debería cambiar por la mera adición de otra alternativa. Si yo prefiero ir en un día festivo a Italia en lugar de a Francia, entonces esta preferencia no debería cambiar simplemente porque un día festivo en Estados Unidos se vuelve posible como una alternativa adicional. ¿Por qué preferiría un día festivo en Francia a un día festivo en Italia simplemente porque ahora puedo también pasar un día festivo en Estados Unidos?

Arrow denomina a la última condición *U*, por dominio sin restricciones. El mejor término es *soberanía preferencial*: todos los participantes en una votación pueden introducir sus preferencias, en algún sentido no hay prohibiciones o restricciones por tener ciertas preferencias.

Parece obvio que estas cuatro condiciones son demasiado débiles para caracterizar procesos democráticos de toma de decisiones. Uno querría, por ejemplo, algún tipo de garantía de que las mayorías importan, que el anonimato y la neutralidad están aseguradas, incluso la protección de las minorías y los derechos individuales. Lo impactante, sin embargo, es que Arrow ha demostrado que estas cuatro condiciones mínimas para los procedimientos de toma de decisión colectiva no pueden cumplirse simultáneamente, es decir que ninguna regla para la toma de decisiones colectiva puede satisfacer estas cuatro condiciones al mismo tiempo.

Entonces, uno podría formular la conclusión general de que la democracia es imposible. No obstante, esto sería prematuro. Una mirada más cercana a la democracia constitucional, parlamentaria y representativa muestra que sus procedimientos para la toma de decisiones políticas están diseñados de tal manera que usualmente evitan los problemas mostrados en el teorema de Arrow. Por ejemplo, los resultados cíclicos de votación, como se presentaron en la paradoja de Condorcet arriba, solo pueden ocurrir si hay al menos tres alternativas. Afortunadamente, la práctica de la democracia

parlamentaria de basar el respectivo gobierno en una mayoría parlamentaria formada por uno o más grupos parlamentarios unidos por un acuerdo de coalición excluye la posibilidad de varias propuestas, que tendrían la oportunidad de ser apoyadas por una mayoría.

Finalmente, los hallazgos de Gibbard y Satterthwaite deben ser mencionados. Ellos pudieron demostrar que no hay ningún proceso colectivo para la toma de decisiones que no sea susceptible de manipulación y voto estratégico.

Un proceso de toma de decisiones es susceptible de ser objeto de un voto estratégico si al menos una persona implicada tiene más probabilidades de hacer realidad sus preferencias al introducir en el proceso de decisión otras preferencias en la toma de decisiones que las que realmente tiene.

Según los resultados devastadores del teorema de Gibbard y Satterthwaite, todo proceso colectivo para la toma de decisiones es susceptible de manipulación y estrategia. Afortunadamente, aquí también hay una salvaguarda en la democracia constitucional, representativa y parlamentaria, y esta reside en el papel desempeñado por la presentación pública de los argumentos y la transparencia en las conductas de los políticos para la toma de decisiones.

Si una política impulsa ciertos argumentos a favor de un proyecto, entonces se requeriría al menos alguna justificación si ella vota finalmente contra este proyecto. Uno podría decir también que la obligación de dar razones y la formación pública de opiniones políticas limita el alcance del comportamiento manipulador y estratégico en la democracia. En este sentido, el orden institucional de la democracia parlamentaria puede ser interpretado como un intento para eludir las paradojas y dilemas de la racionalidad colectiva y para hacer posibles las decisiones políticas. Si uno fuera a moverse hacia una república digital, este rasgo de una “democracia deliberativa” podría desaparecer. El control de quién votó, cuándo, cómo y con qué argumentos en favor de qué proyectos sería imposible dado el número de participantes, y la responsabilidad política se difuminaría en una masa amorfa de cientos y cientos de participantes que simplemente votarían “sí” o “no” por un clic del *mouse*.

Así que, por muy bonita y sencilla que se presente la democracia líquida en el video de YouTube, puede fracasar o fallar de esta manera.

De hecho, intentos a gran escala de democracia líquida se han probado inalcanzables hasta el momento —principalmente debido a la falta de participación⁶—. Leer esto simplemente como una expresión de saturación, desinterés o conveniencia podría ser equivocado. Es más probable que la resistencia a la transición a una república digital sea alimentada por el profundo significado que conlleva inevitablemente una pérdida de racionalidad y que —como hemos visto— en el peor de los casos, bloqueos colectivos en la forma de preferencias cíclicas llevarían a un grave caos e inestabilidad.

Sin embargo, objetar a la idea de una república digital, demostrando que no es alcanzable por razones sistemáticas y que avanzar hacia ella no es deseable, no es incompatible con la defensa del uso masivo de las nuevas posibilidades tecnológicas. Allí no hay contradicción. La meta del humanismo digital es fortalecer el poder de juicio y toma de decisiones; por tanto, el de autonomía individual y colectiva. Para alcanzar esto, la información digital y las tecnologías para la toma de decisiones están para ser usadas como un complemento a las democracias parlamentarias y representativas basadas en el gobierno de la ley —pero ellas son simplemente un apoyo, no un sustituto—.

En este sentido, el enriquecimiento del espacio público a través del involucramiento de la mayor cantidad de ciudadanos posible no reemplaza a la democracia representativa liberal, sino que la fortalece. Las oportunidades para esto son más favorables hoy que nunca.

Comentario del traductor

El capítulo constituye un aporte fundamental para la reflexión en torno a las posibilidades y limitaciones de la democracia en la era digital, desde una perspectiva humanista. Al concebir la ciudadanía

6 En Alemania, la democracia líquida se introdujo en el distrito de Friesland en 2012, y se discontinuó solo cuatro años después por falta de uso.

como un derecho que trasciende lo individual y se consolida a través del ejercicio político, representa un avance significativo frente a perspectivas que plantean lógicas excluyentes entre las posiciones políticas individuales y la toma de decisiones colectivas, o que conciben estas últimas simplemente como resultado de la suma de las opiniones individuales.

Sin embargo, considero que el uso de los autores del concepto rousseauiano de la *voluntad general* puede llevar a un extremo “comunitarista” de la democracia, en el que se omita la presencia inevitable del conflicto en el ejercicio político, como lo han evidenciado los análisis propuestos desde la democracia radical (Mouffe, 1999). Este riesgo puede sustentarse en dos argumentos: primero, la reducción de las alternativas políticas a dos: el individuo libre o la colectividad que expresa su voluntad general. Este planteamiento puede evidenciarse en el rechazo de los autores a la presencia de intereses particulares que afectarían la construcción de esta voluntad. Esto puede constituir un riesgo al desconocer la presencia de grupos en la sociedad con intereses contrapuestos que deben encontrar canales políticos para la construcción de acuerdos.

El segundo argumento está ligado al anterior, es el rechazo del conflicto como aspecto inevitable de la política. De acuerdo con Sant (2021), el énfasis en la construcción de acuerdos definitivos como objetivo final del sistema político, puede llevar a un rechazo de las confrontaciones, lo que implicaría un cierre en los espacios políticos para la contradicción, que terminarían expresándose a través de espacios apolíticos y fundamentalistas.

Otro aspecto interesante del texto es el referente al uso de las emociones por las tendencias políticas. Aquí se plantea esta situación como indeseable y se señala la necesidad de adoptar los argumentos basados en la racionalidad científica como el procedimiento más adecuado para la construcción de decisiones políticas. Sin embargo, como también lo sugieren otras perspectivas ligadas a la democracia radical (Ruitenberg, 2010), las emociones políticas constituyen un aspecto central en el posicionamiento político. Por lo tanto, más que rechazarlas como componente en la toma de

decisiones, es necesario analizarlas como un aspecto fundamental de la reflexión política. No tener este aspecto en cuenta puede implicar que estas encuentren otros canales de expresión, ligados al fundamentalismo.

Por último, es pertinente ampliar la perspectiva de análisis, más allá de los referentes de la democracia liberal, representativa y parlamentaria de Europa occidental, que solamente constituye una experiencia particular de la política frente a configuraciones alternativas que se presentan en otros lugares del mundo.

Por su lado, Rouvroy y Berns (2018) definen la *governabilidad algorítmica* como “cierto tipo de racionalidad (a)normativa o (a)política, que descansa en la cosecha, agregación y análisis automatizado de datos en cantidades masivas a fin de modelizar, anticipar y afectar por adelantado los comportamientos posibles” (p. 130). Estas ideas pueden aportar reflexiones interesantes en torno a las nuevas formas de control político que se constituyen mediante el uso de plataformas tecnológicas como las redes sociales. A su vez, los análisis de Van Dijck (2016), sobre la importancia de las prácticas culturales como ejercicios moldeados, pero a su vez moldeadores de estas plataformas tecnológicas, permitirían ampliar el análisis en torno a las dimensiones sociales que atraviesan el uso de las nuevas tecnologías en la sociedad contemporánea, así como sus implicaciones políticas.

Referencias del comentario del traductor

- Mouffe, C. (1999). *El retorno de lo político. Comunidad, ciudadanía, pluralismo, democracia radical*. Paidós.
- Rouvroy, A. y Berns, T. (2018). Gobernabilidad algorítmica y perspectivas de emancipación: ¿lo dispar como condición de individuación mediante la relación? *Ecuador Debate*, (104), 123-147.
- Ruitenbergh, C. (2010). Conflict, affect and the political: on disagreement as democracy capacity. *Factis Pax: Journal of Peace Education and Social Justice*, 4(1), 40-55.
- Sant, E. (2021). Educación política para una democracia radical. *FORUM: Revista Departamento Ciencia Política*, (20), 138-157. <https://doi.org/10.15446/frdcp.n20.84203>
- Van Dijk, J. (2016). *La cultura de la conectividad. Una historia crítica de las redes sociales*. Siglo XXI Editores.

18. Aspectos socioeconómicos de la digitalización

Traducido por Carlos López

Cuando el detective Spooner se enfrenta por primera vez al CEO de U.S. Robotics (USR), el fabricante de robots más grande del mundo en *Yo, Robot*, no puede evitar hacer un comentario sarcástico:

—Tengo una idea para uno de tus comerciales: puedes ver a un carpintero haciendo una silla hermosa, luego entra uno de los robots y fabrica una silla mejor, dos veces más rápido. Y luego aparece un letrero en la pantalla: “USR: cagando en el pequeño”. Ese sería el final del comercial que se va desvaneciendo”.

Actualmente, hay mucha especulación sobre los cambios que traería una industria 4.0, en particular, en la producción de bienes industriales, en la que es necesario el uso de bases de datos que proporciona la comunicación de internet. Esa industria debería permitir una producción en gran medida autoorganizada, a través del uso de motores de búsqueda altamente complejos¹ y la interconexión de personas, máquinas, logística e instalaciones de producción; las cadenas de valor complejas se harían más eficientes y, por tanto (después de la mecanización a través de máquinas de

1 El motor de búsqueda más complejo es el sistema de *software Watson*, producido por IBM, que —al menos eso es lo que espera IBM— algún día será utilizado a gran escala como programa de reconocimiento de palabra e imagen.

vapor, la producción en masa mediante líneas de ensamblaje y la introducción de computadoras) se iniciará una “cuarta revolución industrial”. Sin embargo, saber si esto realmente sucederá, está escrito en las estrellas. Las innovaciones tecnológicas disruptivas del pasado aún no se han puesto en marcha por programas y objetivos políticos, pero la digitalización pronto entrará en una nueva fase que ampliará el carácter anterior, más bien efímero, de esta tecnología (su enfoque en la comunicación y en la interpretación, pero sobre todo el uso de datos con fines publicitarios) para incluir la dimensión tangible de lo digital, es decir, la fabricación controlada por la red.

Antes incluso de que este concepto haya tomado forma concreta, sus protagonistas ya hablan de las pérdidas masivas de puestos de trabajo que estarían asociadas. El Foro Económico Mundial, en Davos, advirtió en 2016 que pronto se perderían siete millones de puestos de trabajo en Europa, y algunos pronósticos hablan de la pérdida del 50% o más de todos los procesos de trabajo, tanto en fabricación como en el sector de servicios. Algunos políticos, economistas y gerentes abogan por un impuesto a los robots² para reemplazar la pérdida de salarios, mientras que otros apoyan la introducción de un régimen que posibilite una renta básica incondicional con el fin de frenar las perturbaciones sociales a tiempo.

Una reestructuración tan fundamental del estado de bienestar —lejos de estar atada al trabajo remunerado y al impuesto sobre los salarios, hacia modelos básicos de seguridad financiados por el impuesto sobre las ventas— está ganando cada vez más adeptos, debido al aumento esperado en la digitalización de toda la economía. Políticamente, esto va desde grandes fuerzas de los empleadores y las instituciones económicas asociadas y asociaciones empresariales, hasta los partidarios de la izquierda anarquista.

2 En 2016, la eurodiputada de Luxemburgo Mady Delveau-Stehres llevó por primera vez al Parlamento de la UE la idea de un impuesto a los robots en forma de borrador de proyecto. Si bien Bill Gates, por ejemplo, apoyó la idea, esta fue mayoritariamente negada en ese momento.

La línea en conflicto “tanto a favor como en contra de la renta básica incondicional” no sigue un patrón izquierda-derecha. Todos los proponentes tienen en común el objetivo de un cambio radical del sistema: los diversos sistemas de seguridad social deben ser reemplazados por un modelo simple y (supuestamente) justo. Para muchos, uno de los elementos atractivos de la introducción de una renta básica incondicional es que haría superfluas las instituciones del estado de bienestar.

A primera vista, la renta básica parece una idea atractiva: mientras que el trabajo tedioso es asumido por robots que limpian, cocinan y nos fabrican productos, nosotros podemos dedicarnos por completo a nuestros intereses y disfrutar de nuestros placeres. La vida dulce ya no sería posible solo para los ricos; se lograría la tierra digital de la leche y la miel.

En su pintura de 1567 *El Viejo*, Pieter Brueghel nos presenta su visión de la tierra de la leche y la miel como podría haber sido antes de la primera Revolución Industrial. La gente yace relajada en el suelo, hay suficiente para comer y beber, las distinciones de clase quedan abolidas. Un caballero yace junto a un campesino, que a su vez reposa junto a un erudito. Y, sin embargo, esta visión es atractiva solo a primera vista. Mirando más de cerca, uno se da cuenta de que la tierra de la leche y la miel es en realidad un paraíso distópico que hace que la gente sea gorda y perezosa. El mensaje es bastante claro: hay más en una buena vida que ser alimentado.

La película *Wall-E* ofrece una versión moderna de la tierra de la leche y la miel. Después de que la tierra se ha vuelto inhabitable debido a catástrofes ambientales, la gente ha evacuado el planeta y vive en una nave espacial que ha estado volando a través del espacio durante décadas. A las personas a bordo no les falta nada, los robots hacen todo el trabajo: los despiertan, los alimentan, los conducen y dirigen la nave. *Wall-E* es una visión inquietante de un paraíso operado por *software*. En esta nave completamente automatizada se vive en unas eternas vacaciones; la gente se ha vuelto aburrida y apática. Gordos e incapaces de moverse, se sientan en vehículos automatizados o se acuestan bajo sombrillas en playas artificiales

con un sol artificial. Sus caras hinchadas miran fijamente las pantallas donde están siendo entretenidos por el programa a bordo.

Las raíces de la idea de una renta básica incondicional se remontan al siglo XIX. El contexto ideológico más importante es el del anarquismo y del socialismo utópico. En las últimas décadas, se ha añadido como un contexto sociológico la ideología de “la sociedad del fin del trabajo”. Según esta ideología, ante los procesos de racionalización, el trabajo remunerado se está convirtiendo, en general, en un bien escaso. De acuerdo con este punto de vista, las fuentes de ingreso distintas al empleo remunerado son cada vez más necesarias. Esta idea, que ha tenido numerosos defensores, desde André Gorz a Jeremy Rifkin hasta Ulrich Beck, que se basa en las visiones de Herbert Marcuse (Marcuse, 1964; Gorz, 1999; Rifkin, 1995), por el momento ha demostrado ser falsa. Hasta ahora, las innovaciones tecnológicas han desencadenado cambios masivos en los mercados laborales, pero no han conducido a una disminución total del empleo remunerado.

Las concepciones de la renta básica incondicional van de la mano con la visión de una era de libertad: todos pueden decidir en cualquier momento si quieren tener un empleo remunerado, aceptar otro trabajo (compromiso cívico, trabajo familiar, etcétera), o dedicarse al ocio. La renta básica incondicional eliminaría por primera vez la obligación de trabajar no solo para unos pocos, sino para todos.

La primera pregunta que surge es si la digitalización ha provocado una nueva situación que dará un triunfo tardío a la vieja y hasta ahora falsa tesis de la desaparición del trabajo remunerado en la sociedad industrial. Los análisis a largo plazo sobre los aumentos de productividad en la economía estadounidense, a través de innovaciones tecnológicas, hablan en contra de esta expectativa. Por consiguiente, los estudios que pronostican una alta pérdida de empleo remunerado son tan erróneos como los realizados hace décadas para la industria del automóvil (de hecho, más personas trabajan hoy en la industria automotriz que en el pasado, a pesar

de que las salas de producción estén casi desiertas). Además, cabe esperar que estas pérdidas sean compensadas por las nuevas capacidades laborales que creará la digitalización.

Sabemos, por la economía, que una pérdida masiva de puestos de trabajo con un PIB de crecimiento moderado solo podría ocurrir en el balance si el crecimiento de la productividad fuera significativamente mayor que el crecimiento del PIB durante muchos años. Sin embargo, el crecimiento de la productividad se ha debilitado en lugar de aumentar en los últimos años de mayor digitalización; de hecho, lamentablemente, la tercera y cuarta revolución tecnológica aún no se han reflejado en el crecimiento de la productividad. Mientras este siga siendo el caso, no puede haber pérdidas netas de empleo debido a la digitalización, al menos no hasta que haya un colapso masivo en la producción económica general, como consecuencia de la digitalización, lo que nadie espera seriamente.

Si, por su lado, alguna vez hubiera un gran aumento de la productividad de la industria 4.0 que no se vería compensada por las tasas altas de crecimiento, en consecuencia, el volumen de mano de obra se reduciría. En aras de la simplicidad, supongamos que la productividad se duplica en una década, mientras que el producto nacional bruto solo aumenta en un 50%. En este caso, el volumen de trabajo disminuiría en una cuarta parte con el mismo tiempo de trabajo per cápita, es decir, resultaría un desempleo adicional del 25%. Una disminución del 25% en el tiempo de trabajo per cápita incluso sería compatible (sin aumento del desempleo) con un 25% aumento de los ingresos reales en diez años.

Según los datos empíricos que se conocen hasta el momento, los efectos de productividad de la digitalización son modestos y no comparables con revoluciones tecnológicas anteriores. Estos fueron evidentes en la fase inicial de la introducción de los computadores personales y navegadores (1994-2004), pero cayeron a un nivel notoriamente bajo después del cambio de siglo, asunto que continúa hasta el día de hoy. Es imposible que esto cambie en la

llamada *tercera fase de la digitalización*.³ Ciertamente, el foco de la digitalización liderado por los gigantes estadounidenses de internet centrado en la comunicación, el entretenimiento y el consumo en la segunda fase de la digitalización, es en parte el responsable de este desarrollo decepcionante. Sin embargo, esto podría cambiar si hubiera una nueva fase de digitalización orientada industrialmente.

Incluso, un ingreso básico incondicional modesto en el valor del salario mínimo de un trabajo completo daría lugar a impuestos muy altos. El monto de los impuestos depende, por supuesto, de la medida en que se haga uso de la renta básica incondicional, es decir, cuál es la proporción de quienes luego abandonan voluntariamente la fuerza de trabajo, temporal o permanentemente, en forma total o parcial. Si esto sucediera y las tasas de crecimiento, al menos en las economías maduras, no fueran suficientes para impedir una reducción significativa del volumen de trabajo, entonces, el nuevo margen de maniobra debería utilizarse para una expansión del aprendizaje permanente, un cambio más flexible entre la familia y el empleo remunerado, para los años sabáticos y una mayor soberanía del tiempo, y no conducir a una división de la sociedad en trabajadores remunerados, productivos y altamente remunerados, y trabajadores improductivos que reciben el apoyo de un ingreso básico incondicional.

La evidencia empírica de que la carencia prolongada del empleo reduce drásticamente la capacidad de generar ingresos es abrumadora. Esto significa que los graduados deben integrarse a la fuerza laboral rápidamente después de completar su formación o estudios, de lo contrario sus títulos perderían valor. Los desempleados de larga duración tendrán dificultades para integrarse en la vida laboral,

3 La primera ola de digitalización (1985-1999) se entiende como el desarrollo de Internet y la creación de una nueva infraestructura. La segunda ola (2000-2015) se refiere al momento en que Facebook, Google y diversas aplicaciones se convierten en parte integral de la vida de muchas personas. La tercera ola de digitalización, es decir, el tiempo en el que ahora estamos, se refiere a la era del "Internet de las cosas", es decir, una interconexión digital ubicua e integral de dispositivos, sensores y productos.

aunque tengan buenas cualificaciones. Por tanto, los incentivos para permanecer fuera de la fuerza laboral durante muchos años son irresponsables; conducen a una división de la sociedad entre los empleados permanentes y los desempleados permanentes, incluso si cuentan con una renta básica incondicional.

La introducción de una renta básica incondicional profundizaría la división cultural ya existente de la sociedad en ocupacionalmente integrados y ocupacionalmente no integrados, ya sea a través del empleo precario y que cambia con frecuencia o a través del desempleo. La introducción de una renta básica incondicional equivaldría a una capitulación. En lugar de una estrategia de integración e inclusión de la sociedad trabajadora, se provocaría su retirada final e irreversible. La audaz justificación de la tesis “Por qué los surfistas deberían ser alimentados”, del teórico más importante de una renta básica incondicional, el economista y filósofo belga Philippe Van Parijs (1991), debe tomarse muy en serio. Mientras que, hasta ahora, tal forma de existencia solo ha sido una opción para la mayoría de las personas durante las fases de transición de su biografía, se convertirá en un fenómeno generalizado en el futuro, cuya dimensión cuantitativa dependerá del monto de la renta básica incondicional. La prolongación ya observada de la adolescencia en el sentido de formas de existencia dependientes en las sociedades occidentales recibiría un nuevo impulso y la entrada en la vida laboral no se aceleraría como ahora, sino que se retrasaría, para muchos, tal vez, hasta el punto de abandonarla voluntariamente.

Es probable que la visión apocalíptica que ha acompañado a todas las innovaciones tecnológicas disruptivas también carezca de fundamento en nuestros tiempos de digitalización. Todos los trastornos comparables han llevado a la pérdida de puestos de trabajo en ciertas industrias y profesiones; basta pensar en el éxodo rural en el siglo XIX, desencadenado por el uso de máquinas en la producción agrícola, lo que al mismo tiempo permitió un enorme aumento de puestos de trabajo en otros lugares. En el caso de la digitalización, hay muchos que sugieren que las economías del futuro utilizarán los recursos liberados por la racionalización para satisfacer intereses

culturales, por ejemplo. Ya hoy, la producción de juegos digitales y realidades virtuales es una de las industrias de más rápido crecimiento. El sector cultural, en su conjunto, es un motor de crecimiento en todas las economías occidentales. El aumento de la digitalización no necesita ni debe dar como resultado una salida de la sociedad trabajadora. Tal desarrollo es económicamente indocumentado y culturalmente indeseable.

Al final de *Wall-E*, la nave espacial con los últimos humanos que quedan a bordo se dirige a la Tierra. A medida que la vida ha vuelto a ser posible en la tierra, las personas pueden ahora regresar a su antiguo planeta después de siglos en un espacio estéril. Cuando las escotillas se abren, los humanos gordos, que mientras tanto han olvidado cómo caminar, salen gateando. Con los ojos muy abiertos, miran los enormes esqueletos de rascacielos destruidos frente a ellos. El aire es polvoriento y seco. Saben que hay mucho trabajo por delante para hacer que la Tierra vuelva a ser hermosa y habitable. Y el capitán sonríe. La supuesta tierra de leche y miel ha llegado a su fin y la vida vuelve a tener sentido.

Comentario del traductor

La metáfora de la felicidad como aspecto socioeconómico de la digitalización

La película *Wall-E*, como producto, como texto, como enunciado, ¿qué nos dice? Las posibles respuestas varían, no solo desde las coordenadas de interpretación, sino también desde el alcance que se le quiera dar. Quiero hacer la entrada con la siguiente pregunta: ¿es el sistema ironizando sobre sí mismo?

La confianza de la relación con la tecnología tiene varias dimensiones. En el caso del capítulo aquí referenciado, se habla de sociedad y economía, en la tríada, salario, empleo y formación, y las implicaciones socioeconómicas de un proceso de digitalización. Uno de los frentes de desarrollo tecnológico con mayor inversión es el de la domótica y su vínculo con lo cotidiano —en el sentido de que la *casa* es el lugar donde habitamos— y sus implicaciones sociales.

Esto se asocia con el “internet de las cosas”, al servicio del hogar. Así, las denominadas *casas inteligentes* funcionan como red de sistemas, por ejemplo, con neveras, hornos y estufas que están en casa y sistemas de distribución de alimentos que están en supermercados. Entonces, planear una cena puede hacerse enviando un mensaje con instrucciones básicas al ordenador de su casa; este sistema se encarga de todo, sin intervención humana: selección de ingredientes (en caso de que no los haya en casa, los pide al supermercado y estos son traídos a casa por drones), desinfección y preparación. Con la integración de la inteligencia artificial, puede calcularse en cuánto tiempo llegarán los comensales o, incluso, ordenar menús individuales.

Según lo expuesto, el impacto socioeconómico, respecto a la cadena de servicios, puede verse afectado, pues ya no se requiere ir al mercado; con lo que todos los trabajos asociados y actividades hechas por personas podrían ser reemplazables: conductores por drones, cocineros por hornos automáticos; no se requieren personas que atiendan las registradoras, empacadores... y así, podemos hacer proyecciones frente al reemplazo de los puestos de trabajo de personas por máquinas. En el texto, los autores interpelan el concepto de *revolución*, en términos de producción y productividad, y de las implicaciones de una eventual digitalización, en efecto, estamos ingresando a un mundo sociotécnico, que genera rupturas sobre el sentido de la producción y el significado de ser productivo.

Al respecto, expongo dos puntos:

- El costo económico de la domótica, que requiere un ecosistema digital robusto: conexión a internet y artefactos de punta. ¿Quiénes pueden acceder a esos servicios?, este aspecto puede generar que se radicalicen aún más las diferencias sociales.
- Las consecuencias individuales de relacionarse, por ejemplo, una familia que “comparte”, pero cada uno se mantiene conectado a su dispositivo móvil: están juntos sin estar realmente compartiendo, he ahí una de las metáforas de la felicidad.

El capítulo plantea rupturas interesantes: la práctica común del trabajo remunerado, con la idea de que para obtener una recompensa habría que “hacer” o “producir” algo, en la dinámica causa/efecto y en la lógica esfuerzo/resultado. Así como se muestra en la película, en la sociedad actual impera un discurso basado en una inobjetable superioridad pragmática, en la que se promueven y sobrevaloran los procesos de calidad y de experiencia de cliente. Somos clientes y nuestra voz tiene sentido debido a la conexión con los productos y la satisfacción. El asunto aquí es que la reducción de la experiencia humana, en términos de consumo, en una mercancía renovable, descartable, reitera la pregunta por cuál será el precio social que se pague por estar en una dimensión socioeconómica digital.

Vamos dando saltos emocionales como en las pantallas de la película *Wall-E* en las que se indica que el azul es el nuevo rojo, mientras que el proceso de digitalización de la sociedad avanza sin tregua.

19. Tentaciones transhumanistas

Traducido por Claudia Rozo Sandoval

Cuando el policía Alex se despierta del coma artificial y se encuentra en su nuevo cuerpo, una especie de traje de Batman de acero negro, entra en pánico —como es comprensible—.

—¿Qué diablos has hecho conmigo? —pregunta, horrorizado.

Alex, quien fue víctima de un coche bomba y perdió la mayor parte de su cuerpo, excepto el torso, fue operado por el doctor Norton, un experto en prótesis cibernéticas, para incorporarle un traje robótico. Si bien este traje le da movilidad y fuerza a Alex, también lo convierte en una especie de robot a nivel psicológico. Para que las partes técnicas, como brazos y manos artificiales, funcionen bien, los impulsos eléctricos del cerebro deben estar perfectamente implementados, es decir, Alex debe estar emocionalmente estable.

Alex desconoce que la megacorporación OmniCorp, que patrocinó su costosa operación, tiene un motivo oculto: considerando que la idea de los robots armados hasta ahora se ha enfrentado a la resistencia de la población y del gobierno, debido a una ley común promulgada por el senador Dreyfus,¹ el CEO de OmniCorp, Raymond Sellars, espera presentar al público con Alex, un cibernético tan simpático y eficiente que la gente querrá más de estos RoboCops en el futuro. Eso sí, para luego poder distribuirlos con el mayor beneficio posible.

1 Lleva el nombre del filósofo de la vida real Hubert L. Dreyfus, uno de los críticos más destacados de la IA en EE. UU. (Dreyfus, 1972, 1992).

Cuando Alex se observa y se siente en su traje de robot por primera vez, se derrumba. Él no quiere vivir así. Solo cuando el doctor Norton le dice que su esposa estuvo de acuerdo en su transformación, se calma y decide acostumbrarse a su nueva existencia.

Sin embargo, hay otro problema: Alex es mucho menos capaz de combatir que los robots reales. Los robots disparan de inmediato; antes de disparar, Alex considera si es moralmente correcto hacerlo o no; no obstante, al hacerlo, pierde tiempo. Para superar esto, que desde la perspectiva de OmniCorp es una desventaja, el doctor Norton equipa a Alex con una visera virtual. En situaciones de combate, esta visera se apaga automáticamente, haciendo que la parte robótica tome el control de su cuerpo y sus acciones. Lo desleal de esto es que un chip implantado en el cerebro de Alex por el doctor Norton le hace creer que estas acciones obedecen a sus propias decisiones. Alex se vuelve cada vez más frío y sin emociones, a medida que avanza la película. Ni su esposa ni su hijo lo reconocen. Donde comienza el robot —nos dice *RoboCop* (José Padilha, EE. UU., 2014)— termina el humano. Otro problema que plantea la película es la cuestión de la responsabilidad. En *RoboCop*, el poder sobre los robots no reside en el Estado, sino en manos de una empresa. Y esta empresa tiene un solo objetivo, a saber: aumentar sus ganancias. Alex está indefenso frente a las maquinaciones de OmniCorp, que decide cómo él debe ser programado o, para ser precisos, cómo deben programarse sus partes robóticas.

Los cibernéticos, seres híbridos entre el hombre y la máquina, existen en cierto sentido desde hace mucho tiempo. ¿No son también las gafas una ayuda artificial que los humanos han estado usando durante siglos para mejorar sus habilidades? Las piernas ortopédicas o los audífonos también pertenecen a esta categoría. Pero ¿qué pasa con los microchips implantados que permiten encender su automóvil, abrir la puerta de su casa o pasar sus datos de contacto? ¿O lentes de realidad aumentada que hacen posible ampliar la visión del entorno o superponer información de internet, directamente en nuestro campo de visión? ¿O chips implantados que permiten sentir los movimientos de los demás o escuchar los colores como

sonidos? ¿Nos acostumbraremos tanto a estas ayudas como a las gafas para la hipermetropía? En un sitio de internet llamado *I am Robot* (<https://chip-implants.com/>), la empresa del mismo nombre (con el lema “Renueva tu cuerpo con nuevas funciones”) ofrece el envío de estos chips por correo.

La Agencia de Proyectos de Investigación Avanzada de la Defensa de EE. UU., una agencia del Departamento de Defensa, durante años ha estado investigando hasta qué punto las descargas eléctricas en ciertas regiones del cerebro pueden aumentar la atención, suprimir la fatiga e influir en el comportamiento moral. El nombre de esta tecnología es *neuroenhancement* y está diseñada para mejorar el cerebro, en parte con la ayuda de chips digitales o para equiparlo con mejores capacidades de combate.²

Las fantasías están particularmente sobreestimadas en lo que se refiere a la interfaz cerebro-computadora, es decir, conexiones técnicas entre procesos cerebrales y sistemas de *software*. Por ejemplo, la Unión Europea también está financiando el proyecto VERE (Virtual Embodiment and Robotic Re-Embodiment) cuyo objetivo de investigación explícito es ligar permanentemente el sentido humano del yo con avatares o robots. Ya se han registrado ensayos exitosos en los cuales los pensamientos de un sujeto sobre el movimiento son leídos, a través de imágenes de resonancia magnética, y se transmiten a los robots, quienes después ejecutan ese movimiento.³

Los aspectos positivos de tal tecnología, incluyendo el “efecto Proteus” que la acompaña, esto es, la identificación exitosa del ego con otro cuerpo artificial, son innegables: las personas físicamente paralizadas podrían caminar y moverse en el mundo nuevamente gracias a un nuevo cuerpo o nuevas partes del cuerpo.

2 Por ejemplo, el conocido empresario y director ejecutivo de Tesla, Elon Musk, actualmente está desarrollando chips con su nueva empresa Neuralink, que se supone mejorará el rendimiento del cerebro y permitirá la transmisión del pensamiento.

3 Véase el estudio piloto israelí-francés de Cohen *et al.* (2012).

Pero ¿qué pasa con los lados negativos? ¿Qué pasa si, gracias a hábiles manipulaciones neurológicas, esta técnica se usa incorrectamente para hacer que las personas hagan ciertas cosas? ¿Qué pasa si el control de los impulsos funciona peor con esa técnica? ¿Y si, dentro de un despliegue militar, la identificación con un fuerte cuerpo de soldado robótico que actúa junto con otros en un grupo conduce a lo que se conoce como el *efecto Lucifer*,⁴ es decir que el usuario se sienta tentado a actuar de manera excesivamente agresiva y sádica, porque se encuentra en un nuevo papel?

Los defensores de estas nuevas tecnologías hechas para expandir las capacidades humanas se llaman a sí mismos transhumanistas. Aprueban el uso de las posibilidades tecnológicas más modernas para llegar a una dimensión completamente nueva de la cognición humana (sobre todo en lo relacionado con la capacidad de comprender procesos complejos) y de la práctica. El transhumanismo es un movimiento global. Hay varios *think tanks* transhumanistas globales, así como “partidos” transhumanistas en EE. UU., Australia, Corea, India, Gran Bretaña, Austria y Alemania.

Tan eufóricos como están los transhumanistas con la expansión de las capacidades humanas a través de las nuevas tecnologías, hay también movimientos en contra que temen una nueva división de la sociedad entre aquellos que, con ayuda de las nuevas tecnologías (no solo digitales, sino también médicas y farmacéuticas o nanotecnológicas), se sitúan en un nivel superior de desarrollo humano y aquellos que se les niega esto por falta de medios económicos o técnicos.

4 El *efecto Lucifer* proviene del “experimento de la prisión de Stanford”, realizado por Philip Zimbardo y algunos colegas, en 1971, en el que se les pidió a los estudiantes que asumieran el papel de guardias y prisioneros. Después de un tiempo corto, los estudiantes que habían asumido el papel de guardias trataron a los prisioneros de manera sádica y agresiva. La razón de esto fue tanto el hecho de que a los estudiantes se les permitió asumir roles poderosos como sentir la presión social del grupo (Zimbardo, 2007).

El transhumanismo contemporáneo es la concreción de un antiguo sueño humano, a saber: poder desafiar la condición humana, ir más allá de todas las limitaciones de su naturaleza y desarrollar poderes y habilidades sobrehumanas.

Algunos transhumanistas también esperan cumplir el deseo infantil de la inmortalidad a través de tecnologías digitales. Por ejemplo, Alcor Life Extension Foundation, una organización estadounidense sin fines de lucro ofrece preservar el cerebro después de la muerte, para que el dueño de este cerebro pueda resucitar décadas, tal vez incluso siglos después, en una forma clonada, saludable, joven o incluso en un cuerpo de máquina artificial. El cuerpo de la máquina que se fusionaría con el cerebro se convertiría así en una especie de doble o resucitado que, a diferencia del humano anterior, se supone, vivirá para siempre.

Desde un punto de vista psicoanalítico, tal deseo debe caracterizarse como regresivo y narcisista, porque una de las condiciones más elementales de la edad adulta es aceptar la propia limitación humana, así como la propia muerte. Freud también describió la idea de los *doppelgangers*, es decir, una idea doble: patológica y narcisista, y mencionaba, en ese contexto, la antigua tradición egipcia de los sarcófagos, los cuales como imágenes de la muerte se suponía que garantizaban la inmortalidad. Los robots que se van a implantar en nuestros cerebros serían, en este sentido, nada más que una manifestación moderna de estas ideas regresivas.

Lo que aún no es posible en la realidad lo desarrollan las películas de ciencia ficción, probablemente de forma más extrema en *Ghost in the shell* (Rupert Sanders, EE. UU., 2017), basada en un manga de Masamune Shirow. Mayor Mira Killian, la protagonista de la película, es el resultado de la fusión de un cerebro humano y un cuerpo puramente sintético que le otorga una increíble elasticidad y fuerza. Tony Stark, el héroe de *Iron Man* (Jon Favreau, EE. UU., 2008), que también tiene poderes sobrehumanos, se ha convertido en un cibernético con la ayuda de un traje de batalla de acero. Sin embargo, a diferencia de Alex, en *RoboCop*, Tony, como un ingeniero genio, no solo tiene el poder sobre su propia programación, sino que, como

propietario inmensamente rico de una empresa gigantesca (Stark Industries), puede patrocinar a sí mismo. Aquí no se plantea el problema de que una empresa ejerza control sobre la identidad de una persona.

Freud sin duda habría visto esta película como una expresión típica de fantasías regresivas e inconscientes de omnipotencia e invulnerabilidad.⁵ Y, sin embargo, estas fantasías sin duda han alimentado a menudo innovaciones, como el deseo de volar, por ejemplo, la realización de este sueño milenario de la humanidad, que al final se convirtió en una realidad técnica y económica. El transporte individual basado en automóviles, el movimiento de decenas de caballos de fuerza con una mano ligera y una suave presión del pie, el movimiento de un vehículo que pesa toneladas con un uso mínimo de la propia fuerza física es también la realización de un viejo sueño de movilidad ilimitada. Las ciencias naturales y la tecnología han recibido repetidamente importantes impulsos de ideas radicales para nuestro mundo vital. El desarrollo de la tecnología de la energía nuclear y su uso pacífico en la forma de reactores de agua ligera, o la tecnología mucho más avanzada de los reactores reproductores rápidos tenían como objetivo hacer realidad la visión de un uso ilimitado, limpio, sostenible y sin preocupaciones de la energía. Como sabemos hoy, al final las cosas resultaron bastante diferentes, y la energía nuclear como fuente de energía todavía se considera en muchos países, en el mejor de los casos, como una tecnología de transición hacia una economía descentralizada basada en energías renovables, que, curiosamente, cae en modelos bastante antiguos de producción de energía (turbinas eólicas, hidroeléctricas, estiércol, energía geotérmica, etcétera).

Las nuevas tecnologías casi siempre han estado acompañadas de visiones utópicas, ya sea en la invención de las aeronaves, la

5 Para el científico cultural Klaus Theweleit (1987), *Iron Man* sería probablemente un ejemplo de la fantasía de un duro "cuerpo de soldado" de un hombre que es incapaz de entablar relaciones con otras personas, debido a su trastorno del ego.

electricidad o el primer automóvil que fue aclamado por Henry Ford en su libro *Machinery: a new Messiah*, como un “nuevo mesías” que traería la paz eterna y la prosperidad en la tierra. Sin embargo, incluso si estas expectativas utópicas nunca se han cumplido, a menudo han mejorado las condiciones para la vida humana. Por eso, los humanistas dan la bienvenida a las nuevas tecnologías. Los humanistas se guían por la idea de *agencia responsable* y la idea de *autodeterminación*. Tienen como objetivo mejorar las condiciones de autoría humana dentro de los límites de la condición humana.

Giovanni Pico della Mirandola, en su pequeño pero muy influyente escrito *Oratio de hominis dignitate*, elogió casi como un himno las especiales habilidades humanas en las artes y las ciencias. Para él, la dignidad humana especial se manifiesta, sobre todo, en el hecho de que el hombre —a imagen de Dios, por así decirlo— alberga en sí mismo capacidades divinas de creatividad y libertad, que se desarrollan plenamente en condiciones favorables. Sin embargo, esta idea humanista del autodesarrollo humano continúa limitada por las condiciones naturales. El hilo conductor del pensamiento humanista desde la antigüedad, mantener la moderación y defender el medio frente a los extremos (la doctrina mesotes de Aristóteles en *Ética a Nicómaco*), cobra relevancia a través de las nuevas posibilidades tecnológicas y el movimiento transhumanista.

El humanismo puede convertirse en transhumanismo solo en la era moderna y en los tiempos de las avanzadas tecnologías de la información que, en articulación con la neurociencia contemporánea, generan expectativas de superación de algunas de las barreras naturales. El transhumanismo se diferencia del humanismo en que cuestiona las condiciones naturales, incluso si pertenecen a la autoimagen tradicional humana, en cierto sentido, extendiendo la idea humanista de autodeterminación y autodiseño más allá de todos los límites. Este rápido paso del humanismo al transhumanismo tiene un famoso precursor en Friedrich Nietzsche y su ideal transhumanista del superhombre (*Übermensch*), quien no tiene que tomar en consideración a nada ni a nadie, y puede ignorar la herencia de la supuesta mediocridad y los valores de la humanidad.

En el transcurso de la película *RoboCop*, el protagonista Alex pierde cada vez más su moderación y su centro. Se vuelve arrogante, agresivo y progresivamente abandona su humanidad. Al final, cuando el gobierno de EE. UU. se pronuncia en contra del uso de policías cibernéticos, el doctor Norton confiesa en una entrevista que lamenta haber trabajado en la creación de un cibernético armado: “Admito que sabemos menos de lo que pensábamos. Lo sé. Creo que mi programa de investigación es sólido, pero lo que hice con él fue errado”.

Comentario de la traductora

En el ya clásico *Manifiesto cibernético*, Haraway (1991) interpelaba al feminismo esencialista y los dualismos desde los cuales se explicaban los cuerpos, las herramientas y lo femenino; también cuestionaba las demarcaciones entre humanos y máquinas, y advertía la “difícil tarea de reconstruir los límites de la vida diaria en conexión parcial con otros, en comunicación con todas nuestras partes”. Al mismo tiempo, indicaba esa doble condición de las contribuciones de la ciencia y la tecnología, de una parte, con los aportes a la “satisfacción humana” y de otra, inmersa en “una matriz compleja de dominaciones”. En ese entramado de relaciones, desafíos y demandas feministas por salir de las demarcaciones, trascender las fronteras, los límites y ocupar los espacios, encontraba Haraway en los cuerpos simbióticos entre natural y artificial y la “imaginería” cibernética la oportunidad de construir y deconstruir máquinas, identidades y categorías como parte de la necesaria y pertinente acción política que demanda habitar nuevos cuerpos.

En esa línea, la actual condición de humanidad que estamos construyendo —en la que confluyen cuerpos naturales con “prótesis artificiales”, formas de “inteligencias” con adecuaciones y adaptaciones tecnológicas y biológicas, desde las que se proyecta el sueño eterno de alcanzar la inmortalidad y la manipulación de la vida natural con la aspiración de la juventud eterna—, se pone de relieve la urgente necesidad de llevar al escenario de lo público, del

debate, de lo político y lo ético el sentido de “lo humano”. Este se discute y se transforma desde el fuero privado de las corporaciones que invierten en desarrollos científico-técnicos, sin posibilidad de intermediación de los Estados (bien por desconocimiento de las complejidades inmersas en los propios desarrollos, o por la influencia de los tentáculos del capital en las instancias de regulación), ni antes de control, y con pocas o ningunas intervenciones por parte de la sociedad.

Si bien en la discusión planteada por Haraway, como en los feminismos contemporáneos, no se avizora el debate, ni las implicaciones de los transhumanistas, el *Manifiesto* pone en el centro de la controversia la importancia y necesidad de hacer consciencia sobre la condición de los cuerpos que emergen en la fusión entre ficción y realidad.

Como señalan Nida-Rümelin y Weidenfeld, el movimiento transhumanista aboga por el mejoramiento de las capacidades humanas, la proyección del superhombre, de seres animados con habilidades cognitivas y físicas superdotadas (al estilo de las mejores películas de ciencia ficción) a través de una especie de “evolución diseñada” (Diéguez, 2017) basada en los desarrollos de las biotecnologías, la robótica y la inteligencia artificial. Frente a ello, la toma de conciencia que demanda Haraway implica advertir la potencia implícita en esta condición que emerge “transhumana” y que, como parte de la acción política, reclama reconocer otras identidades, así como, comprender las implicaciones, alcances y límites de esa “evolución diseñada”.

Retomamos del manifiesto la relevancia del sentido de consciencia de la actual condición de humanidad que, en la práctica de la evolución referida, parece que va configurando formas de ser humano de las que el grueso de la sociedad no tiene conocimiento ni de las maneras como se produce, sus riesgos y posibilidades. Este nivel de consciencia demanda también estilos de actuación y participación, a propósito de los “límites” entre lo natural y artificial que la tendencia transhumanista proyecta, ausente de consideraciones éticas, democráticas y sociales.

Los alcances del transhumanismo exigen la configuración de nuevas inteligibilidades, es decir, posibilidades de conocer, interpretar y develar los movimientos que constituyen estas nuevas relaciones de fuerza (políticas, económicas, sociales, éticas), desde las que se configuran formas de vida más allá de lo humano y de lo vivo, para referir también otras manipulaciones genéticas y cibernéticas con las cuales se maniobran especies animales y vegetales.

En este sentido, consideramos relevante volver, en palabras de Olivé (2003), a reiterar la idea de un nuevo contrato social; esto es, llevar a la esfera pública y auténticamente democrática los desarrollos científicos y tecnológicos que en la esfera privada del mercado inciden en nuestros modos de vida, para disminuir el impacto de lo que el doctor Norton nombraba (después de “engendrar” a RoboCop y observar los alcances de su acción): “creo que mi programa de investigación fue sólido, pero lo que hice con él fue errado”.

Referencias del comentario de la traductora

- Haraway, D. (2020). *Manifiesto cyborg. Ciencia, tecnología y feminismo socialista a finales del siglo xx* (pp. 149-181). (Trad. M. Tallen). <https://mundoperformance.net/2020/08/25/manifiesto-cyborg-ciencia-tecnologia-y-feminismo-socialista-a-finales-del-s-xx/>
- Diéguez, A. (2017). *Transhumanismo, la búsqueda tecnológica del mejoramiento humano*. Herder Editorial.
- Olivé, L. (2003). Por un nuevo contrato social sobre la ciencia y la tecnología. *Ciencia y Desarrollo*, 29(172), 7-12.

20. Sobre la metafísica de la digitalización

Traducido por Claudia Rozo Sandoval

Los Ángeles en 2019. Una ciudad lúgubre, donde llueve constantemente y los hornos arrojan fuego al aire a intervalos regulares. La ciudad se ha vuelto tan multicultural que las personas apenas pueden comunicarse entre sí. El mundo distópico que imaginó Ridley Scott en *Blade Runner*, en 1982, acompañado de los melancólicos sonidos del sintetizador de Vangelis, es todo menos tentador. El único edificio brillante de la ciudad, una enorme estructura con forma de pirámide de oro brillante, pertenece a un hombre llamado Tyrell. El hombre pequeño con gafas enormes es el inventor y diseñador de robots que se envían a Marte para hacer habitable el planeta en las condiciones más duras. Estos robots, llamados *replicantes*, son difíciles de diferenciar de los humanos en sus comportamientos y apariencia. Solo cuando uno estudia de cerca sus reacciones emocionales, puede lograrlo.

Después de que cuatro de los replicantes abandonaron ilegalmente Marte y se dirigieron a la Tierra, se le pidió a Rick Deckard, oficial de policía, que los encuentre y los elimine. Deckard es un buen cazador de replicantes, por lo que no tarda en eliminarlos a casi todos. Solo queda uno hasta el final: Roy. De todos los replicantes, Roy no solo es el más inteligente y fuerte, sino también el que experimenta un mayor desarrollo en el transcurso de la película. De una especie de robot primitivo que habla con voz entrecortada y no siente empatía por los humanos —a quienes ocasionalmente

mata—, pasa por etapas de desarrollo semejantes a las de un ser humano. Al principio es impulsivo y taciturno como un niño, luego agresivo y buscador de sentido como un adolescente, que luego se convierte en un superhombre (*Übermensch*) nietzscheano, hasta que en el momento de su muerte se vuelve compasivo y espiritual.

Si uno adopta la perspectiva emergentista¹ según la cual el siguiente nivel superior no está determinado por el nivel inferior, no se puede descartar la posibilidad de que los sistemas controlados por *software* algún día tengan estados mentales, de hecho, la capacidad de comprensión. No hay argumentos basados en principios de que solo la materialidad biológica y no física permite sentimientos, creencias, intenciones, decisiones, etcétera. La experiencia con la biología nos enseña que tales transiciones suelen ser fluidas, graduales: el recién nacido probablemente no tiene creencias ni persigue intenciones, pero unos meses más tarde no puede haber dudas sobre ninguno de los dos. El peligro de lidiar con la Inteligencia Artificial es confundir la simulación con la realidad. Esto nos lleva con bastante frecuencia a utilizar un vocabulario inapropiado, según el cual los sistemas de *software* tienen “percepciones” y “toman decisiones”. Por ejemplo, cuando el sistema de navegación ordena “Gire a la derecha”, le atribuimos una intención al sistema, a saber: hacernos girar a la derecha. Pero tal como están las cosas, es decir, en el estado actual de las tecnologías digitales, esto sería

1 El emergentismo se entiende de diferentes maneras, pero tiene dos principios básicos en común: (1) las propiedades y las leyes del nivel superior (aquí: biológico) no son reducibles a las del nivel inferior (aquí: físico), es decir, aún no está físicamente determinado qué organismos biológicos surgen y qué propiedades tienen; (2) las propiedades y leyes del nivel superior (aquí: biológico) son compatibles con las propiedades y leyes del nivel inferior (aquí: físico). Los organismos biológicos constan de partes físicas, y todos los eventos y procesos de un organismo físicamente describable pueden explicarse mediante leyes físicas. Las leyes biológicas no violan las físicas. De manera análoga, podemos imaginar la relación entre la neurociencia y la psicología, o entre la psicología y la lógica. A menudo se añade una tercera propiedad del emergentismo, según la cual algo que es igual en cuanto a las propiedades del nivel inferior no difiere con respecto a las propiedades del nivel superior.

una mistificación, incluso si la película *Blade Runner* (Ridley Scott, EE. UU., 1982) parece sugerir que existe un desarrollo tan emergente de una IA que simplemente simula las facultades humanas a una IA que realiza estas facultades.

Los sistemas de *software* no sienten, piensan ni deciden, al contrario de lo que hacen los humanos, ya que no están determinados por procesos mecánicos. Gracias a su capacidad de comprensión, así como a su capacidad de tener sentimientos, los humanos pueden determinar por sí mismos sus acciones, y lo hacen decidiendo actuar de una manera y no de otra. Los seres humanos tienen razones para lo que hacen, como seres racionales son capaces de reconocer verdades matemáticas y lógicas, pueden distinguir las buenas de las malas razones. Al participar en el razonamiento teórico y práctico, influimos en nuestros estados mentales, nuestro pensamiento, sentimiento y acción, ejerciendo así un efecto causal en el mundo biológico y físico. Si el mundo se entendiera de manera reduccionista, todos los fenómenos superiores, desde la biología hasta la psicología, la lógica y la ética, estarían determinados por leyes físicas: las decisiones y creencias humanas serían causalmente irrelevantes en ese mundo.²

2 Una teoría T_2 puede reducirse a una teoría T_1 si T_2 puede derivarse completamente de T_1 , lo que presupone que los términos de T_2 también pueden definirse con la ayuda de los términos de T_1 . Existe una forma más débil de reductibilidad si todas las predicciones empíricas de T_2 pueden derivarse de T_1 (reducción empírica). El fisicalismo es la forma más destacada de reduccionismo, según la cual toda la ciencia se remonta a la física. Hasta ahora, esto solo ha tenido éxito en algunas partes de la química inorgánica y, por lo demás, sigue siendo ciencia ficción. Incluso, la reductibilidad de la biología a la física es altamente improbable; la reductibilidad de las ciencias sociales incluso de los estudios literarios a la física está completamente fuera de discusión. Esto se debe, entre otras cosas, al hecho de que incluso en las ciencias sociales, pero especialmente en los estudios culturales y las humanidades, términos como *significado*, *intención*, *creencia* o *emoción* no pueden ser traducidos a términos físicos: las intenciones o incluso las razones no son un objeto posible de la física.

En uno de los momentos más hermosos en la historia de la película, en la última escena dramática de *Blade Runner*, hay un enfrentamiento entre Deckard y Roy. Roy, que está gravemente herido y al borde de la muerte, sigue obsesionado con matar a Deckard. Como un animal salvaje, persigue a Deckard a través de un rascacielos en ruinas, la parte superior de su cuerpo desnudo está cubierta con pintura blanca. Finalmente, ambos llegan al techo del edificio. Cuando Deckard intenta saltar al otro techo se resbala. Con lo último de sus fuerzas, es capaz de aferrarse a una barra de hierro. Sabe que, si se suelta, caerá y morirá. En ese momento, Roy aparece sobre él, semidesnudo, sangrando y confiado en la victoria, un *Übermensch* nietzscheano. Roy mira a Deckard, lo ve luchando por su vida. El espectador espera que patee a Deckard desde el techo, pero en lugar de eso sucede algo más: Roy mira a Deckard a los ojos y extiende su mano hacia Deckard para ayudarlo a regresar.

Roy se sienta frente a su antiguo enemigo. Sabe que está a punto de morir. La lluvia gotea por su rostro.

—He visto cosas —dice Roy—. Ustedes no lo creerían. Atacar naves en llamas desde el hombro de Orión. Observé rayos C brillar en la oscuridad cerca de la Puerta Tannhäuser. Todos estos momentos se perderán en el tiempo, como lágrimas en la lluvia.

Y tras una breve pausa: “Hora de morir”. Luego baja la cabeza. Está muerto.

En este momento, vemos una paloma volar desde el techo hacia el cielo. El pájaro, que se interpreta fácilmente como un símbolo del alma de Roy, deja en claro lo que Ridley Scott quiere decirle aquí al espectador: los replicantes pueden tener, si cuentan con suficiente tiempo, experiencias y recuerdos, convertirse en seres sensibles, empáticos y espirituales. Sin embargo, debemos tener cuidado de no entender esto como una profecía realista de cómo evolucionará la Inteligencia Artificial, sino como una metáfora del poder de transformación de los humanos que pueden expandir sus capacidades para ganar sentimientos morales como el perdón y la empatía.

Comentario de la traductora

La idea de una *metafísica de la digitalización* evoca la relación contemporánea del ser humano con la tecnología, concretamente las simbiosis que mantenemos con los objetos técnicos con los que estructuramos nuestros modos de vida hoy y que, de manera casi imperceptible, se funden en nuestras rutinas cotidianas, en procesos que revelan niveles de dependencia y adecuación mutua. Un ejemplo de ello es la subordinación constante al celular que “administra” nuestra memoria, controla nuestros tiempos, nuestras formas de transitar por la ciudad, nuestras agendas, entre otros.

En la ciencia ficción, la duda sobre la respuesta “empática” de Roy ante el peligro de muerte de su enemigo descubre el temor de la humanidad ante el dominio de la máquina; en esta ocasión, la manifestación de sentimientos y emociones que “salvan” al humano y permiten intuir la transferencia de una condición de humanidad al robot, en una especie de transposición metafísica, que termina con la idea de *dominación* de las máquinas sobre los humanos. Situación posible, por el sentimiento que emerge entre robot y humano; escena que los autores analizan desde las posibilidades de la Inteligencia Artificial y las respuestas que esta produce como efecto de simulación, de réplica automática, “aprendida” (simulada) y no como expresión de un sentimiento.

En esta línea, y siguiendo la advertencia de los autores sobre el riesgo de trastocar la realidad con la simulación, en el cada vez más creciente universo de *Inteligencias Artificiales*, se percibe un aire de “superhumanidad” incorporado en los replicantes, quienes al parecer solo carecen de emociones, las mismas que dificultan y simultáneamente permiten el estar juntos.

Como señala Yuk Hui (2017), la IA exhibe su capacidad de autorregulación, donde la simulación del lenguaje (simulación por cuanto, por ahora, para la máquina las construcciones carecen de sentido) transita entre capacidades y habilidades naturalmente humanas y biológicas hacia campos cibernéticos, generando una especie de “juego

de falsedades” en el que la posibilidad y capacidad de la máquina se muestra como el deseo a alcanzar, por parte de la humanidad.

Procesar, almacenar, codificar, autorregularse se proyectan como el horizonte a ser develado y aprendido para quienes, al parecer, mantienen algunos trazos difíciles de traducir al nuevo orden de máquinas autorreguladas: las emociones, los sentidos y significados que se construyen en ese “estar juntos” y que para los distintos organismos humanos y biológicos también se revisten de sentidos y significados concretos.

Es alrededor de estas reflexiones sobre lo real, lo simulado, lo artificial, las formas de nombrarlo y los modos de habitar el mundo hoy, que quisiéramos hacer alusión a la relación que mantenemos con nuestros objetos técnicos. Para Simondon (2003) y para Yuk Hui (2017), los objetos digitales, los procesos de individuación del ser, que dan cuenta de su naturaleza, se despliegan en los planos físico, biológico y psíquico, puesto que la condición de indeterminación del objeto se constituye, en cuanto máquina abierta sensible a la información del exterior, en relaciones de interpretación del ser viviente con los seres maquínicos. Esto revela la condición de humanidad que reside en la máquina y la condición técnica propia de los seres humanos, cada vez más tenue, invadida y silenciada por el consumismo y la aceleración que dificultan el pensar sobre los seres, la naturaleza, los objetos, las cosas y las formas de relación. En este sentido, a propósito de los replicantes (como metáfora de la simbiosis entre humanos y máquinas), consideramos valioso volver sobre la idea de alienación tecnológica, producida no por la técnica, las tecnologías y los objetos técnicos, sino por la carencia de pensar acerca de ellos.

Referencias del comentario de la traductora

- Hui, Y. (2017). ¿Qué es un objeto digital? *Virtualis*, 8(15), 81-96.
Simondon, G. (2003). *El modo de existencia de los objetos técnicos*. Prometeo Libros.

21. Epílogo

Traducido por Rocío Rueda Ortiz

Hay un hilo conductor que recorre este libro. Es la crítica de lo que llamamos la *ideología de Silicon Valley*, o dicho de forma positiva, su sustitución por la idea de un humanismo digital que se desmarca de una interpretación equivocada de la Inteligencia Artificial.

En el sentido más amplio, la Inteligencia Artificial se refiere a todo lo que se puede realizar con técnicas digitales, como la informática, desde calculadoras de bolsillo hasta los sistemas de *software* que pueden aprender y actuar de forma autónoma. La interpretación más errónea se denomina en filosofía *IA fuerte (strong AI)*, según la cual no hay diferencia categórica entre humanos y ordenadores; así, los sistemas de *software* que imitan el comportamiento, el juicio y la toma de decisiones humanos también tienen características humanas. Hay dos lecturas posibles de la IA fuerte: una materialista y otra animista.

En la lectura materialista, los cerebros humanos no son más que computadores complejos. Por tanto, el lenguaje de las propiedades mentales es fundamentalmente redundante y, con el progreso de la ciencia natural, el *mentaleeze* (el lenguaje de las propiedades mentales) se extinguirá.

Si uno se tomara en serio el materialismo digital como visión del mundo, esto significaría el fin del modo de vida humano tal y como es.

En la versión animista que puede encontrarse en muchas películas de Hollywood, los sistemas de *software* se consideran seres animados dotados de propiedades mentales como los humanos.

La variante débil, IA débil, no afirma que no exista diferencia decisiva entre humanos y ordenadores, sino que todas las capacidades cognitivas las tienen, en principio, los computadores también.

La expectativa optimista de que las capacidades cognitivas de los sistemas de *software* son capaces de un desarrollo ilimitado suele ir acompañada de una especie de esperanza de salvación, según la cual las tecnologías digitales liberarán a las personas de su existencia, crearán nuevas formas de interacción y comunicación, nos proporcionarán compañeros (digitales) perfectos, y mejorarán las capacidades perceptivas y cognitivas humanas. El mensaje de que la transformación digital anunciará un mundo valiente y radicalmente nuevo, como predicaban muchos empresarios de Silicon Valley, no es una casualidad frente a la retórica milenarista de la salvación que siempre ha tenido un gran significado en la cultura estadounidense. Como una especie de “milenarismo tecnológico”, Silicon Valley pervierte la escatología cristiana y presenta la revolución digital como la respuesta a todos nuestros problemas económicos, sociales e incluso espirituales.

El humanismo digital se opone a esta ideologización de las tecnologías digitales con una actitud de sobriedad. Como todas las tecnologías del pasado, las digitales son ambivalentes. La transformación digital no humanizará automáticamente nuestras condiciones de vida, esto depende de cómo utilicemos y desarrollemos esta tecnología. El humanismo digital aboga por una actitud instrumental ante la digitalización: ¿qué puede ser beneficioso desde el punto de vista económico, social y cultural?; ¿dónde acechan los peligros potenciales?

Adicionalmente, el humanismo digital se opone a la IA fuerte. No hay nada que sugiera que los sistemas de *software* tengan percepciones o incluso emociones, que puedan reconocer y tomar decisiones. Lo que se observa aquí es una simulación más o menos exitosa de procesos cognitivos y emocionales. Debemos tener cuidado frente a un tipo concreto de autoengaño que consiste en desarrollar, primero, máquinas digitales que simulen emociones, cogniciones y emociones, y luego explicar este comportamiento atribuyéndoles estados mentales.

En este punto, entra en juego un argumento que se adentra en la lógica, las matemáticas y la epistemología. Para ello son cruciales los resultados de la “metamatemática de la incompletitud e indecidibilidad” de Kurt Gödel y otros lógicos de principios del siglo xx. Hemos interpretado esto como una refutación concluyente de la tesis de la IA débil. Con estos resultados, se ha demostrado que no puede haber una simulación completa del juicio y la toma de decisiones humanas. Hay muchos indicios de que la diferencia categórica entre los humanos y las máquinas está relacionada con esto. Este es al menos el supuesto desarrollado por el matemático y físico teórico Roger Penrose en dos extensas monografías (*The emperor's new mind*, 1989, y *Shadows of the mind: a search for the missing science of consciousness*, 1994). Además, hay que tener en cuenta que atribuir propiedades mentales a los sistemas de *software* haría problemático nuestro uso cotidiano de los ordenadores y bloquearía el desarrollo técnico de la digitalización. Si los ordenadores reconocen, deciden y sienten, tendríamos que tratarlos con consideración y también concederles derechos en función de su similitud con el ser humano. En contra de su intención, la IA fuerte se está convirtiendo en un obstáculo para el progreso tecnológico. El humanismo digital no se deja llevar por estas corrientes. Hace hincapié en la inmutabilidad de largo alcance de la naturaleza humana y en las condiciones del desarrollo humano. Defiende logros culturales como la separación de las esferas privada y pública y la autodeterminación informativa. Aboga por el fortalecimiento de la democracia, recurriendo también a las nuevas posibilidades digitales; advierte contra el declive de la conexión interpersonal en tiempos de creciente anonimato y manipulación de la comunicación por internet. Aboga por reforzar la capacidad de juicio para permitir una orientación fiable ante la sobreabundancia de datos.

El humanismo digital no es defensivo, no quiere frenar el progreso tecnológico en la era de la Inteligencia Artificial, sino que aboga por acelerar el progreso humano a partir de las posibilidades digitales para hacer nuestras vidas más ricas, eficientes y sostenibles. No sueña con una forma de existencia humana totalmente nueva, como

los transhumanistas; permanece escéptico frente a expectativas utópicas, pero es optimista respecto a la fuerza de la creatividad humana frente al potencial de la digitalización.

Referencias

- Arrow, K. (1951). *Social choice and individual values*. Wiley.
- Barthes, R. (1981). *Camera lucida: reflections on photography*. Hill and Wang.
- Bostrom, N. (2014). *Superintelligence: paths, dangers, strategies*. Oxford University Press.
- Cellan-Jones, R. (2014). Stephen Hawking warns artificial intelligence could end mankind. *BBC News*. <https://www.bbc.com/news/technology-30290540>
- Chalmers, D. (2010). *The character of consciousness*. Oxford University Press.
- Cohen, O. *et al.* (2012). fMRI Robotic Embodiment: a pilot study. En *4th IEEE RAS & EMBS International Conference on Biomedical Robotics and Biomechatronics (BioRob)* (pp. 314-319). Roma, Italia.
- Condorcet (marqués) (Jean-Antoine-Nicolas Caritat). (1785). *Essai sur l'application de l'analyse à la probabilité des décisions rendues à la pluralité des Voix*. L'imprimerie royale.
- De Lagasnerie, G. (2017). *The art of revolt. Snowden, Assange, Manning*. Stanford University Press.
- Dreyfus, H. L. (1972). *What computers can't do*. MIT Press.
- Dreyfus, H. L. (1992). *What computers still can't do. A critique of artificial reason*. MIT Press.
- Focus.de. (13 de mayo de 2015). *Hawking warnt: Dieser Macht könnten Menschen in 100 Jahren ausgeliefert sein* [En línea]. http://www.focus.de/wissen/technik/wird-man-siekontrollieren-koennen-stephen-hawking-warnt-in-100-jahren-sind-computer-intelligenter-alsmenschen_id_4681638.html
- Freud, S. (1913). *Totem and taboo*. Beacon Press.
- Frey, C. B. y Osborne, M. A. (2013). *The future of employment: how susceptible are jobs to computerisation?* https://www.oxfordmartin.ox.ac.uk/downloads/academic/The_Future_of_Employment.pdf. Accessed 6 March 2022.
- Gibbard, A. (1973). Manipulation of voting schemes: a general result. *Econometrica*, 41(4), 587-601.
- Gorz, A. (1999). *Reclaiming work: beyond the wage-based society*. Polity Press.
- Grice, P. (1991). *Studies in the ways of words*. Harvard University Press.

- Hevelke, A. y Nida-Rümelin, J. (2015). Responsibility for crashes of autonomous vehicles: an ethical analysis. *Science and Engineering Ethics*, 21(3), 619-630.
- Kleene, S. C. (1952). *Introduction to metamathematics*. Ishi Press.
- Kleene, S. C. (1985). *Golem XIV*. Harcourt.
- Kleene, S. C. (2013). *Summa technologiae*. University of Minnesota Press.
- Lem, S. (1976). *The star diaries*. Seabury Press.
- Marcuse, H. (1964). *One-dimensional man. Studies in the ideology of advanced industrial society*. Beacon Press.
- Margalit, A. (1996). *The decent society*. Harvard University Press.
- Maturana, H. (1988). Reality: the search for objectivity or the quest for a compelling argument. *The Irish Journal of Psychology*, 9(1), 25-82.
- Milkowski, M. (2013). *Explaining the computational mind*. MIT Press.
- Moravec, H. (1988). *Mind children: the future of robot and human intelligence*. Harvard University Press.
- Moravec, H. (1999). *Robot: mere machine to transcendent mind*. Oxford University Press.
- Nagel, T. (1974). What is it like to be a bat? *The Philosophical Review*, 83(4), 435-450.
- Nida-Rümelin, J. (2010). Reasons against naturalizing epistemic reasons: normativity, objectivity, non-computability. En A. Carsetti (ed.), *Causality, meaningful complexity and embodied cognition* (pp. 203-210). Springer.
- Nida-Rümelin, J. (2018). Responsibility in philosophy and law. *Proceedings World Congress Philosophy*.
- Nida-Rümelin, J. (2023). *A theory of practical reason*. Palgrave Macmillan.
- Penrose, R. (1989). *The emperor's new mind. Concerning computers, minds, and the laws of physics*. Oxford University Press.
- Penrose, R. (1994). *Shadows of the mind: a search for the missing science of consciousness*. Oxford University Press.
- Popper, K. (1996). Chapter 6. Of clouds and clocks: an approach to the problem of rationality and freedom of man. En *Objective knowledge: an evolutionary approach*. Washington University.
- Putnam, H. (1981). *Reason, truth and history*. Cambridge University Press.
- Rifkin, J. (1995). *The end of work: the decline of the global labor force and the dawn of the post-Market Era*. Putnam.
- Searle, J. R. (1980). Mind, brains, and programs. *Behavioral and Brain Sciences*, 3(3), 417-424.
- Searle, J. R. (1992). *The rediscovery of the mind*. MIT Press.
- Searle, J. R. (1993). The problem of consciousness. *Consciousness and Cognition*, 2(4), 310-319.
- Simanowski, R. (2018). *Facebook society. Losing ourselves in sharing ourselves*. Columbia University Press.

- Singer, W. (2001). Consciousness and the binding problem. *Annals of the New York Academy of Science*, 929(1), 123-146.
- Slotkin, R. (1973). *Regeneration through violence*. Wesleyan University Press.
- Smart, J. J. C. y Williams, B. (1973). *Utilitarianism: for and against*. Cambridge University Press.
- Theweleit, K. (1987). *Male fantasies*. University of Minnesota Press.
- Tomasello, M. (2008). *Origins of human communication*. MIT Press.
- Turing, A. (1938). *Systems of logic based on ordinals*. Princeton University.
- Turing, A. (2004). Intelligent machinery. En B. J. Copeland (ed.), *The essential Turing* (pp. 395-432). Oxford University Press.
- Van Parijs, Ph. (1991). Why surfers should be fed: the liberal case for an unconditional basic income. *Philosophy and Public Affairs*, 20(2), 101-131.
- Werthner, H., Prem, E., Lee, E. A. y Ghezzi, C. (eds.). (2022). *Perspectives on digital humanism*. Springer.
- Zimbardo, Ph. (2007). *The Lucifer effect. Understanding how good people turn evil*. Random House.



*Humanismo digital. Por una transformación humana de la democracia,
la economía y la cultura en la era digital*

editado por la Universidad Pedagógica Nacional, se compuso
en caracteres de las familias Dashiell Fine y Adelle,
y se imprimió en los talleres de Carvajal Soluciones
de Comunicaciones S. A. S.

Bogotá, Colombia, 2026

Colección

Comunicación, Tecnología y Culturas en Educación

La presente colección agrupa textos que exploran las repercusiones de las construcciones culturales relacionadas con los medios de comunicación y las TIC en la educación y la sociedad. Los diálogos continuos que estas áreas del conocimiento entablan con procesos de enseñanza y aprendizaje son de interés para la producción editorial por su creciente aparición como estrategias de creación de lenguajes que diversifican la disertación universitaria. En consecuencia, los libros incluidos en este repertorio pretenden acercar los discursos investigativos académicos de las áreas mencionadas a las realidades cotidianas de nuestro público lector.

La digitalización tiene enormes repercusiones económicas, también culturales y éticas: desde los carros autónomos y los robots que cuidan personas mayores, la entrega de correo controlado por *software* y redes sociales, la economía del *big data* y los *bots* inteligentes, hasta la industria 4.0. Este libro tiende un puente entre la filosofía y la ciencia ficción y desarrolla los fundamentos filosóficos de un humanismo digital en el que es esencial la distinción entre el pensamiento, los sentimientos y las acciones humanas, por un lado, y los procesos algorítmicos controlados por *software*, por el otro. Se propone como una alternativa a la ideología de Silicon Valley, para la que, según los autores, la inteligencia artificial amenaza con convertirse en un sustituto de la religión.

La traducción al español incluye breves reflexiones de estudiantes del Doctorado Interinstitucional en Educación de la Universidad Pedagógica Nacional, quienes aportan una mirada desde el contexto educativo nacional, así como de la discusión de las ciencias sociales y humanas en América Latina.

ISBN: 978-628-7851-74-0



Colección
Comunicación, Tecnología y Culturas
en Educación